Online ISSN : 0975-4172
Print ISSN : 0975-4350
DOI : 10.17406/GJCST

Global Journal

of Computer Science and Technology: C

Software & Data Engineering

Extended Linked Clustering Algorithms

A Review of Real World Big Data

Materialized view in Data Warehousing

Highlights

Clustering on Spatial Data Sets

Extended Linked Clustering Algorithms

Discovering Thoughts, Inventing Future

VOLUME 18 ISSUE 3 VERSION 1.0

Global Journal of Computer Science and Technology: C
Software & Data Engineering

Global Journal of Computer Science and Technology: C
Software & Data Engineering

Volume 18 Issue 3 (Ver. 1.0)

Open Association of Research Society

Incorporation No.: 0423089
License No.: 42125/022010/1186
Registration No.: 430374
Import-Export Code: 1109007027
Employer Identification Number (EIN):
USA Tax ID: 98-0673427

Global Journals Inc.

*(A Delaware USA Incorporation with "Good Standing"; **Reg. Number: 0423089**)*

Sponsors: Open Association of Research Society
Open Scientific Standards

Publisher's Headquarters office

Global Journals® Headquarters
945th Concord Streets,
Framingham Massachusetts Pin: 01701,
United States of America
USA Toll Free: +001-888-839-7392
USA Toll Free Fax: +001-888-839-7392

Offset Typesetting

Global Journals Incorporated
2nd, Lansdowne, Lansdowne Rd., Croydon-Surrey,
Pin: CR9 2ER, United Kingdom

Packaging & Continental Dispatching

Global Journals Pvt Ltd
E-3130 Sudama Nagar, Near Gopur Square,
Indore, M.P., Pin:452009, India

Find a correspondence nodal officer near you

To find nodal officer of your country, please email us at *local@globaljournals.org*

eContacts

Press Inquiries: *press@globaljournals.org*
Investor Inquiries: *investors@globaljournals.org*
Technical Support: *technology@globaljournals.org*
Media & Releases: *media@globaljournals.org*

Pricing (Excluding Air Parcel Charges):

Yearly Subscription (Personal & Institutional)
250 USD (B/W) & 350 USD (Color)

Editorial Board

Global Journal of Computer Science and Technology

Dr. Abdurrahman Arslanyilmaz

Computer Science & Information Systems Department
Youngstown State University
Ph.D., Texas A&M University
University of Missouri, Columbia
Gazi University, Turkey
Web: cis.ysu.edu/~aarslanyilmaz/professional_web

Dr. Chutisant Kerdvibulvech

Dept. of Inf. & Commun. Technol.,
Rangsit University
Pathum Thani, Thailand
Chulalongkorn University Ph.D. Thailand
Keio University, Tokyo, Japan

Dr. Sukhvinder Singh Deora

Ph.D., (Network Security), MSc (Mathematics),

Masters in Computer Applications

Dr. Sotiris Kotsiantis

Ph.D. in Computer Science, University of Patras, Greece
Department of Mathematics, University of Patras, Greece

Dr. Ramadan Elaiess

Ph.D.,

Computer and Information Science

Dr. Manpreet Singh

Ph.D.,
(Computer Science)

Nicla Romano

Professor in Cellular and Developmental Biology;
Cytology and Histology; Morfogenesis and Comparative Anatomy

Dr. Muhammad Abid

M.Phil,
Ph.D Thesis submitted and waiting for defense

Dr. K. Venkata Subba Reddy

Ph.D in Computer Science and Engineering

Loc Nguyen

Postdoctoral degree in Computer Science

Faisal Mubuke

M.Sc (IT), Bachelor of Business Computing, Diploma in Finanicial services and Business Computing

Jiayi Liu

Physics, Machine Learning,
Big Data Systems

Dr. Yuanyang Zhang

Ph.D in Computer Science

Asim Gokhan Yetgin

Design, Modelling and Simulation of Electrical Machinery;
Finite Element Method, Energy Saving, Optimization

Anup Badhe

Bachelor of Engineering (Computer Science)

Dr. S. Nagaprasad

M.Sc, M. Tech, Ph.D

Contents of the Issue

Global Journal of Computer Science and Technology: C
Software & Data Engineering
Volume 18 Issue 3 Version 1.0 Year 2018
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Spark Big Data Analysis of World Development Indicators

By Kunal Pritwani, Knox Wasley & Jongwook Woo

California State University

Abstract- We would like to analyze different development indicators such as life expectancy of a country, patent applications by residents, and trademark of applications which serves as great analyzation for the Business Analysts, Financial Analysts, and Data Scientists. Data set is collected from the World Bank websites, for this analysis, which has world development indicators with respect to the country. The data is analyzed based on the yearly timeline, geographical locations on the map and also top 10 countries for a particular world development indicator. Moreover, it has found that the countries where the life expectancy is high the people are more creative and the patent applications are created on a huge scale. Also, the trademark applications are more where the life expectancy is higher. This analysis provides insights on the world development indicators. In the paper, data analysis is done on a huge dataset by using Spark on Hadoop Big Data cluster and its visualization charts are presented.

Keywords: *big data, spark, databricks, life expectancy, trademark applications, patent applications, tableau hadoop, world development indicators data analysis.*

GJCST-C Classification: *H.3.m*

Strictly as per the compliance and regulations of:

Spark Big Data Analysis of World Development Indicators

Kunal Pritwani [α], Knox Wasley [σ] & Jongwook Woo [ρ]

Abstract- We would like to analyze different development indicators such as life expectancy of a country, patent applications by residents, and trademark of applications which serves as great analyzation for the Business Analysts, Financial Analysts, and Data Scientists. Data set is collected from the World Bank websites, for this analysis, which has world development indicators with respect to the country. The data is analyzed based on the yearly timeline, geographical locations on the map and also top 10 countries for a particular world development indicator. Moreover, it has found that the countries where the life expectancy is high the people are more creative and the patent applications are created on a huge scale. Also, the trademark applications are more where the life expectancy is higher. This analysis provides insights on the world development indicators. In the paper, data analysis is done on a huge dataset by using Spark on Hadoop Big Data cluster and its visualization charts are presented.

Keywords: big data, spark, databricks, life expectancy, trademark applications, patent applications, tableau hadoop, world development indicators data analysis.

I. Introduction

Our goal was to analyze the world development indicators which are important factors in big data analytics. This kind of data is analyzed by big name analysts for big money as this kind of analysis provides insights on different aspects of development indicators. The outcome of this analysis will help Business Analysts, Financial Analysts, and Data Scientists to compare between different development indicators and select the world development indicator that would have a better impact on the economy of the country.

Big Data is defined as non-expensive frameworks that can store a large scale data and process it in parallel [4, 5]. A large scale data means really a big data, this data cannot be processed using traditional computing techniques. Data is getting generated everyday through social media, websites, mobile applications etc. To analyze and store data we use Hadoop, which is an open source framework which provides distributed storage on the commodity hardware. Hadoop has two major components which are MapReduce and HDFS (Hadoop Distributed File System).

Authors α σ ρ: Department of Computer Information Systems, College of Business and Economics, California State University, Los Angeles.
e-mails: kpritwa@calstatela.edu, kwasley@calstatela.edu, Andjwoo5@calstatela.edu

Spark and tableau are adopted for data analyzation and visualization. Technically, to process a huge chunk of data a fast processor is required to process the big data. Hadoop is a technique which creates a cluster on the commodity hardware and stores and process the data on multiple nodes while having multiple copies of the same node on the cluster. Old Hadoop clusters use map reduce technique to map and store the data but Apache Spark is a newer version which is faster and runs on top of Hadoop architecture and it runs in memory.

Apache Spark runs 100 times faster than Hadoop but it doesn't have its own HDFS. So it uses HDFS as its file system and runs on top of Hadoop by using memory. Spark uses RDD (Resilient Distributed Datasets) which replaces the Map Reduce functionality to write the data to physical storage every time.

II. Related Works

Miranda and Michael does the data analysis using statistical techniques to find the correlation between different columns. But, we have used spark to manipulate and visualize the data to get useful insights [Chen, Ching 2000]. Life expectancy is analyzed by selecting the multiple columns using statistical techniques to find the correlation and by using scatter plots for visualization [Chen, Ching 2000]. We simply used geographical visualization to show top earning states. Paul uses basic approach for analysis of top countries for patent applications by using bar graph and we used geographical visualization with time series analysis using historical data [6]. Besides, Spark computation is less time consuming to process the results.

We have used Big Data Spark platform to store and analyze the data and BI tool such as tableau for visualizations. By analyzing the 247 countries data of 54 years, we have different results as we analyzed a very huge dataset. We have the detailed analysis for 247 countries and they have analysis for around top 10 countries. We have found that top countries where people are being more creative and innovative due to high life expectancy which helps the economic growth of the country. Spark helps to process the queries and gives the results fast and also spark has a very less lines of code as compared to other technologies.

III. Methods

First, we collected the data from an online community dedicated to data scientists where the dataset comprises of historical data of 5,656,458 rows which contain over a thousand annual indicators of economic development from hundreds of countries around the world. Further, by using the Spark technique to find different terminologies like we would like to analyze different development indicators like Life expectancy of a country, Patent applications by residents and Trademark of applications. Detailed Analysis of world development indicators has been performed using data visualization tools.

a) Specification of Data Set

The data is collected from an online community kaggle. We have historical data of about 5,656,458 rows which contain over a thousand annual indicators of economic development from hundreds of countries around the world. To be precise, there are 1344 indicators and 247 countries in this dataset spanning of 54 years. The data size is 574 MB and file is in CSV (Comma Separated Values) format.

b) Platforms

Data Analysis tools used are Apache Spark cluster on Databricks cloud platform, and visualization tool Tableau 9.2 is used for detailed data analysis for daily and yearly records.

i. *Cloud Computing: Databricks*

In order to collect and analyze data, cloud computing service from Databricks is adopted. Databricks provides a very fast data platform which runs on top of apache spark that helps to easily create big data advanced analytics solution. It can *be connected directly to the existing storage apache clusters and Databricks services in the cloud. It provides highly integrated workspace to* create dashboards by using notebooks. Also it provides a functionality to use third party Business Intelligence tools and create custom spark applications with spark production jobs. Basically Databricks provides a very easy to use cloud platform which reduces the use of high end and expensive hardware. There is no need to install the Apache Spark environment which is a huge time saver for Data scientists, Data analysts and Data engineers. It also provides option to choose different spark versions like Spark 1.6.1(Hadoop 1).

There are various programing languages available on Databricks in form of different notebooks like Scala, python, SQL and R. The instances are highly scalable which can be modified as per the user. Also the visualization can be done instantly instead of exporting it to other visualization tools. Notebook access can be given to multiple users to edit or read [2].

Figure 1. Shows the different languages spark API supports like R, SQL, Python, Scala and Java. Also shows that spark supports data frames, streaming, machine learning and GraphX usability.

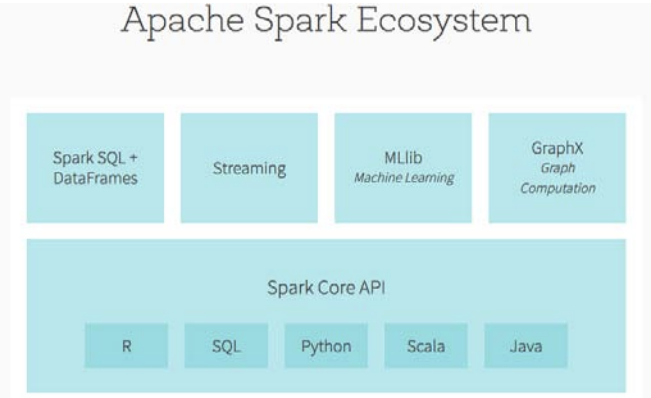

Figure 1: Apache Spark Ecosystem

In the Spark service of Databricks, the code is written in the ipython notebook. The spark cluster is accessed by the notebook, on which the query processing is done. In to the cluster, we have to upload the data file, in this case its 'indicators_csv'. We can change the data_type also during the creation of the table. We have to create RDDs using SQL context and run the queries which are discussed in the analysis part.

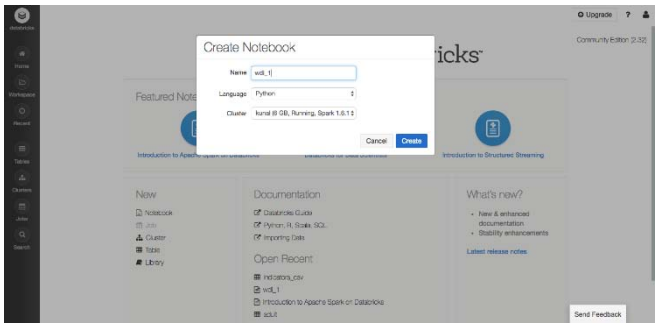

Figure 2: Creating a notebook in Databricks

Figure 2. Shows a screenshot of a notebook creation where it gives you the option of selecting a language, In this case it's python also called as 'PySpark'. Also gives you the option of selecting a cluster in this case its Spark 1.6.1. This cluster is Apache Spark Version (Spark 1.6.1) of Databricks. Memory is 6GB with 0.88 Cores CPU cores and 1 CPU master node.

Figure 3: Table format in Databricks

Figure 3. Shows a table which was imported 'indicators_csv' and here we can also modify the datatype of the column. The PySpark code is built, where Databricks automatically takes a default datatype.

ii. *Visualization: Tableau*

Tableau is adopted for visualizing the result data set that is computed in Spark, which is a business intelligence tool. It is easy to use and produce interactive visualizations to get the insights using data analytics techniques. Tableau provides to the traditional small data set a user friendly and powerful environment for Data Scientist, Data Analyst and Data Engineers. It can produce visualization from relational database, cloud databases and excel files. However, it cannot compute huge data set. The data analysis code is built and run in Spark in order to generate and find out insights, which is the result data set and small amount of data. Thus, Tableau can take the result data set of the data analysis in Spark and it can produce the graphs in the next sections. It provides a number of types of graphs like bar graphs, pie charts, line graphs, geographical maps, Gantt chart etc. Tableau is used to get the hidden insights from the data which can help to improve the world by implementing changes to the world development indicators [9].

c) *Terminology*

i. *Life expectancy at birth, total (years)*

Derived from male and female future during childbirth from sources, for example, (1) United Nations Population Division. Total populace Prospects, (2) Census reports and other factual productions from national measurable workplaces, (3) Eurostat: Demographic Statistics, (4) United Nations Statistical Division. Populace and Vital Statistics Report (different years), (5) U.S. Enumeration Bureau: International Database, and (6) Secretariat of the Pacific Community: Statistics and Demography Program [3].

ii. *Patent applications, residents*

World Intellectual Property Organization (WIPO), WIPO Patent Report: Statistics on Worldwide Patent Activity. The International Bureau of WIPO accepts no accountability concerning the change of these information [3].

iii. *Trademark applications, total*

Trademark applications documented are applications to enlist a trademark with a national or local Intellectual Property (IP) office. A trademark is an unmistakable sign which distinguishes certain merchandise or administrations as those created or gave by a particular individual or venture. A trademark gives assurance to the proprietor of the check by guaranteeing the select appropriate to utilize it to distinguish products or benefits, or to approve another to utilize it as an end-result of installment. The time of security fluctuates, however a trademark can be reestablished uncertainly past as far as possible on installment of extra charges [3,11].

IV. Detail Data Analysis Results

a) *Life expectancy at birth, total (years)*

This formula selects columns CountryName and Value with a filter on Indicator Name as "Life expectancy at birth, total (years)" Results are stored in 'results' RDD and then displayed using Spark Display command [8].

➔ Results= sqlContext.sql('Select CountryName, Value from indicators_csv where IndicatorName = "Life expectancy at birth, total (years)" order by Value desc')

➔ Display (results)

Figure 4. Shows the geographical view of life expectancy on the map. Life expectancy has a high value and the lighter regions have the less value. United States is dark green so it means that the Life expectancy is good in United States and In Africa the area is light green which means the life expectancy is less.

Figure 5 shows the top countries which have high value for life expectancy. In this case San Marino has the highest average value in world for life expectancy as 81.49.

Figure 6. Shows the Lowest Life Expectancy at birth, Top 10 countries. In this case it shows that Seirra Leone has the lowest value for life expectancy as 38.68.

Figure 7. Shows the life expectancy of United States from 1960 to 2013. The trend shows that the life expectancy is increasing which increases the United States Growth.

b) *Patent applications, residents*

This formula selects columns CountryName and Patent applications, residents with the filter on column IndicatorName as "Patent applications, residents". Results are stored in 'results' RDD and then displayed using Spark Display command [8]. Refer the code at Github[12].

Figure 8. Shows Patent applications by residents, Top 10 countries. In this graph we can see that Japan and USA has the highest Average value all around the world.

Figure 9. Shows Patent applications by residents, Lowest Top 10 countries. In this graph we can see that Aruba has the lowest Average value all around the world.

Figure 10. Shows the geographical view of Patent applications by residents on the map. Patent applications which have a high value are darker regions and the lighter regions have the less value. Japan, United States are dark blue so It means that the Patent applications are more in United States, China and Japan and in Africa the area is light blue which means the life expectancy is less.

Figure 11. Shows the Patent applications by residents of United States from 1960 to 2013. The trend shows that the Patent applications is increasing which increases the United States Growth.

c) Trademark applications, total

This formula selects columns CountryName and Trademark applications, residents with a filter on IndicatorName as "Trademark applications, total". Results are stored in 'results' RDD and then displayed using Spark Display command [8].Refer the code at Github[12].

Figure 12. Shows Trademark applications for Top 10 countries. In this graph we can see that China, Japan and USA has the highest Average value all around the world.

Figure 13. Shows the geographical view of life expectancy on the map. Trademark applications have which have a high value are darker regions and the lighter regions have the less value. China, Japan, United States are dark yellow so It means that the Trademark applications are more in China and In Africa the area is light which means less number of Trademark applications.

Figure 14. Shows the Trademark applications of United States from 1960 to 2013. The trend shows that the Trademark applications are increasing which increases the United States Growth.

Figure 4: Life Expectancy at birth on the Map

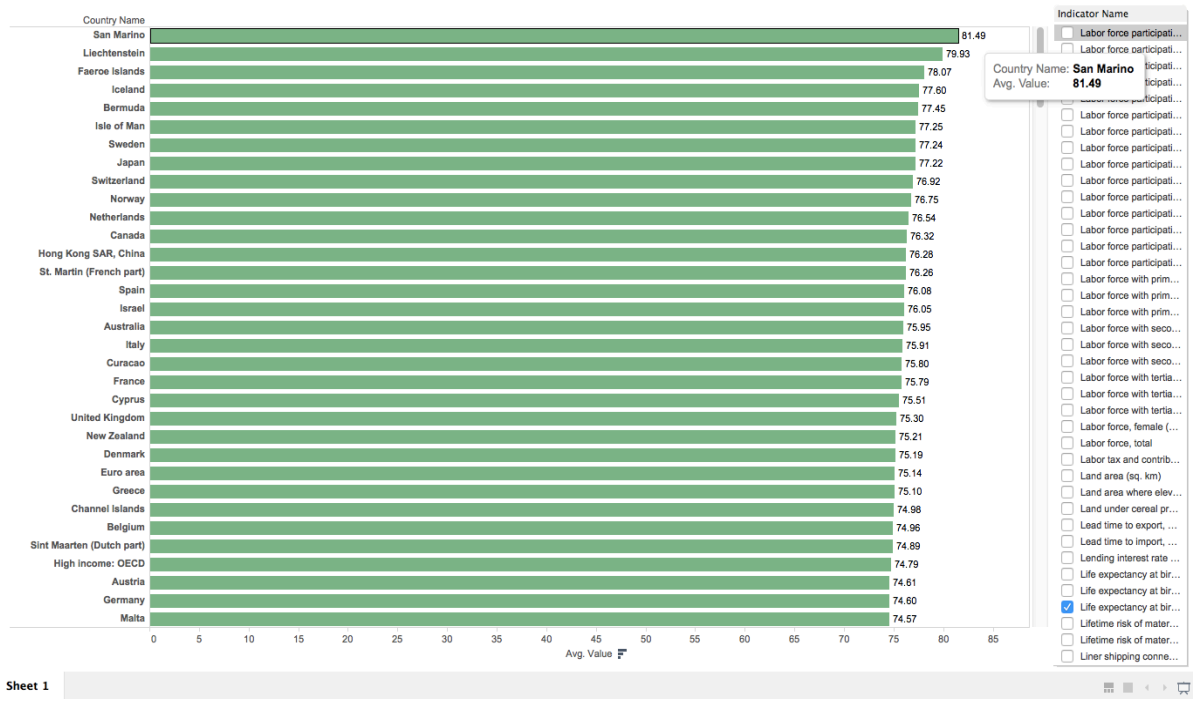

Figure 5: Life Expectancy at birth, Top 10 countries

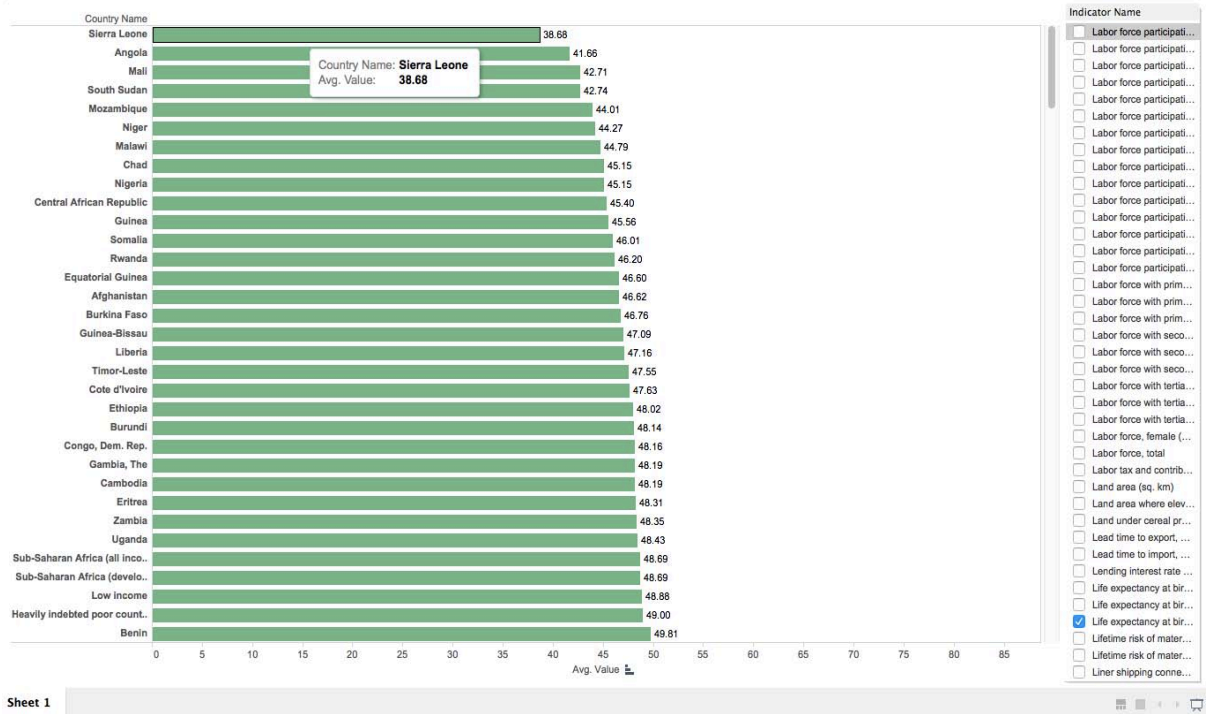

Figure 6: Lowest Life Expectancy at birth, Top 10 countries

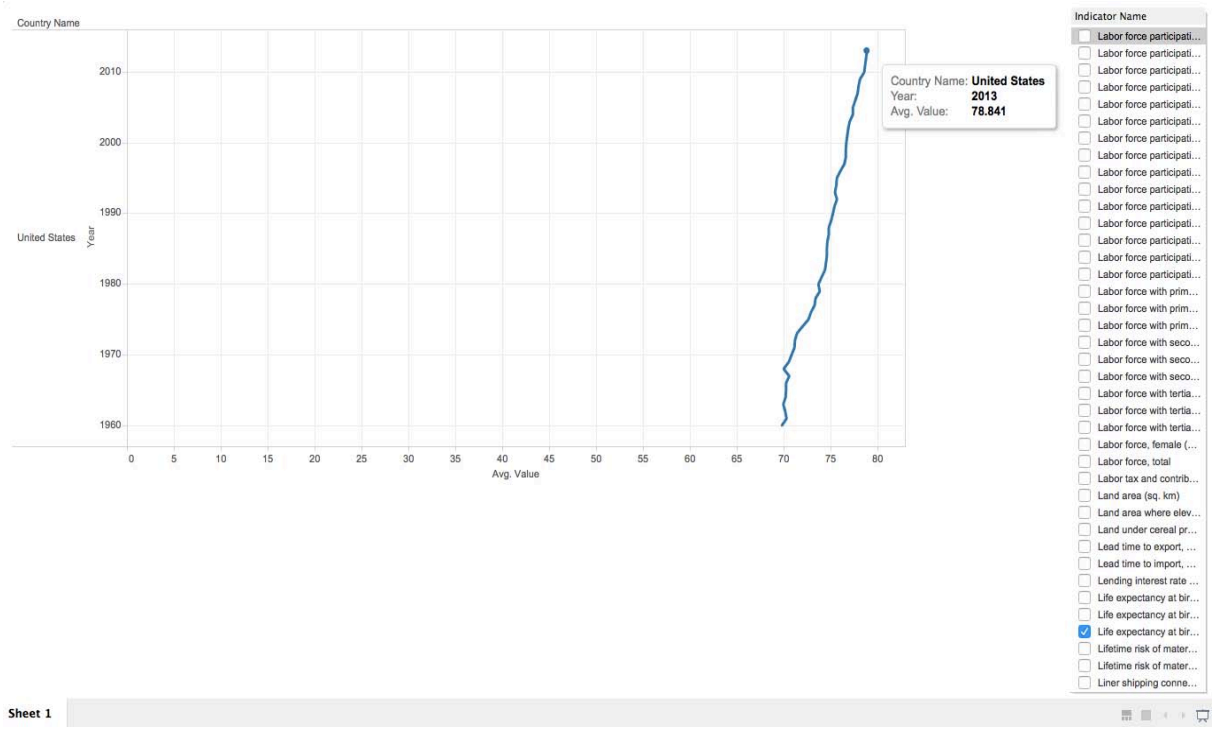

Figure 7: Life Expectancy at birth, from 1960 to 2013(United States)

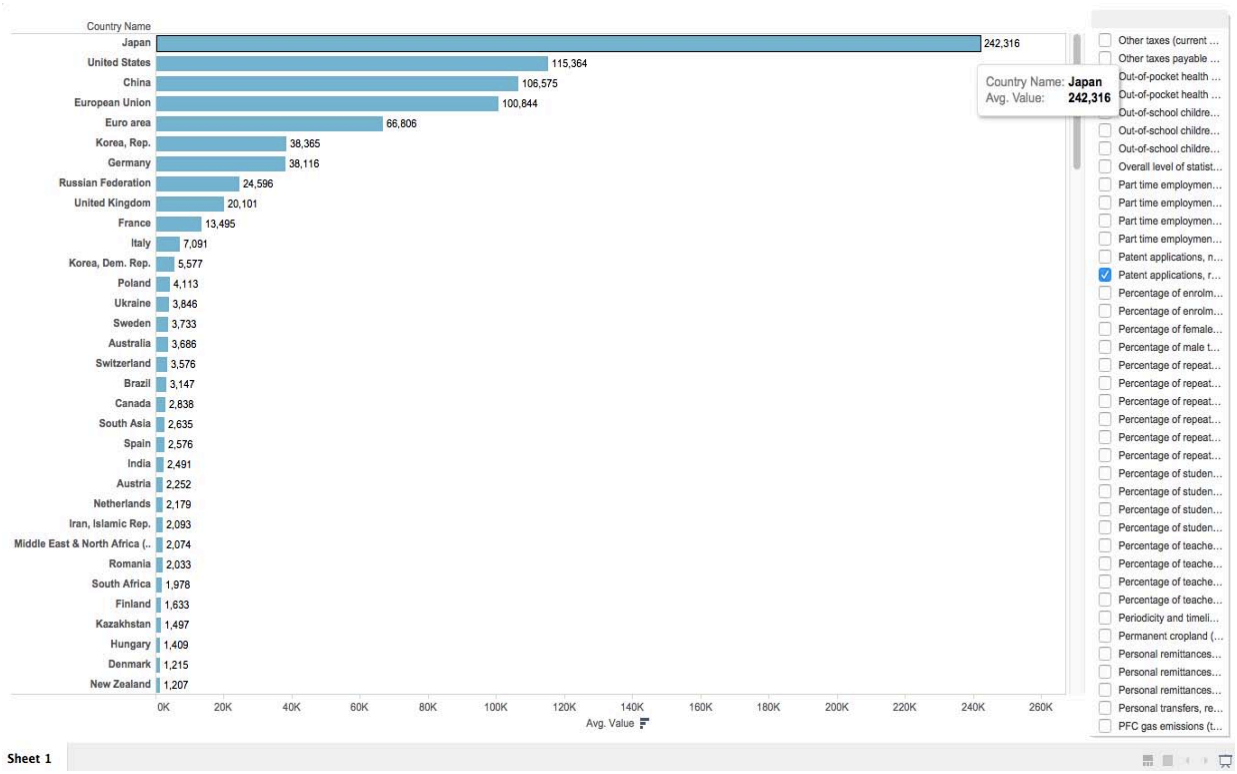

Figure 8: Patent applications by residents, Top 10 countries

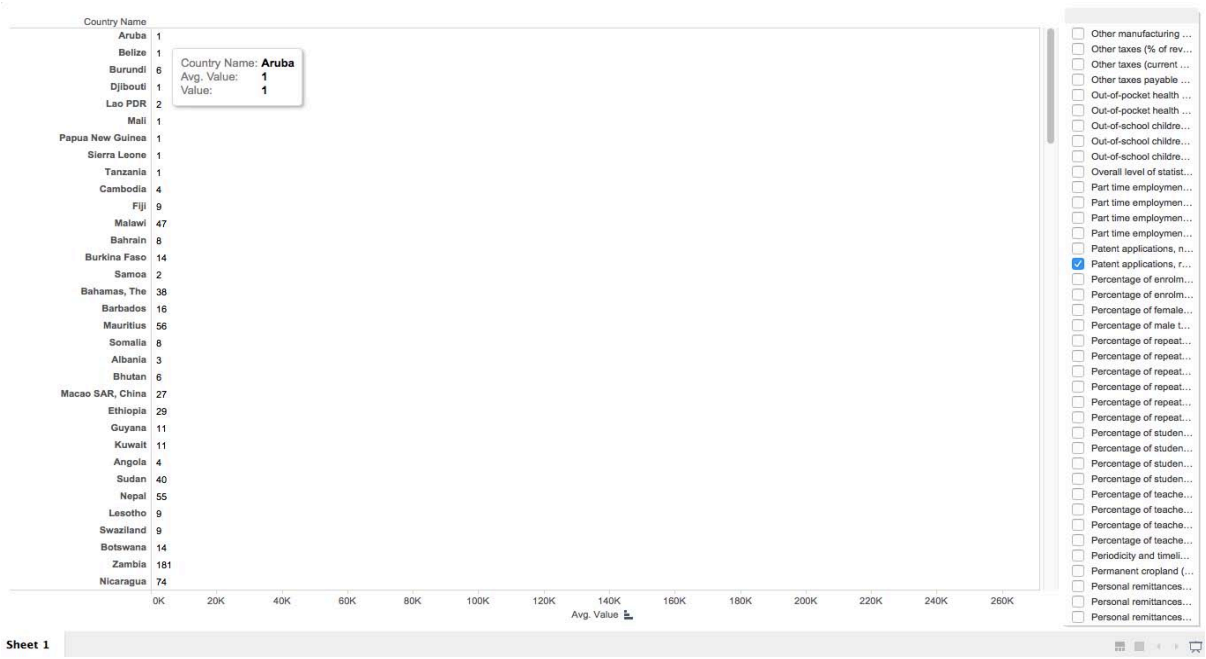

Figure 9: Patent applications by residents, Lowest Top 10 countries

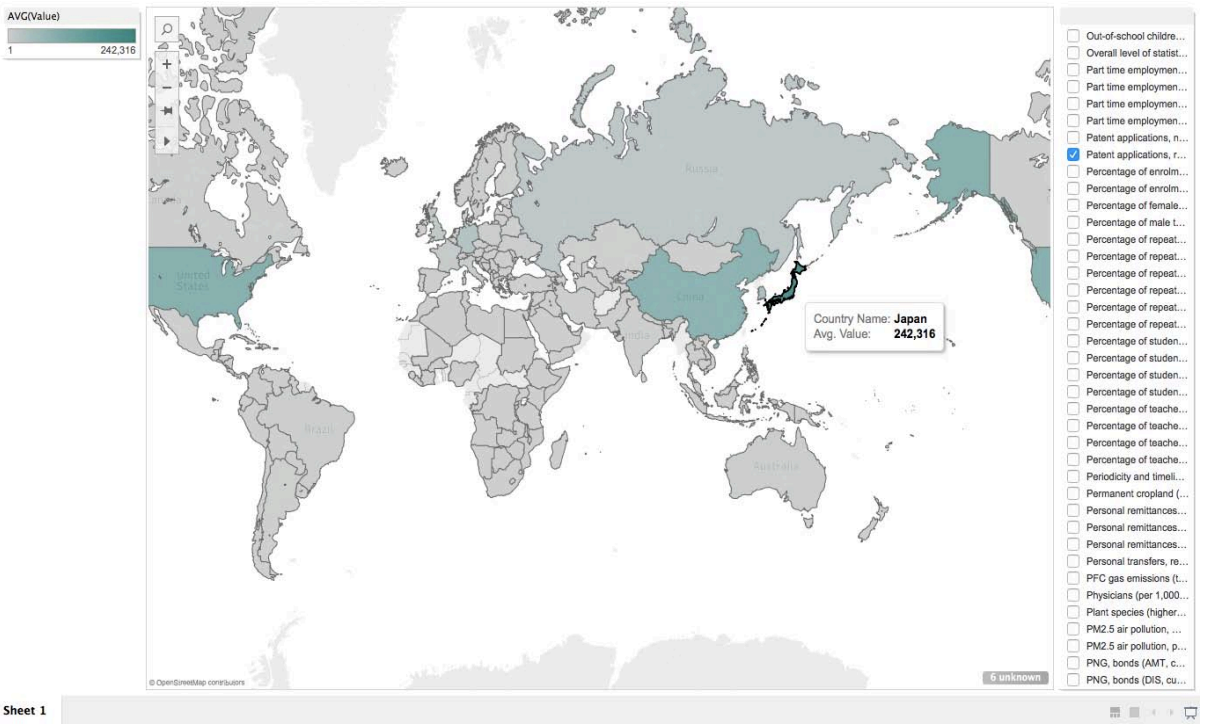

Figure 10: Patent applications by residents on map

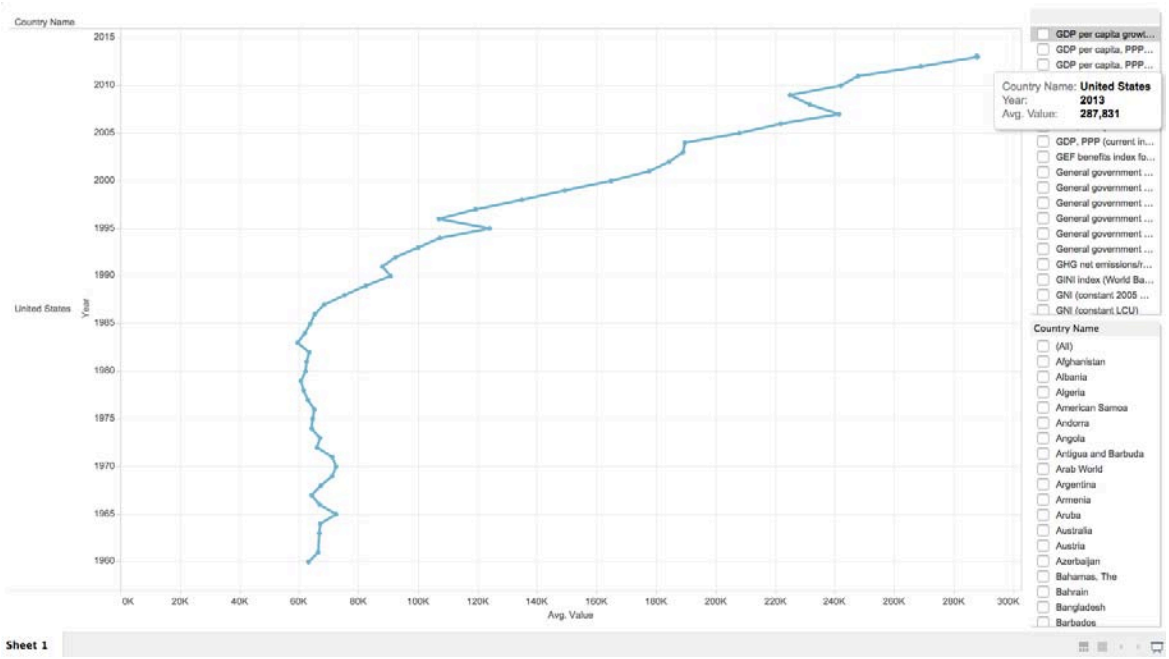

Figure 11: Patent applications by residents, from 1960 to 2013(United States)

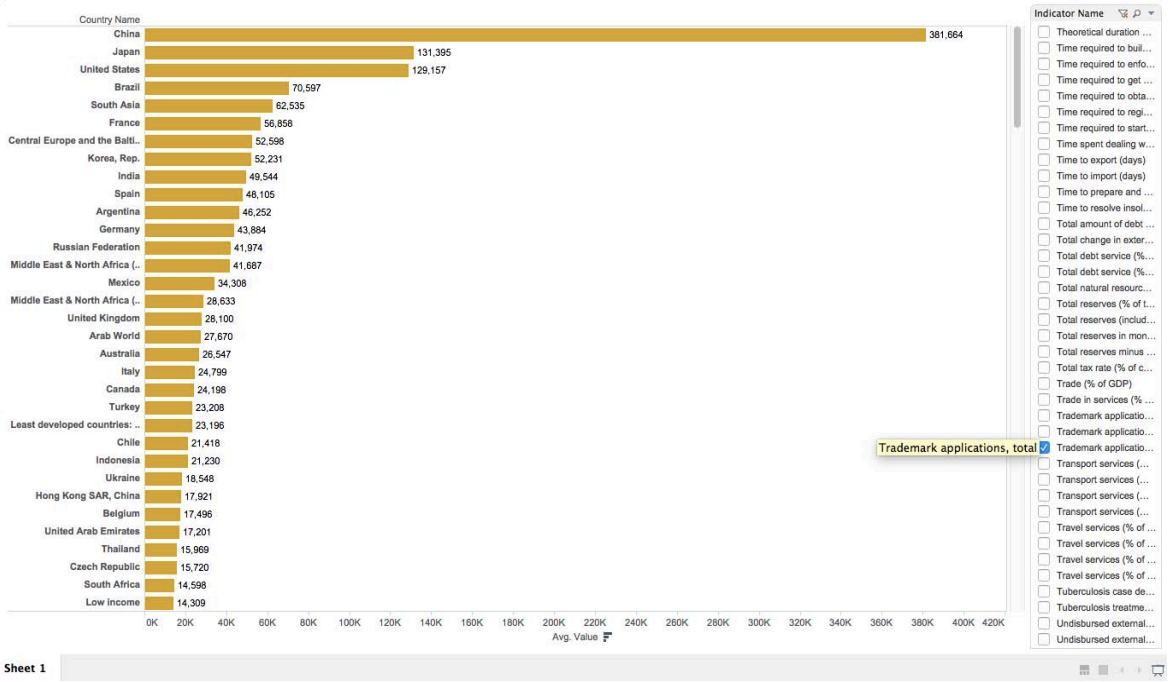

Figure 12: Trademark applications, Top 10 countries

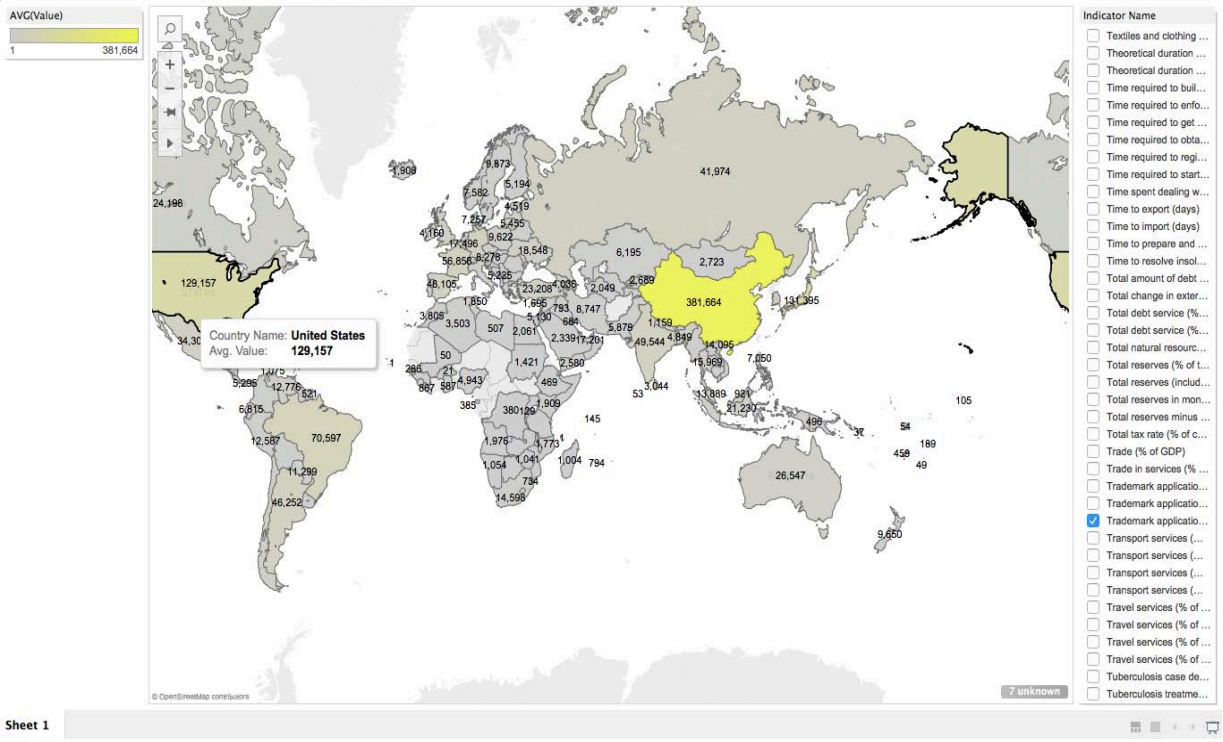

Figure 13: Trademark applications on map

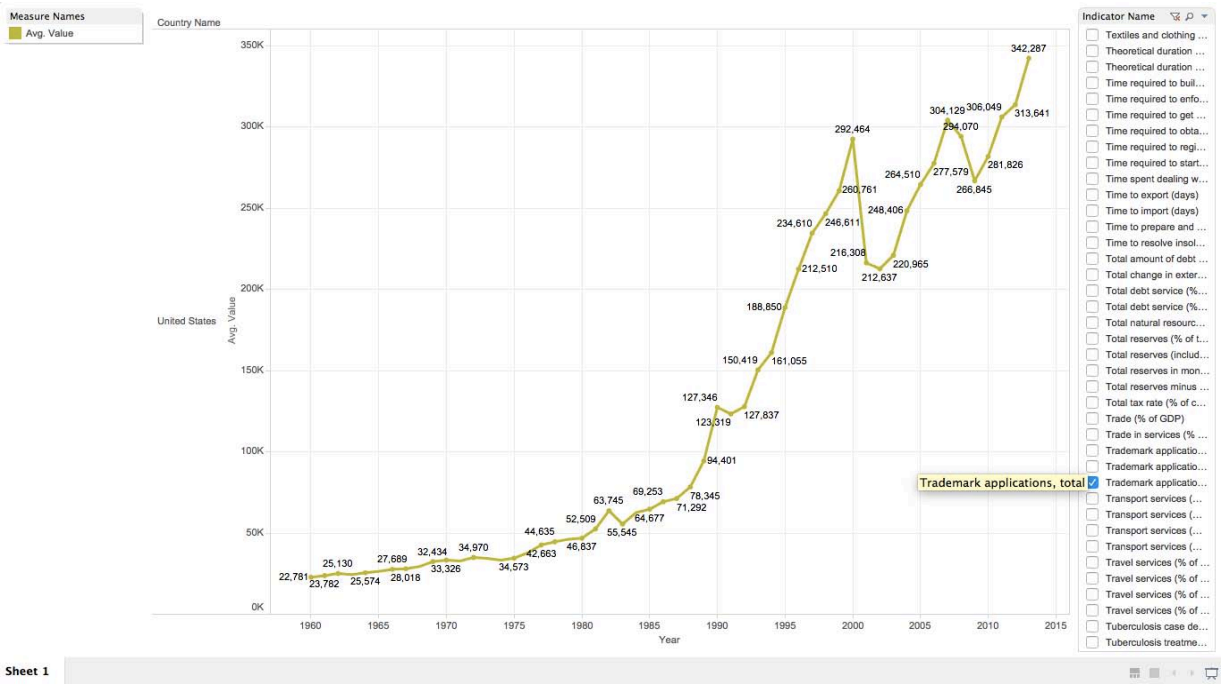

Figure 14: Trademark applications, from 1960 to 2013(United States)

V. Conclusion

To conclude, world development indicators reflect the growth of the country using big data analytics. This kind of insight will be charged huge sum by data analyst for what we just presented with the analysis insights. It has found that the countries where the life expectancy is high, the people are more creative and the patent applications are created on a huge scale. Also the trademark applications are more where the life expectancy is higher. This analysis provides insights on the world development indicators. These analysis is helpful for decision making at a higher level where the growth factors of the country are planned.

It is found that life expectancy and creativity of the people are correlated. The more people live, the more creative they become. They will create more creative applications and trademarks for the betterment of the country and world. Technology plays a huge role in finding these insights as it helps the analysts for decision making. Big data technology is the future for decision making systems as the data is getting bigger and bigger every day and we need to analyze, process, store and fault tolerance this big data. Databricks helps users to easily create Apache Spark Hadoop clusters and run the queries on huge chunks of data. Tableau provides a wide variety of data visualization options to gather some use full insights like "finding a gold in Data Lake" so that better decisions can be made for the development of the country and world.

Acknowledgement

This research was supported by Amazon AWS research grant.

References Références Referencias

1. Chen, M., and Ching, M. (2000, December 18). "A STATISTICAL ANALYSIS OF LIFE EXPECTANCY ACROSS COUNTRIES USING MULTIPLE REGRESSION". Retrieved April 15, 2017, from http://www.seas.upenn.edu/~ese302/Projects/Project_2 pdf
2. Databricks. (2016). "What is Apache Spark?" Retrieved November 02, 2016, from https://databricks.com/spark/about
3. Worldbank. (2013). "Life expectancy at birth, total (years)". Retrieved November 10, 2016, from http://data.worldbank.org/indicator/SP.DYN.LE00.IN
4. Woo, J., and Xu Y. 2011. "Market Basket Analysis Algorithm with Map/Reduce of Cloud Computing", The 2011 international Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA 2011), Las Vegas (July 18-21, 2011).
5. Woo, J. (2013) "Market Basket Analysis Algorithms with MapReduce", DMKD-00150, Wiley Interdisciplinary Reviews Data Mining and Knowledge Discovery, Oct 28 2013, Volume 3, Issue 6, pp445-452, ISSN 1942-4795.
6. Muggeridge, P. (2014, May 17). "Which countries file the most patent applications?". Retrieved April 16, 2017, from https://www.weforum.org/agenda/2015/09/which-countries-file-the-most-patent-applications/
7. Worldbank. (2013). "Patent applications, residents". Retrieved December November 14, 2016, from http://data.worldbank.org/indicator/IP.PAT.RESD
8. Spark Apache. (2016). "Spark SQL Data Frames and Datasets Guide". Retrieved November 29, 2016, from http://spark.apache.org/docs/latest/sql-programming-guide.html
9. Tableau Software. (2016). "Business intelligence for your people". Retrieved November 02, 2016, from https://www.tableau.com/resource/business-intelligence
10. Worldbank. (2013) "Trademark applications, total". Retrieved November 19, 2016, from http://data.worldbank.org/indicator/IP.TMK.TOTL
11. Trademark Economics. (2014). "Trademark applications - total in World". Retrieved November 27, 2016, from http://www.tradingeconomics.com/world/trademark-applications-total-wb-data.html
12. Github, "https://github.com/pritwanikunal/Big-Data-Analysis-of-World-Development-Indicators-using-Apache-Spark"

Global Journal of Computer Science and Technology: C
Software & Data Engineering
Volume 18 Issue 3 Version 1.0 Year 2018
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Incremental Maintenance of a Materialized View in Data Warehousing: An Effective Approach

By Dr. Sanjay S Solank

JSPMs Abacus Institute of Computer Application

Abstract- A view is a derived relation defined in terms of base relations. A view can be materialized by storing its extent in the database. An index can be made of these views and access to materialized view is much faster that recomputing the view from scratch. A Data Warehouse stores large amount of information collected from a different data sources. In order to speed up query processing, warehouse usually contains a large number of materialized views. When the data sources are updated, the views need to be updated. The process of keeping view up to date called as materialize view maintenance. Accessing base relations for view maintenance can be difficult, because the relations may be being used by users. Therefore materialize view maintenance in data warehousing is an important issue. For these reasons, the issue of self-maintainability of the view is an important issue in data warehousing. In this paper we have shown that a materialized view can be maintained without accessing the view itself by materializing additional relations at the data warehouse site.

Keywords: optimized view, ETL, incremental maintenance, view maintenance process, DMWS, view synchronization, expression tree.

GJCST-C Classification: H.2.7

Strictly as per the compliance and regulations of:

Incremental Maintenance of a Materialized View in Data Warehousing: An Effective Approach

Dr. Sanjay S Solank

Abstract- A view is a derived relation defined in terms of base relations. A view can be materialized by storing its extent in the database. An index can be made of these views and access to materialized view is much faster that recomputing the view from scratch. A Data Warehouse stores large amount of information collected from a different data sources. In order to speed up query processing, warehouse usually contains a large number of materialized views. When the data sources are updated, the views need to be updated. The process of keeping view up to date called as materialize view maintenance. Accessing base relations for view maintenance can be difficult, because the relations may be being used by users. Therefore materialize view maintenance in data warehousing is an important issue. For these reasons, the issue of self-maintainability of the view is an important issue in data warehousing. In this paper we have shown that a materialized view can be maintained without accessing the view itself by materializing additional relations at the data warehouse site. We have developed a cost effective approach to reduce the burden of view maintenance and also proved that proposed approach is optimum as compared to other approaches. Here incremental evaluation algorithm to compute changes to materialized views in relational is presented.

Keywords: optimized view, ETL, incremental maintenance, view maintenance process, DMWS, view synchronization, expression tree.

I. Introduction

It has been observed that in most typical data analysis and data mining applications, timeliness and interactivity are more important considerations than accuracy; thus, data analysts are often willing to overlook small inaccuracies in the answer, provided that the answer can be obtained fast enough. This observation has been the primary driving force behind the recent development of approximate query processing techniques for aggregation queries in traditional databases and decision support systems [4], [5]. Numerous approximate query processing techniques have been developed: The most popular ones are based on random sampling, where a small random sample of the rows of the database is drawn, the query is executed on this small sample, and the results are extrapolated to the whole database. In addition to simplicity of implementation, random sampling has the compelling advantage that, in addition to an estimate of the aggregate, one can also provide confidence intervals of the error, with high probability. Broadly, two types of sampling-based approaches have been investigated: 1) pre-computed samples, where a random sample is pre-computed by scanning the database and the same sample is reused for several queries and 2) online samples, where the sample is drawn "on the fly" upon encountering a query. So the selection of these random samples in distributed environments for query processing is addressed in [6]. Data warehouses (DW) [6] are built by gathering information from data sources and integrating it into one virtual repository customized to users' needs. One important task of a Data Warehouse Management System (DWMS) is to maintain the materialized view upon changes of the data sources, since frequent updates are common for most data sources. In addition, the requirements of a data source are likely to change during its life-cycle, which may force schema changes for the data source. A schema change could occur for numerous other reasons, including design errors, the addition of new functionalities and even new developments in the modeled application domain. Even in fairly standard business applications, rapid schema changes have been observed. In [10], significant changes (about 59% of attributes on the average) were reported for seven different applications over relational databases. A similar report can also be found in [15]. These applications ranged from project tracking, sales management, to government administration.

In situations that real-time refreshment of the data ware-house content is not critical; changes to the sources are usually buffered and propagated periodically such as once a day to refresh the view extent. Two benefits are possible. One is to gain better maintenance performance. The other is that there are less conflicts with DW read sessions. In a data update only environment, most view maintenance (VM) algorithms proposed in the literature [17, 1, 14] group the updates from the same relation and maintain such a large delta change in a batch fashion. However, these algorithms would fail whenever source schema changes occur, which are also common as stated above. One obvious reason is that the data updates in this group may be schema inconsistent with each other if there are some schema changes in between. On the other hand, work has begun on incorporating source schema changes into the data warehouse, namely, view synchronization (VS) [8] aims at rewriting the DW view definition when the source schema has been changed.

Author: Professor, JSPM's Abacus Institute of Computer Application, Pune. e-mail: sanjay.solanki123@gmail.com

To handle the delete of any schema information of a data source, VS tries to locate an alternative source for replacement to keep the new view semantically as close to the original view as possible. Thereafter, view adaptation (VA) [12] incrementally adapts the view extent to keep the new view consistent. Such algorithms are also not sufficient to batch a group of mixed data updates and schema changes, since there could be a number of schema changes interleaved with some data updates. In this paper, we propose a solution strategy that is capable of batching a mixture of both source data updates

II. Definition of Terms

View evaluation can be represented by a tree, called an expression tree[5,9]. An expression tree is a tree, where the leaf nodes represent base relations and non-leaf nodes represent binary expressions in the relational algebra. The unary relational algebraic expressions are associated along the edges. A view or a query is optimized by the query optimizer before executing it. A query optimizer takes an expression tree as input and produces an output, called an optimized expression tree, which determines the internal sequence of operations for executing a query. Thus, an optimized expression tree defines a partial order in which operations must be performed in order to produce the result of the view.

Depth: The depth of leaf nodes, that is data base relations is 0. The depth d of a node is defined as max(depth of descendents)+1.

Height: The height of the optimized expression tree is defined as the maximum depth of any node in the tree.

Given a node i in the expression tree, its parent is denoted by ↑i, and op(i) and op(↑i) are the expressions associated with i and i, respectively. The children of node i are denoted by i' and i'' where i' is a sibling of i'' and vice versa. IR_i denotes the intermediate result of node i. The auxiliary relation associated with node I is denoted AR_i in the case where only one relation is needed, and by AR^1i and AR^2i when two are needed. The key of IR_i is denoted by K_i, and the keys of $IR_{i'}$ and $IR_{i''}$ are denoted by $K_{i'}$ and $K_{i''}$, respectively. Insertion and deletion of tuples are denoted by $\triangle$ and ∇ respectively. The symbol δ either an inserted set or a deleted set of tuples. The instance of a relation, say Ri, before and after an update is denoted by Ri^{old} and Ri^{new} respectively, similiary for an auxiliary relation AR and a materialized view V.

III. Example & Simplification

Consider a data warehouse for a large research organization which has got many departments and each department has many research groups. Suppose this data warehouse is collecting data from four base relations whose schemas are as follows:

R1: emp_rschr(rschr_id,rname,deptno,major) This relation gives the researchers id, name, department and major.

*R2:*emp_paperpublish(rschrid,paper_id,paper_title,source_of_publiscation, year_of_publish)

This gives researchers id,paper id, paper title, source of publication and year of publish.

R3: emp_manager(rschr_id,deptno)

This relation contain one record for each manager and his department. Assume that each department has one manager. Since a manager is also a researcher, relation emp_rschr has a tuple for each manager.

R4: emp_groupleader(rschr_id,deptno)

This relation contains information about th research group name and who is leading this group. Since a group leader is also a researcher, relation emp_rshcr has a tuple for each group leader.

Suppose a user of the organization is interested in materializing and maintaining the following view:

'Researchers other than managers and group leaders along with their departments who have published more than 10 papers in the year 2010.'

In SQL, it is defined as a sequence of view definitions:

```
Create view mngr_or_groupleader (rschr_id, deptno) as select rschr_id, deptno from emp_rschr
  UNION
  (select rschr_id, deptno from emp_groupleader)
// This view is for finding manager and group leader
Create view rschr_ex_ manager_or_groupleader (rschr_id, deptno) as select reshr_id, deptno from emp_rschr where NOT EXISTS (select *from mngr_or_grouple ader where emp_rschr.id=mngr_or_groupleader.id)
//This view is for finding researcher, those are not manager or group leader.
Create view rschrpaperview2010 (rschr_id, paper_id, deptno) as select emp_paperpublish.rschr_id, paper_id, deptno from rschr_ex_manager_or_group leader, emp_paperpublish where rschr_ex_ manager_or_ group leader.rschr_id=emp_paperpublish.rschr_id and year='2010'.
//This view gives the researcher those who have published paper in the year 2010.
Create view rschrpaperview(rschr_id,deptno) as
  Select rschr_id, deptno from rschrpaper view2010 group by rschr_id having count(*)>10;
// This view gives the researcher who published more than 10 research paper in the year 2010.
```

As base relations are updated, changes representing the researchers data come into the warehouse. Most warehouse do not apply the changes immediately. Instead, changes are deferred and applied

to the auxiliary relations incrementally. Deferring the changes allows analysts that query the warehouse to see a consistent snapshot of the data throughout the day, and can make the maintenance more efficient. Figure 1 shows the optimized expression tree for the above view. Here, the nodes at leaf level are base relations and non-leaf nodes are expressions. Each non-leaf node in the tree corresponds to a relational algebraic expression given above.

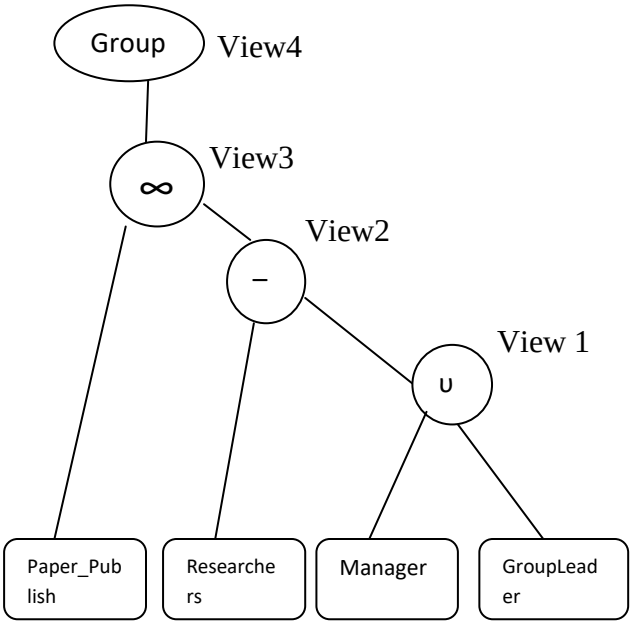

Figure 1: Expression tree

Suppose Researchers or Paper_Public relations are updated. In this case we materialize the two auxiliary relations View2 and View3. The contents of these views are derived while computing the view first time. By materializing these two auxiliary relations in the warehouse, the view is self-maintainable along with these auxiliary relations. Suppose new researchers joined the organization, therefore, one tuple for each new researcher in emp_rschr relation has to be inserted. These insertions will led to generate tuples that to be inserted in rschr_ex_manager_groupleader. Since these new researchers have not published any paper at the time of joining, these tuples cannot join with any tuples of emp_paper_publish, thus there will no change in the materialized views. Therefore, all auxiliary relations and materialized views are self_maintainable. Now consider another case where a set of tuples is inserted in emp_paper_publish relation, say R. Then, we first compute the research paper those are published in year 2010 and then it is join with rschr_ex_managergroupleader view. Lastly the intermediate result is grouped in the final auxiliary relation by performing count operation. In this case also, the view and auxiliary relations are self-maintainable.

IV. Procedure of Materialize Views Maintenance

The materialize view maintenance process can be divided into two functions: 1. Propagate and 2. Refresh. The work of computing the auxiliary relations happens within the propagate function, which can take place without locking materialize views so that the warehouse can continue to be made available for querying by analysts. Materialize views are not locked until the refresh function, during which time the materialize views are updated from the auxiliary relations.

The propagate function involves updating the auxiliary views incrementally from deferred set of changes. The final auxiliary view represents the net changes to the materialize views due to the changes in the underlying data sources.

The refresh function applies the net changes represented in the final auxiliary relation to the materialize views. This process carried out after a specific time interval or when the system has free cycles. So none of the data warehouse users or operations are affected by the view maintenance process. None of the query has to pay for view maintenance. The materialize view maintenance process totally hidden by users and running transactions. Whenever an interested change happens in the underlying data source, simply this desire change is stored in the auxiliary relations by comparing and joining it with others relations if required. This change is passed to the higher level auxiliary relations. Again the change is integrated and circulated to final auxiliary relation. Lastly the change is refreshed into the data warehouse when the refresh trigger is occur.

a) Analytical Cost Model

In this section we show the performance results of our materialize view maintenance method. The results are based on the following cost model.

i. *Cost Model*

The overall view maintenance cost of materialized views includes the cost of propagate the changes and the cost of refresh operations. Let $V_1,V_2,\ldots,V_m$ be the m materialized views. Let $B_1, B_2,\ldots,B_n$ be the n base relations and $A_1,A2,\ldots A_i$ be the i auxiliary relations. Let f_{u1}^{B1} ,.....,f_{un}^{Bn} be the update frequency to the base relations. Let $C^{ij}_{B->A}$ be the cost of propagating an update on base relation B_i to auxiliary relation A_j and $C^{jk}_{A->V}$ be the cost of refresh of auxiliary A_j to materialized view V_k. The overall cost of maintaining the views when keeping both the materialized views and the auxiliary relations is:

$$C_{MV}+AR=\sum_{i=1}^{i=n}\left(f_{ui}^{Bi}\right)*\left(\sum_{j=1}^{j=1} C-> A\binom{n}{k}x^k a^{n-k}\right)^k a^{n-k}$$

The total view maintenance cost with no auxiliary relations is:

$$C_{MV}=\sum_{i=1}^{i=n}\left(f_{ui}^{Bi}\right)*\left(\sum_{k=1}^{i=n} C\right.$$

It is obvious that the cost of maintaining the materialized views directly from base relations is much more than the cost of maintaining materialized views through auxiliary relations.

V. Evaluation

To verify the feasibility and effectiveness of our view maintenance strategies and corresponding optimization framework, we have implemented the proposed techniques using Oracle 9i. All experiments were performed on a workstation with Pentium D 3.2 GHz processor, 1 GB of memory and 160 GB disks, running Windows XP.

Relation R1 contain 500000 records, R2 contains 25000 records, where as in R3 there are records of individual manager of a department and in R4 holds the records of group leaders.

We considered two types of changes:

Update-Generating changes: Insertions and deletions of an equal number of tuples over existing researchers and paper publishers. These changes mostly cause updates amongst the existing tuples in materialized view.

Insertion-Generating changes: Insertions over new researchers those who published certain number of research papers. These changes cause only insert into paper publish table.

The insertion-generating changes are very meaningful since in many data warehousing applications the only changes to the fact tables are insertions of tuples for new dates, which leads to insertions into materialized views.

Figure 2 shows four graphs illustrating the performance advantage of using incremental materialized view maintenance method which uses auxiliary views to store intermediate results. The view maintenance time is split into two functions propogate and refresh. While computing the intermediate result the data warehouse is remain free to the user.

Figure 2 (a) and (b) plot the variation in elapsed time as the size of the change set changes(delta relation), for a fixed size 500000 records in emp_rschr relation and 250000 records in emp_paperpublish relation.

We found that the incremental materialize view maintenance using auxiliary relations wins for both types of changes, but it wins with a greater margin for the update generating changes. The refresh time is going down by 20% in figure 2(b).

Figure 2(c) and (d) plot the variation in elapsed time as the size of the emp_paperpublish relation (source relation) changes, for a fixed size of 50000 records in change set(delta relation).

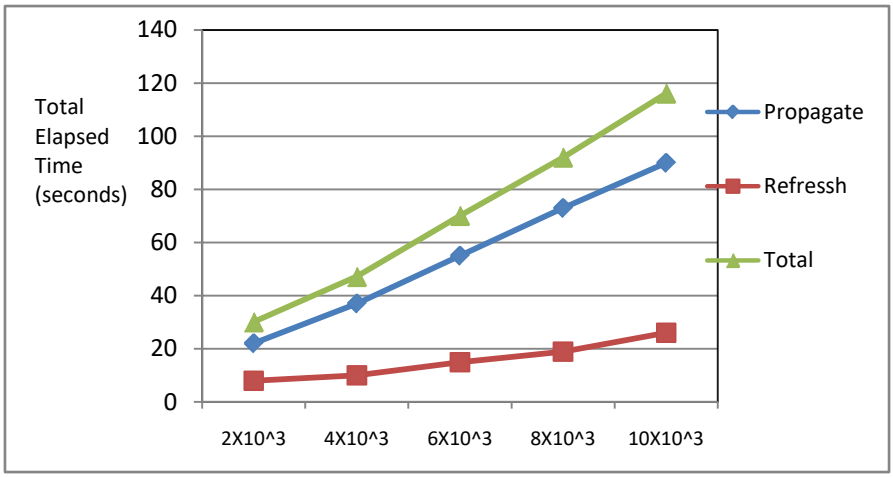

Figure 2 (a): Varying change set size for insert generating changes

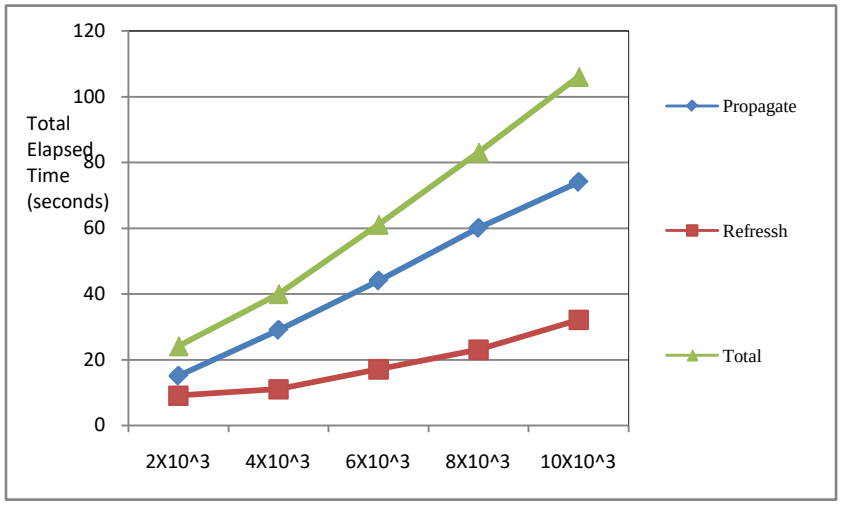

Figure 2(b): Varying change set size for update generating changes

VI. Conclusions

We have investigated one of the significant problems of a data warehouse, that is, materialized view maintenance and how to make warehouse materialized views self maintainable without accessing the data from underlying data sources. The study shows that it is possible to make warehouse views self maintainable by materializing additional auxiliary relations, which contain intermediate results, at a data warehouse site. Using efficient incremental materialize view maintenance technique it is possible to reduce the cost of view maintenance. Proposed materialize view maintenance technique using auxiliary relation and dividing the maintenance process into two steps: propagate and refresh require less maintenance time as compared to counting algorithm. Here the propagate function works implicitly and whenever the data warehouse is ideal the refresh function integrate the data into data warehouse views. The entire maintenance process is hidden from the data warehouse users.

References Références Referencias

1. A Segev and J. Park, "Maintaining Materialised Views in Distributed databases", In Proceedings of the IEEE International Conference on Data Engineering, 1989.
2. Segev and W. Fang," Currency based updates to distributed materialized Views", In proceedings of the IEEE International Conference on Data Engineering, 1990.
3. Abdulaziz S. Almazyad & Mohammad Khubeb Siddiqui," Incremental View Maintenance: An Algorithmic Approach", Internatioinal Journal of Electrical & Computer Sciences IJECS-IJENS Vol. 10, No. 03, 2009.
4. Bin Liu & Elke A. Rundensteiner, "Optimizing Cyclic Join View Maintenance over Distributed Data Sources", IEEE Transactions on Knowledge and Data Engineering, Vol. 18, No. 3, March 2006.
5. D. Agarwal, A. E. Abbadi, A. Singh and T. Yurek, "Efficient View Maintenance at Data Warehouses", Proc. ACM SIGMOD, pp. 417-427, 1997.
6. D. Lomet and J. Widom," Special Issue on Materialized Views and Data Warehousing", IEEE Data Engineering Bulletin 18(2), June 1995
7. E. N. Hanson, "A performance analysis of view materialization strategies", In SIGMOD pages 440-453, 1987.
8. GianLuca Moro and Claudio Sartori," Incremental View Maintenance on Multi-Source", In proceedings of IEEE, 2001.
9. Gray C. H. Yeung and William A. Gruver, "Multi agent Immediate Incremental View Maintenance for Data
10. Warehouses", IEEE Transaction on Systems, Man & Cybermetics- Part A: Systems & Human, Vol. 35, No. 2, March 2005.
11. Hao He, Junyi Xie., Jun Yang, Hai Yu," Asymmetric Batch Incremental View Maintenance", In the Proceedings of the 21st International Conference on Data Engineering, 1084-4627/05, 2005.
12. Heney E. Korth and Abraham Silberschatz, "Database System Concepts", McGraw Hill, 1986.
13. J. A. Blakeley, P.A. Larson and F. W. Tompa, "Efficient Updating Materialized Views", Proc. ACM SIGMOD, pp. 61-71, May 1986.
14. J. Chen, X. Zhang, S. Chen, K. Andreas and E. A. Rundensteiner, "DyDa: Data Warehouse Maintenance under Fully Concurrent Environments", Proc. ACM SIGMOD Demo Session, p.619, 2001.
15. J. Hammer, H. Garcia-Molina, J. Widom, W. Labio & Zhuge, "The Stanford Data Ware housing Project", IEEE Data Engineering Bulletin, June1995.
16. Jingren Zhou, PerAke Larson and Hicham G. Elmongui: Lazy Maintenance of Materialized Views",

in Proceddings of 33rd International conference on VLDB 2007, Vienna, Austria.

17. L. S. Colby, A. Kawaguchi, D. F. Lieuwen, I. S. Mumick and K. A. Ross, "Supporting Multiple View Maintenance Policies", In Proceeding ACM SIGMOD International Conference on Management of Data, 1977.
18. Latha S. Colby, Timothy Griffin, Leonid Libkin, Inderpal Singh Mumick and Howard Tricky, "Algorithms for Defered View Maintenance", In proceedings of ACM SIGMOD, 1996, Canada.
19. M. Adiba & B. Line\dsay, "Database Snapshots, "In Proceedings of the sixth International Conference on Very Large Databases, pages 86-91, Montreal, Canada, October 1980.
20. M. Mohnia, "Avoiding re-computation: View Adaptation in Data Warehouses", In Proc. Of 8th International Database Workshop, Hong Kong, pages 151-165, 1997.
21. N. Hyun, "Efficient View Self-Maintenance", Proceeding of ACM workshop on Materialized views: Techniques & Applications", Canada, June 7, 1996.
22. N. Roussopoulos, "An Incremental Access Method for Viewcache: Concept, Algorithms and Cost Analysis", ACM Trans. On Database Systems, 16(3):535-563, 1991.
23. O. Wolfson, H. M. Dewan, S. J. Stolfo and Yemini, "Incremental Evaluation of Rules & Its Relationship to Parallesim", In Proceedings ACM IGMOD, International Conference on Management of Data, pages 78-87, 1991.
24. R. Hull & G. Zhou, "A framework for supporting data integration using the materialized & virtual approaches", In SIGMOD Int'l Conference, Canada, June 4 -6,1996.
25. R. Ramakrishan, K. A. Ross, D. Srivastava and S. Sudarshan, "Efficient Incremental Evaluation of Queries with Aggregation", In International Logic Programming Symposium 1994.
26. S. Chaudhuri and U. Dayal, "An Overview of Data Warehousing and OLAP Technology", In ACM SIGMOD Record, volume26, pages 65-74, 1974.
27. S. Chen, B. Liu and E.A. Rundensteiner, "Multiversion Based View Maintenance over Distributed Data Sources", ACM Trans. Database Systems (TODS), vol.29, no. 4, pp. 675-709, 2004.
28. S. Solanki and Dr. Ajay Kumar, "A Comparative Study of Materialized View Maintenance Techniques in Data Warehousing", IJRIME, Vol. 1, Issue 2, August 2011.
29. S. Solanki and Dr. Ajay Kumar, "A Comprehensive Study of Data Warehousing", IJMR, Vol1, Issue 1, January 2012.

Global Journal of Computer Science and Technology: C
Software & Data Engineering
Volume 18 Issue 3 Version 1.0 Year 2018
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Clustering on Spatial Data Sets using Extended Linked Clustering Algorithms

By K. Lakshmaiah

Sir Vishveshwaraiah Institute of Science & Technology

Abstract- Various clustering algorithms (CA) have been reported in literature, to group data into clusters in diverse domains. Literature further reported that, these CAs work satisfactorily either on pure numerical data or on pure categorical data and perform poorly on mixed numerical and categorical data. Clustering is the process of creating distribution patterns and obtaining intrinsic correlations in large datasets by arranging the data into similarity classes. The present work pertains to reviewing the available research papers on clustering spatial data. In a web perspective, a detailed inspection of grouped patterns and their belonging to well known characters will be very useful for evolution of clusters. The review work is split into spatial data mining, clustering on spatial data sets and extended linked clustering. This review work will enable the researchers to make an in depth study of the till date research work on above areas and will pave way for developing extended linked clustering algorithms with a view to find number of clusters on mixed datasets to produce results for several datasets.

Keywords: *extended linked clustering, optimum criterion, coherence, databases, spatial datasets.*

GJCST-C Classification: *I.5.3*

Strictly as per the compliance and regulations of:

Clustering on Spatial Data Sets using Extended Linked Clustering Algorithms

K. Lakshmaiah

Abstract- Various clustering algorithms (CA) have been reported in literature, to group data into clusters in diverse domains. Literature further reported that, these CAs work satisfactorily either on pure numerical data or on pure categorical data and perform poorly on mixed numerical and categorical data. Clustering is the process of creating distribution patterns and obtaining intrinsic correlations in large datasets by arranging the data into similarity classes. The present work pertains to reviewing the available research papers on clustering spatial data. In a web perspective, a detailed inspection of grouped patterns and their belonging to well known characters will be very useful for evolution of clusters. The review work is split into spatial data mining, clustering on spatial data sets and extended linked clustering. This review work will enable the researchers to make an in depth study of the till date research work on above areas and will pave way for developing extended linked clustering algorithms with a view to find number of clusters on mixed datasets to produce results for several datasets. This work is likely to assist in deciding which clustering solution to use to obtain a coherent data solution for a particular character experiment. Further it could be used as an optimal tool to guide the clustering process towards better and character interpretable meaningful solutions. The major contribution of present work is to present an in depth literature review of research in the areas of Spatial data mining, clustering on spatial data sets and extended linked clustering with a view to assist researchers to develop optimum extended linked clustering and to develop optimum extended linked clustering algorithms for clustering process, towards better and character interpretable meaningful solutions.

Keywords: extended linked clustering, optimum criterion, coherence, databases, spatial datasets.

I. Introduction

Data mining (DM) is the process of effectively extracting data in the form of knowledge discovery, which provides useful and helping guide for information processing that can be utilized in varieties of applications [1]. Different types of DM techniques are augmented for application in the fields of science, research, medicine etc. Data bases (DB) comprise of terabytes, which are capable of storing huge mine of heterogeneous data. For effective DM pre mining and post mining procedures are prescribed. DM models consist of predictive models and descriptive models. AIS algorithm was the first association rule mining algorithm. The main limitation of this is that, it requires large space and is there by not efficient. Apriori algorithm is more efficient than AIS. Spatial data mining (SDM) is the process of arriving at potentially helpful patterns from large spatial data sets. Spatial data base can be best understood as the data relating to price range of the houses with nearby spatial features like beaches, different geographical regions in a city etc. Image spatial data base deals with image database dealing entirely with pictures or images [2]. The algorithms used for SDM include generalization – based methods. This is based on the concept of data from more than a few evidences from a concept level to its higher concept level and tracking knowledge from the widespread database. Collecting data and using it for knowledge discovery are two independent factors. It is often said that, we are data rich, but information poor. We require tools for automatic summarization of data, discovery of patterns in the data for analyzing and interpretation [3]. Many techniques are being used for data mining with Geographic Information Systems (GIS) for carrying out spatial data analysis of geographic data. The two approaches in common use are, first comes first learning on spatial data bases, where as the second is based on spatial statistics [4]. In SDM we must consider the spatial relations between objects. SDM is used in geo marketing, environmental studies, risk analysis etc. another equally important area in SDM is visual data mining (VDM) which applies visual human perception for mining large data sets. This mostly comprises of presenting huge data simplified to a graphical format [5], resulting in discovery of valuable patterns in very large data bases. In the fields of clustering and visualization pixel maps, a new way of displaying dense point sets are often used. However in SDM the issues and challenges involved need careful analysis [6]. The voluminous data and its handling and analysis poses a major challenge in SDM. Another challenge is extraction of implicit knowledge not explicitly stored in spatial data bases. The Introduction of computer capabilities and emergence of I.T, have led to enormous amount of data relating to science and engineering [7], which is normally made available through internet. This has led to transforming many areas from data poor to data rich stage.

Clustering algorithms (CA) are proposed by researchers in the fields DM [8]. The advantage is that, the clustered data is easy to understand and normally it does not confine to the shapes of the clusters. Various algorithms on clustering namely DBSCAN, VDBSCAN, DVBSCAN, ST-DBSCAN etc are proposed. In simple

Author: e-mail: klakshmaiah78@gmail.com

words clustering is grouping of objects or data into meaningful sub classes. Among various CA, density based algorithms are more efficient in identifying clusters with varied densities. However there exist constraints on application of CA in DM which will be presented in subsequent sections. Recent research in clustering includes, proposing Adaptive flocking algorithms for spatial clustering [9]. This works on use of new Swarm Intelligence (SI) techniques. Many authors have presented huge survey on application of above algorithms in clustered SDM [10]. Most of the clustered SDM algorithms were applied to clustered SDM in the fields of spatial cancer databases [11]. The various assumptions and requirements needed for applying the clustered SDM are [12] huge data to exist for applying algorithms, and the developed algorithms should be capable of handling irregular shapes, insensitive to bulky amounts of noise etc. Clustering algorithms are modified as two density based spatial clustering algorithms, especially when very huge large databases are to be handled [13]. The accomplishments and research needs of spatial data mining focus on location prediction, spatial outlier detection, co-location mining etc [14]. Researchers have also proposed and presented efficient and effective clustering methods for DM [15]. Novel methods based on delaunay triangulation were tried in the area of clustered SDM [16].

A recent development in mining spatial structural data mining is link mining (LM). This technique deals with mining richly structured data sets, where the objects are linked in some way [17]. The links include certain patterns, which could not be analyzed in traditional DM techniques. Web and hypertext mining, social networks, security and law enforcement data bases, bibliographic citations etc are best mined using linked clustered algorithms. LM is an emerging area and is an instance of multi-relational data mining. LM encompasses a range of tasks including descriptive and predictive modeling. With the introduction of link concepts, new issues such as number of links, types of links, inferring the existence of links etc arise. In the area of applying LM to web page DM, the algorithms used are based on citation relation between web pages.

Linked data base mining model of bibliographic description [18] is derived from ideas based on schema bib extended group and was found useful in above mining task. Though linked clustered SDM is an emerging area of research in DM, approaches to visualizing linked data is [19] assuming greater dimensions. These studies are mostly confined to semantic web community. The main issue involved is the lack of technical knowledge and an understanding of the semantic technology in the use of web data. This linked data base mining is an extended concept of linked clustered SDM and may be termed as Extended Linked Clustered SDM (ELC SDM). These applications result in further challenges in the areas of mobile devices and the reduction in cost of producing sensors. When a Uniform Resource Identifier (URI) is de referenced, a response is returned and is characterized by an extended hypertext markup linked clustered representation of the resource, managing the life cycle of linked data. Extended linked data with LOD2 stack forms the latest research in the field of extended linked clustered SDM [20]. The LOD2 stack is an integrated distribution of aligned tools, supporting the life cycle of linked data from extraction via enrichment, interlinking and fusing. Recent applications of ELCSDM are seen in educational linked data bases.

In the area of Web databases related to education, highlighting how such extended links form a globally addressable network of resources for education [21] is very important. Adopting the linked DM to extended linked DM needs a minor integration effort, to improve the global cohesion of education networks. DM techniques have now emerged to the extent of application of extended linked clustered SDM techniques, for arriving at meaningful and useful conclusions from the huge vast data base.

II. Literature Review

The literature review is considered from three aspects. First one deals with general methods of data mining. The second is on clustering spatial data sets and third is on linked clustering with Spatial reference to extended linked clustering. The detailed survey is given below.

Amitkumar patnaik etal[1], worked on different types of data mining techniques for powerful data mining ranging from commercial to scientific applications. Their studies included the areas of warehouse and online analytical processing, along with various data mining models. They have tried different data fields.

M. Hemaltha etal [2], made extensive survey on knowledge discovery in spatial data mining. Spatial data mining is the process of discovering, motivating and obtaining previously unknown, but potentially helpful patterns from large spatial datasets. Extracting interesting and useful patterns from spatial datasets is more tricky than extracting the parallel patterns from established numeric and definite data, due to the complexity of spatial data types, spatial relationships, and spatial auto correlations. They focused on the sole features, that distinguish spatial data mining from traditional data mining. Major activities and research needs in spatial data mining research were discussed by them. They listed the applications and techniques, issues and challenges on spatial data mining. They concluded that, spatial data mining is a promising field with rich research results and many challenging issues.

Neelamadhab padty etal [3], made extensive work on data mining applications and future scope, and published their work. They focussed on variety of techniques among different areas. Their work concentrated on MNC data collection and mining the data, collected from different places and countries. They have used data ware house, by improving the effectiveness of managerial decision making. They mainly worked on mining huge data. Their paper concentrated on, number of applications of data mining, which formed a basis for further research in this area.

Karine etal [4], made extensive survey of spatial data mining methods from available data bases and developed statistical point of views. Their work included review of data mining methods, combined with GIS for conducting spatial analysis of geographic data. Their conclusions included, listing of main differences between the two common approaches, namely first come first learning, and spatial statistics and also the elements they have, in common.

Danial etal [5], developed pixel based visual mining of geo spatial data. In many application domains, data is collected and referenced by geo-spatial location. Spatial data mining, or the discovery of interesting patterns in such databases, is an important capability in the development of database systems. A noteworthy trend includes increasing size of data sets in common use, such as records of business transactions, environmental data and census demographics. These data sets often contain millions of records, or even far more. This situation creates new challenges in coping with scale. For data mining of large data sets to be effective, it is also important to include humans in the data exploration process and combine their flexibility, creativity, and general knowledge with the enormous storage capacity and computational power of today's computers. Visual data mining applies human visual perception to the exploration of large data sets. Presenting data in an interactive and graphical form often fosters new insights, encouraging the formation and validation of new hypotheses to the end of better problem-solving and gaining deeper domain knowledge. They gave a short overview of visual data mining techniques especially for analyzing geo-spatial data. They provided examples for effective visualizations of geo-spatial data in important application areas such as consumer analysis and census demographics.

Krzysztof etal [6], did extensive work on spatial data mining and reviewed the progress made so far along with associated issues and challenges. Since huge amount of data exists in various applications, analysis of this huge data far exceeds human ability. Data mining is extended to spatial data bases. They have summerised recent works on spatial data mining from spatial data generalization to spatial data clustering, mining spatial association rules etc. They concluded that spatial data mining is a promising field, with fruitful research results and many challenging issues.

Jiawei etal [7], identified research challenges for data mining in science and engineering .With the advent of IT and CSE fields fast developing ,data is collected and stored in a massive scale. This data is made available globally through networks. This has led to, developing data rich data base, calling for new data intensive methods, to conduct research in science and engineering. Their work focused on issues including (1) information network analysis, (2) discovery, usage, and understanding of patterns and knowledge, (3) stream data mining, (4) mining moving object data, RFID data, and data from sensor networks, (5) spatio temporal and multimedia data mining, (6) mining text, Web, and other unstructured data, (7) data cube-oriented multidimensional online analytical mining, (8) visual data mining, and (9) data mining by integration of sophisticated scientific and engineering domain knowledge.

M. Parimala etal [8], made a survey on density based clustering algorithms for mining large spatial data bases. Density based clustering algorithm is one of the primary methods for clustering in data mining. The clusters which are formed, based on the density are easy to understand and it does not limit itself to the shapes of clusters. They gave a detailed survey of the existing density based algorithms namely DBSCAN, VDBSCAN, DVBSCAN, ST-DBSCAN and DBCLASD based on the essential parameters needed for good clustering algorithms. They analyzed the algorithms, in terms of the parameters essential for creating meaningful clusters.

Gianluigi etal [9], developed an adaptive flocking algorithm for spatial clustering. This algorithm was based on the use of new swarm intelligence techniques (SI). SI is a new emerging area where a problem can be solved by using a set of biologically inspired agents exhibiting a collective intelligent behaviour. They have applied this algorithm to two synthetic data sets and its performance was comparable with other algorithms.

Jiawei etal [10], brought out current status and future of frequent data mining, in their research paper. Frequent pattern mining has been a focused theme in data mining research for over a decade. Abundant literature has been dedicated to this research and tremendous progress has been made ranging from efficient and scalable algorithms for frequent item set mining in transaction databases, to numerous research frontiers, such as sequential pattern mining, structured pattern mining, correlation mining, associative classification, and frequent pattern-based clustering, as well as their broad applications. They provided a brief overview of the current status of frequent pattern mining and discussed a few promising research directions. They made it clear that, frequent pattern mining

research has substantially broadened the scope of data analysis and will have deep impact on data mining methodologies and applications in the long run. However, there are still some challenging research issues that need to be solved before frequent pattern mining can claim a corner stone approach in data mining applications.

Rituchauhan etal [11], worked on data clustering methods for discovering clusters in spatial cancer databases. They outlined data analyzing tools and data mining techniques to analyze medical data as well as spatial data. The spatial data base is formed by grouping the objects into clusters. Their study focused on discrete and continuous spatial study, focused on discrete and continuous database on which clustering techniques are applied to form clusters. Classical clustering and hierarchical clustering on the spatial data sets to generate efficient clusters formed their main work. Their experimental results were reported which exhibited certain facts that are evolved and cannot be otherwise retrieved from raw data.

Sundararajh etal [12], studied on spatial data clustering algorithms in data mining. Heavy and huge databases have produced interests in the area of data mining. Useful information can only be obtained after clustering the data. Through this process, the hidden patterns or useful subgroups can be identified. They used spatial clustering approach for investigations. They developed fast working and effective algorithms for extraction of information, trends etc from the database. They presented the essential features of clustering algorithms which include scalability, ability to recognize irregular shapes, insensitive to bulky noises etc. The major contribution of their research work was to help researchers to come up with needy techniques to cluster the spatial data effectively.

Xin etal [13], compared density based spatial clustering algorithms for large datasets. The two density methods chosen by them were, density based spatial clustering algorithms and density based clustering algorithms. The two methods are described in detail and a comparison of algorithms was made.

Shastistekar etal [14], worked on spatial data mining, which is the process of discovering interesting and previously unknown, but potentially useful patterns from large spatial datasets. They have identified the research needs of spatial data mining, in the areas of location, prediction, spatial outlier detection, co-location mining and clustering.

Rayman etal [15], developed effective clustering methods for spatial data mining. Spatial data mining is the discovery of interesting relationships and characteristics that may exist implicitly in spatial databases. They explored whether, clustering methods have a role to play in spatial data mining. To this end, they developed a new clustering method called CLAHANS, which is based on randomized search. They also developed two spatial data mining algorithms that use CLAHANS. Their analysis and experiments showed that, with the assistance of CLAHANS, these two algorithms are very effective and can lead to discoveries that are difficult to find with current spatial data mining algorithms. Furthermore, experiments conducted, to compare the performance of CLAHANS with that of existing clustering methods showed that, CLAHANS is the most efficient.

Insooking etal [16], used Delaunay triangulation for spatial data mining, with a view to discover significant pattern which may implicitly exist in huge data bases. They have used the SMTIN (spatial data mining by triangulated irregular network) method, which is based on Delaunay triangulation. Its advantages over the Previous ones were described, which include identification of sophisticated patterns and heirarchical structure of cluster distribution, knowledge of prior nature of distribution is not required, requirements of distribution is not required, requires less CPU processing time. It is not ordering sensitive and handles effectively outliers.

Lise etal [17], worked on key challenges on data mining. A key challenge for data mining is tacking the problem of mining richly structured datasets, where the objects are linked in some way. Links among the objects may demonstrate certain patterns, which can be helpful for many data mining tasks and are usually hard to capture with traditional statistical models. Recently there has been a surge of interest in this area, fueled largely by interest in web and hypertext mining, and also by interest in mining social networks, security and law enforcement data, bibliographic citations and epidemiological records.

Carol [18], worked on developing OCLCs linked data models of bibliographic description.

This document describes a proposed alignment between BIBFRAME and a model being explored by OCLC with extensions proposed by the Schema Bib Extend project, a W3C-sponsored community group tasked with enhancing Schema.org to the description of library resources. The key result is that the two efforts are complementary except for some common vocabulary required for the most important entities and relationships. The analysis presented was prompted by the call at the end of the December 2012 BIBFRAME Early Experimenters Meeting, for a set of Point or Position papers that worked out technical issues and made recommendations for a number of sketchy, difficult, or controversial aspects of the BIBFRAME model. The description was based on a small dataset presented in the entity in the Appendix, and the analysis was based on a larger dataset derived from the application of a mapping algorithm from MARC to BIBFRAME on all of World Cat.org. This draft is being released as an OCLC report, but it was intended to be read as a working paper for the BIBFRAME community.

Krzysztof etal [19], worked on visualizing linked data, and approaches for achieving the same. Their survey covered large, distributed and interlined networks of information fragments contained within desparate data sets as provided by unique data publishers. The data was published in a format which was machine readable. This data was linked to other external data. They presented a survey of existing approaches for handling web enabled linked data.

Soren etal [20], did extensive work on managing the life cycle of linked data with LOD2 stack. The LOD2 Stack is an integrated distribution of aligned tools which support the whole life cycle of Linked Data from extraction, authoring/creation via enrichment, interlinking, fusing to maintenance. The LOD2 Stack comprises new and substantially extended existing tools from the LOD2 project partners and third parties. The stack is designed to be versatile, for all functionality web cleared interfaces, which enables the plugging in of alternative third-party implementations. The architecture of the LOD2 Stack is based on three pillars: (1) Software integration and deployment using the Debian packaging system, (2) Use of a central SPARQL end point and standardized vocabularies for knowledge base access and integration between the divergent tools of the LOD2 Stack, and (3) Integration of the LOD2 Stack user interfaces based on REST enabled Web Applications. These three pillars comprise the methodological and technological framework for integrating the very heterogeneous LOD2 Stack components into a consistent framework. In their article they described these pillars in more detail and gave an overview of the individual LOD2 Stack components. The article also included a description of a real-world usage scenario in the publishing domain.

Matheu etal [21], worked on assessing the educational linked data land -scape. They presented a preliminary study of available web datasets related to education, providing an overview of this area and, more importantly, highlighting how such linked datasets form a globally addressable network of resources for education. As expected, a certain level of heterogeneity was found. They also showed how a minor integration effort can improve the global cohesion of such networks of educational web data.

III. Issues and Challenges

1. Since vast huge data base is to be handled, it is very difficult to identify the initial parameters like number of clusters, shape and density of clusters.
2. Since the shape of clusters may be in random manner, discovery of clusters among arbitrary shapes poses a challenge.
3. A good efficiency is to be achieved among large data bases.
4. The various algorithms like a) density based spatial clustered algorithms (DBSCN) ,varied density based spatial clustered algorithm (VDBSCAN) etc have to be carefully written to achieve meaningful end results.
5. Analysing linked clustered data is mostly restricted to web community. The lack of technical knowledge limits users in their ability to interpret and make use of webpage data.
6. Careful analysis is to be made before presenting linked data visualization.
7. The extended linked cluster SDM is an extended version of linked clustered SDM and is arrived at through minor modifications of linked clustered data.

IV. Scope and Objective of Present Work

Though DM is not a new concept and has been in use since a long time, with the advent of I.T enabled computer services, huge databases are created and handling this voluminous data has called for newer techniques, revised algorithms in the field of DM. This resulted in analyzing clustered linked spatial data and the latest trend is handling web based URL through extended linked clustered spatial data. The scope of present work is to make a review of the research work carried out in above areas and to develop extended linked clustered spatial data mining (ELCSBM).

V. Formulation of Problem

Voluminous data collected and stored is drawn from different geographical areas, having or not having similarities and links among them. Handling this data meaningfully and efficiently calls for a systematic analysis of data through spatial clustering the data, grouping the same as per the links present among them and extending these procedures ultimately for web enabled data though extended linked clustered SDM, has now become the center of research in the field of data mining. Hence the formulation of the problem.

VI. Present Work

The present work consists of making a detailed literature review on data mining with special reference to SDM, clustered SDM, linked clustered SDM along with a detailed understanding of the various algorithms used on issue basis. The work ends up with listing modified algorithms to be used for extended linked clustered spatial data mining (ELCSDM) operations. The various results are discussed and the important conclusions are listed.

VII. Algorithm used

The clusters formed based on density of database can be analyzed with ease, without confining to shapes of clusters. The various density based algorithms in use are DBSCAN, VDBSCAN, DVBSCAN,

ST DBSCAN, DBCLASD etc, and are briefly discussed here under.

a. Density based spatial cluster algorithms with website (DBSCAN) discovers clusters with arbitrary shapes with minimum number of input parameters such as radius of the clusters and minimum points required inside the cluster. The related algorithm consists of
   i. Selecting an arbitrary point.
   ii. Retrieving all points which are density reachable from the arbitrary point.
   iii. If the point is a core point, a cluster is formed.
   iv. If it is a border point the next point is considered.
   v. The process is continued till at the points are processed.

This algorithm requires only two input parameters and discovers clusters of arbitrary shapes. It holds good for large SDB.

b. Varied density based spatial clustering algorithms with noise (VDBSCAN) detects clusters with varied density, where the DBSCAN fails, and is capable of selecting several values of input parameters. The related algorithm consists of
   i. Calculating and storing K-dist for each project and partitioning the k-dist point.
   ii. Calculating the number of densities.
   iii. Selecting the parameter for each density.
   iv. Scanning the data for different densities.
   v. Displaying the valid cluster with respect to the corresponding density.

This algorithm helps in finding meaningful clusters having varied densities.

c. Density based algorithm for discovering density varied clusters in large spatial data bases (DVBSCAN) is a pioneer density based clustering algorithm which detects clusters with different shapes and sizes, but fails to detect clusters with varied densities that exist within the clusters.

This algorithm is capable of handling local density variations that exist within the clusters.

d. Distributed based clustering algorithm for mining large spatial data bases (DBCLASD) is a new clustering algorithm which is capable of detecting clusters with arbitrary shapes without calling for input parameters. The efficiency of this algorithm in handling huge database is satisfactory.

Link mining is a newly emerging research area in data mining and is mostly used in hypertext, web mining. It is a multi relational DM technique specializing analysis of links present in the spatial cluster data bases. Link mining does a range of tasks such as descriptive and predictive modeling. To perform these operations link mining requires new data mining algorithms dealing with predicting the number links, predicting type of link between two objects, finding co-reference and subgraph patterns. The algorithms that are commonly used for linked clustered spatial data mining (LCSDM) are given here under.

   i. Select hypertext and webpage classification, which has its roots in information retrieval community.
   ii. Define the features of the links to be searched for, in the web data base.
   iii. Obtain the links and their characteristics.
   iv. Identify the incoming and outgoing links.
   v. Label the category of the web page, based on the features of the link.
   vi. Use the link information such as anchor text and neighboring text around each link and obtain categorization results.

A modified approach based on extended linked clustered SDM (ELCSDM) to mine data present in hypertext and link mining combines techniques from inductive logic programming with statistical learning algorithm to construct features for related documents.

The algorithm for ELCSBM is presented here under.

   i. Instead of using words in a hypertext document, make use of anchor text, neighbouring text, capitalized words and alphanumeric words.
   ii. Using above, propose a combined model for text classification to form links and clusters.
   iii. Select the features of the links to be searched for, in the converted web data.
   iv. Define the features of the links in the web data.
   v. Identify the incoming and outgoing links.
   vi. Label the category of the webpage based on the features of the link.
   vii. Use the link information such as anchor text and neighboring text around each link and obtain categorization results.

A suitable machine language can be chosen and coding can be written, to run the same to visualize the results.

VIII. Results and Discussions

1. A detailed review of literature dealing with classical data mining, spatial data mining, clustered spatial data mining, linked clustered spatial data mining is presented in literature review.
2. The various algorithms used for DBSCAN, VDBSCAN, DVBSCAN, ST-DBSCAN, DBCLASD etc are briefly discussed and presented with their merits and demerits.
3. The algorithms used for identifying linked spatial DBM are discussed separately along with their merits and demerits, to mine useful information from the web based huge data bases.
4. The latest trend in DM relating to web data is, extended/modified linked clustered spatial data

mining (E/MLCSDM). The related algorithms for this technique are also presented.

5. The advantage of this modified/extended algorithms is that, more meaningful correlations and results can be obtained using these extended algorithms.

IX. Conclusions

The major contribution of present work is to review and understand the "as on today research status" on DM starting from classical DM to E/MLCSDM. The proposed algorithms for E/MLCSDM can be altered based on issues , and suitable coding can be written, which when run ,gives useful and meaning full results.

References Références Referencias

1. Amit Kumar Patnaik, "Data Mining and Its Current Research Directions", Information Sciences, Volume 181,Issue 7, 1 April 2011, Pages 1264-1284
2. Dr. M. Hemalatha, "A Recent Survey on Knowledge Discovery in Spatial Data Mining", International Journal of Computer Science Issues, Vol. 8, Issue 3, No. 2, May 2011.
3. Neelamadhab Padhy, "The Survey of Data Mining Applications And Feature Scope", International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.3, June 2012.
4. Karine Zeitouni, "A Survey of Spatial Data Mining Methods Databases and Statistics Point of Views", PRiSM Laboratory-University of Versailles 45, avenue des Etats-Unis - F-78 035 Versailles Cedex.
5. Daniel A. Keim, "Pixel Based Visual Mining of Geo-Spatial Data", AT&T Shannon Laboratory, 180 Park Avenue, Florham Park, NJ 07932-0971, USA.
6. Krzzsztof, " spadial data mining : progress and challenges survey paper", school of computer science,simon fraser university, Burnaby B.C., Canada.
7. Jiawei Han, "Research Challenges for Data Mining in Science and Engineering", University of Illinois at Urbana-Champaign.
8. M. Parimala, "A Survey on Density Based Clustering Algorithms for Mining Large Spatial Databases", International Journal of Advanced Science and Technology Vol. 31, June, 2011.
9. Gianluigi Folino, "An Adaptive Flocking Algorithm for Spatial Clustering", Via Pietro Bucci cubo 41C c/o DEIS, UNICAL, 87036 Rende (CS), Italy.
10. Jiawei Han, "Spatial Clustering methods in data minings: A survey", Burnaby, BC. Canada V5A 1s6.
11. Ritu Chauhan, "Data Clustering Method for Discovering Clusters in Spatial Cancer Databases", International Journal of Computer Applications (0975 – 8887) Volume 10– No.6, November 2010.
12. Sundararajan S, "A Study on Spatial Data Clustering Algorithms In Data Mining", International Journal of Engineering And Computer Science Volume1 Issue 1 Oct 2012 Page No. 37-41.
13. Xin Wang, "A Comparative Study of Two Density-Based Spatial Clustering Algorithms for Very Large Datasets", Department of Computer Science, University of Regina, Regina, SK, Canada S4S 0A2.
14. Shashi Shekhar, "Spatial Data Mining", Department of Computer Science and Engineering, University of Minnesota 4-192, 200 Union ST SE, Minneapolis, MN 55455.
15. Raymond T. Ng, "Efficient and Effective Clustering Methods for Spatial Data Mining", Proceedings of the 20th VLDB Conference Santiago, Chile, 1994.
16. In-So Kang, "A Spatial Data Mining Method by Delaunay Triangulation", Kumjeong-Gu, Jangjeon-Dong, Pusan, Korea 609-735.
17. Lise Getoor, "Link Mining: A New Data Mining Challenge", Dept. of Computer Science/UMIACS University of Maryland College Park, MD 20742.
18. Carol Jean Godby, "The Relationship between BIBFRAME and OCLC's Linked-Data Model of Bibliographic Description: A Working Paper", Senior Research Scientist OCLC Research.
19. Aba-Sah Dadzie, "Approaches to Visualising Linked Data: A Survey", Knowledge Media Institute, The Open University, Milton Keynes, United Kingdom.
20. Soren Auer, "Managing the Life-Cycle of Linked Data with the LOD2 Stack", LOD2 Project ?, c/o Universitat Leipzig, Postfach 100920, 04009 Leipzig, Germany.
21. Mathieu d'Aquin, "Assessing the Educational Linked Data Landscape", Paris, France, ACM 978-1-4503-1889-1. Web Sci'13, May 1 – May 5, 2013.

This page is intentionally left blank

Global Journal of Computer Science and Technology: C
Software & Data Engineering
Volume 18 Issue 3 Version 1.0 Year 2018
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

A Review of Real World Big Data Processing Structure: Problems and Solutions

By Khalid Imtiaz & M. Junaid Arshad

University of Engineering and Technology

Abstract- Information sort and sum in human culture is developing in astonishing pace which is brought about by rising new administrations as distributed computing, web of things and area-based administrations, the time of enormous information has arrived. As information, has been principal asset, how to oversee and use enormous information better has pulled in much consideration. Particularly, with the advancement of web of things, how to handling huge sum continuous information has turned into an extraordinary test in research and applications. As of late, distributed computing innovation has pulled in much consideration with elite, yet how to utilize distributed computing innovation for substantial scale ongoing information preparing has not been contemplated. This paper concentrated the difficulties of huge information firstly and finishes up every one of these difficulties into six issues. Keeping in mind the end goal to enhance the execution of constant handling of substantial information, this paper manufactures a sort of real-time big data processing (RTDP) design considering the distributed computing innovation and after that proposed the four layers of the engineering, and various leveled figuring model.

Keywords: big data; cloud computing; data stream; hardware software co-design; CEP; big data analytics as a service; big data cloud architecture.

GJCST-C Classification: H.3.m

Strictly as per the compliance and regulations of:

A Review of Real World Big Data Processing Structure: Problems and Solutions

Khalid Imtiaz [α] & M. Junaid Arshad [σ]

Abstract- Information sort and sum in human culture is developing in astonishing pace which is brought about by rising new administrations as distributed computing, web of things and area-based administrations, the time of enormous information has arrived. As information, has been principal asset, how to oversee and use enormous information better has pulled in much consideration. Particularly, with the advancement of web of things, how to handling huge sum continuous information has turned into an extraordinary test in research and applications. As of late, distributed computing innovation has pulled in much consideration with elite, yet how to utilize distributed computing innovation for substantial scale ongoing information preparing has not been contemplated. This paper concentrated the difficulties of huge information firstly and finishes up every one of these difficulties into six issues. Keeping in mind the end goal to enhance the execution of constant handling of substantial information, this paper manufactures a sort of real-time big data processing (RTDP) design considering the distributed computing innovation and after that proposed the four layers of the engineering, and various leveled figuring model. This paper proposed a multi-level stockpiling model and the LMA-based application organization technique to meet the continuous and heterogeneity necessities of RTDP framework. We utilize DSMS, CEP, group-based Map Reduce and other handling mode and FPGA, GPU, CPU, ASIC advancements contrastingly to preparing the information at the terminal of information gathering. We organized the information and afterward transfer to the cloud server and Map Reduce the information consolidated with the effective processing abilities cloud design. This paper brings up the general structure for future RTDP framework and computation techniques, is right now the general strategy RTDP framework outline.

Keywords: big data; cloud computing; data stream; hardware software co-design; CEP; big data analytics as a service; big data cloud architecture.

I. Introduction

With the advancement of web of things, different substantial scale continuous information preparing in view of ongoing sensor information are turning into the key of the development of EPC (epcglobal arrange) application at present. The scholarly community, the industry and even the administration establishment have as of now gave careful consideration to enormous information issues and created a distinct fascination.

In May2011, the world-renowned consulting firm McKinley released a detailed report on big data Big data: The next frontier for innovation, competition, and productivity[1], sparking a broad discussion of Big Data. The report gave a detailed analysis of the impact of big data, key technology and application areas. In January 2012, Davos World Economic Forum released a report entitled "Big Data, Big Impact: New Possibilities for Corresponding author e-mail: *pwang@ss.pku.edu.cn* International Development" [2], raising a research boom of Big Data. The report explored how to make a better use of the data to generate good social profits in the new data generation mode, and focused on the integration and utilization of mobile data produced by individual and other data. In March, the U.S. government released "Big Data Research and Development Initiative" [3] to put there search of Big Data on the agenda, and officially launched the "Big Data development plan".

In this manner, the examination of ongoing huge information has incredible application prospect and research esteem. Because of the continuous and the expansive size of information handling and different elements that ongoing enormous information requires make the review for constant huge information preparing testing, for the most part progressively, strength and extensive scale and so on.

Real time: Real-time processing of big data mainly focuses on electricity, energy, smart city, intelligent transportation, and intelligent medical fields. During the information processing it needs to be able to make quick decisions, and feedback relevant instructions to the sensing terminal input within a very short time delay. For instance, in Fire monitoring and rescue system, its processing center needs to be able to analyze and process the data collected by sensors in the site of the incident in a very short period, to make integrated decision by comprehensively considering site information such as the movement of persons and the site form and meantime to issue the corresponding instructions to the site sensing terminals, such as what extinguishing agent used for rescue, how to protect people's safety in the site of the incident and how to help the firemen to rescue. At the same time, the information gathered by sensing terminals and instruction information must arrive information gathering or processing terminal in real time and make relevant decisions. The loss caused by that decision-making information can't be conveyed Stability: In real time is also incalculable, so real-time processing of large data is particularly important. the areas covered by Real-time

Author α σ: University of Engineering and Technology, Lahore Campus Pakistan, Computer Science & Engineering Department.
e-mails: engr.khalidimtiaz@gmail.com, mjunaiduet@gmail.com

processing of big data are mostly closely related to the people such as Smart City 's intelligent transportation systems and high-speed train control system, and mostly highly associated with infrastructure which also determines the real-time data processing system in a large system architecture, hardware and software equipment and other aspects must possess high stability.

Large-scale: As talked about above, continuous huge information preparing frameworks are frequently firmly identified with urban foundation and real national application, so its application is regularly a tremendous scale. For example, shrewd city astute transportation framework, once the biggest ongoing information examination and choice are made, it frequently goes for transportation basic leadership at a commonplace and city level or even a national level, and significantly affects the national life.

Thus, this paper highlights the big data processing architecture under the cloud computing platform. It presents a data storage solution on heterogeneous platforms in real-time big data processing system, constructs a calculation mode for big data processing and points out the general framework for real-time big data processing which provides the basis for the RTDP (Real-Time Data Processing). Section 2 in this paper gives an overview of big data; section 3 discusses the differences and challenges between big data and real-time big data; section 4 points out the current deficiencies of cloud computing technology; section 5 combine cloud computing technology with the feature of real time big data to architect the processing platform for real-time big data; section 6 gives a demo about how the RTDP system is used in smart grid system and finally summarize this article.

II. Big Data Overview

Big data itself is a relatively abstract concept, so far there is not a clear and uniform definition. Many scholars, organizational structure and research institutes gave out their own definition of big data [18] [19] [20]. Currently the definition for large data is difficult to reach a full consensus, the paper references Academician Li Guojies definition for big data: in general sense, Big Data refers to a data collection that cant be obtained within a tolerable time by using traditional IT technology, hardware and software tools for their perception, acquisition, management, processing and service[21]. Real-time data is a big data that is generated in real time and requires real-time processing. Per the definition of Big Data, Big Data is characterized by volume, velocity and variety where traditional data processing methods and tools cannot be qualified. Volume means a very large amount of data, particularly in data storage and computation. By 2010 the global amount of information would rapidly upto 988 billion GB[22].Experts predict that by 2020 annual data will increase 43 times. Velocity means the speed of data grow this increasing, mean while people's requirements for data storage and processing speed are also rising. Purely in scientific research, annual volume of new data accumulated by the Large Hadron Collider is about 15PB [23]. In the field of electronic commerce, Wal-Mart's sells everyday more than 267 million(267Million)products[24].

Data processing requires faster speed, and in many areas data have been requested to carry out in real-time processing such as disaster prediction and rapid disaster rehabilitation under certain conditions need quickly quantify on the extent of the disaster, the regional scope impacted etc. Variety refers to the data that contains structured data table, semi-structured and unstructured text, video, images and other information, and the interaction between data is very frequent and widespread. It specifically includes diverse data sources, various data types, and a strong correlation between the data.

With the advancement of PC and system innovation, and additionally astute frameworks is regular utilized as a part of present day life, enormous information has turned out to be progressively near individuals' everyday lives. In 2008, Big Data issue released by "Nature" pointed out the importance of big data in biology, and it was necessary to build biological big data system to solve complex biological data structure problem [25]. Paper [25] pointed out that the new big data system must be able to tolerate various structures of data and unstructured data, has flexible operability and must ensure data reusability. Furthermore, Big Data plays an important role in the defense of national network digital security, maintaining social stability and promoting sustainable economic and social development [26]. With the development of big data technology, Big Data also plays an important role in creating a smart city, and has important applications in urban planning, intelligent traffic management, monitoring public opinion, safety protection and many other fields [27].

III. Difference and Challenges between Big Data and Real-Time Big Data

Huge information is trademark by multi-source heterogeneous information, broadly circulated, dynamic development, and "information mode after the data"[28] [29]. Notwithstanding having every one of the qualities with huge information, constant enormous information has its own attributes. Contrasted and the huge information, with regards to information reconciliation ongoing enormous information has higher prerequisites in information procurement gadgets, information examination devices, information security, and different angles. The accompanying presents from information

incorporation, information investigation, information security, information administration and benchmarking.

a) Data Collect

With the improvement of web of things [30] and Cyber Physical System (CPS) [31], the ongoing of information handling requires ever more elevated. Under the enormous information condition, various sensors and portable terminals scatter in various information administration frame work which makes information accumulation itself an issue. In RTDP framework, its ongoing information accumulation confronted makes information mix confronting many difficulties.

i. *Extensive heterogeneity*

In huge information framework, the information created by versatile terminals, tablet PCs, UPS and different terminals is frequently put away in reserve, yet in RTDP framework it requires information synchronization which conveys huge difficulties to the remote system transmission. When managing handling heterogeneity, huge information framework can utilize NoSQL innovation and other new stockpiling techniques, for example, Hadoop HDFS. However, the constant requires low in this sort of capacity innovation, where the information is frequently put away once yet read commonly. However, this sort of capacity innovation is a long way from fulfilling the necessity of ongoing enormous information framework that requires information synchronization. Because of broad heterogeneity of enormous information, information transformation must be done amid information incorporations preparing, however conventional information distribution center has clearly deficient to address the issues of time and scale that huge information requires [32] [33] [34].

ii. *Data quality protection*

In the time of huge information, it is a marvel frequently creates the impression that valuable data is being submerged in a substantial number of futile data [6]. The information nature of Big Data has two issues: how to oversee substantial scale information and how to wash it. Amid the cleaning procedure, if the cleaning granularity is too little, it is anything but difficult to sift through the valuable data; if the cleaning granularity is excessively coarse, it can't accomplish the genuine cleaning impact. So, between the amount and quality it requires cautious thought and measured which is more apparent progressively enormous information framework. From one perspective, it obliges framework to synchronize information in a brief span; then again, it additionally requires the framework to make a fast reaction to information progressively. The execution necessities of the speed of information transmission and information investigation are expanding. In addition, the information might be separated at once hub may get to be distinctly basic post handling information. Consequently, how to get a handle on the connection amongst information and precisely decide the convenience and viability of information turns into a genuine test.

b) Data Analytics

Information examination is certainly not another issue. Customary information examination is principally propelled for organized information source, and right now has a total and successful framework. On the premise of the information distribution center, it assembles an information solid shape for online logical preparing (OLAP). Information mining innovation makes it conceivable to discover further learning from a lot of information. Be that as it may, with the entry of the time of enormous information, the volume of various semi-organized and unstructured information quickly develops, which conveys gigantic effect and difficulties to the customary examination strategies and existing procedures are no longer relevant. It for the most part reflects in convenience and file outline under element condition.

i. *Timeliness of information preparing*

In the period of huge information, time is esteem. As time passes by, the estimation of information contained in the information is too weakening. Progressively information frameworks, time is required higher. For instance, in an information preparing of debacle examination, continuous fast prepares, airplane and other high opportuneness execution gadget, time has gone past monetary esteem. Harms brought about by absurd postponement would be difficult to assess. The period of constant huge information proposes another and higher necessity to the courses of events of information handling, for the most part in the choice and change of information preparing mode. Continuous information handling modes fundamentally incorporates three modes: gushing mode, clump mode and a mix of two-a blended handling mode. Albeit as of now numerous researchers have made an incredible commitment to continuous information preparing mode, yet there is no regular structure for constant handling of expansive information.

ii. *Record plan under element condition*

The information design in the period of enormous information might change always as information volume shifts and existing social database file is no longer relevant. Step by step instructions to plan a basic, proficient and ready to rapidly make an adjustment has turned into a one of the real difficulties of enormous information preparing when information mode changes. Current arrangement is essentially fabricated a list by NoSQL databases to take care of this issue, yet they have been notable take care of the demand for constant handling of enormous information.

iii. *Absence of earlier learning*

From one viewpoint, since semi-organized and unstructured information proliferate, it is hard to construct its inside formal relations while dissecting the information; On the other hand, it is troublesome for this information should have been handled continuously to have adequate time to set up from the earlier learning because of the happening to the information stream in the type of a perpetual stream.

c) Data Security

Data privacy issues associated with the advent of computers has been in existence. In the era of big data, the Internet makes it easier to produce and disseminate data, which makes data privacy problems get worse, especially in real-time processing of large data. On the one hand, it requires data transmission real-time synchronization; on the other hand, it demands strict protection for data privacy, which both raise new demands to system architecture and computing power.

i. *Expose hidden data*

With the appearance of the Internet, especially the appearance of social networks, people are increasingly used to leave data footprints. Through data extraction and integration technology, accumulate and associate these data footprints may cause privacy exposure. In real-time big data processing, how to ensure the speed of processing a data as well as data security is a key issue which has troubled many scholars. Data disclosure conflicts with privacy protection by hiding data to protect privacy it will lose the value of data; thus it is essential to public data. Especially by digging accumulated real-time large-scale data, we can draw a lot of useful information, which has a great value. How to ensure the balance between data privacy and data publicly is currently in research and application a difficulty and hot issue. Therefore, the data privacy in the era of big data is mainly reflected in digging data under the premise of not exposing sensitive information of the user. Paper [35] proposed privacy preserving data mining concept, and many scholars have started to focus on research in this area. However, there are is a conflict between the amount of information and the privacy of data, and that's why so far it has not yet a good solution. A new differential privacy method proposed by Dwork may be a way to solve the protection of data privacy in big data, but this technology is still far from practical applications [36].

d) Usability issue of data management

Its difficulties for the most part reflect in two viewpoints: gigantic information volume, complex examination, different outcome shapes; various businesses required by huge information. However, examination specialist's absence of learning of both perspectives generally. Accordingly, the ease of use of constant huge information administration principally reflects in simple to find, simple to learn and simple to utilize [37]. Along these lines with a specific end goal to accomplish ease of use of enormous information administration, there are three essential standards to be minded as takes after:

i. *Visibility*

Deceivability requires the utilization of the information and the outcomes be indicated unmistakably in an exceptionally instinctive manner. The most effective method to accomplish more techniques for vast information handling and instruments rearrangements and mechanization will be a noteworthy test later on. Ultra-substantial scale information representation itself is an issue, while constant perception of huge scale information will spend a considerable measure of registering assets and GPU assets. Subsequently how to upgrade the execution and use of the GPU is an intense test.

ii. *Mapping*

Instructions to coordinate another huge information preparing strategy to handling strategies and techniques individuals have turned out to be usual to and accomplish quick writing computer programs is an extraordinary test to information ease of use later on. For Map Reduce needs SQL-like standard dialect, the scientists built up a more elevated amount dialects and frameworks. Run of the mill agents are the Hadoop Hive SQL [32] and Pig Latin [38], Google's Sawzall [39], Microsoft's SCOPE [40] and DryadLINQ [41] and also MRQL [42], and so on. Be that as it may, how to apply these dialects and frameworks to ongoing enormous information preparing still stay huge difficulties.

iii. *Feedback*

Criticism configuration permits individuals to monitor their working procedures. Works about this perspective is few in Big Information field [43] [44] [45]. In the period of huge information, the inner structure of many apparatuses is exceptionally mind boggling. Also, in programming investigating it is like Black Box troubleshooting for the typical clients and the strategy is unpredictable and additionally need of input. On the off chance that later on human-PC collaboration innovation can be presented in the weight of huge information, individuals can be all the more completely required in the entire investigation prepare, which will successfully enhance the client's criticism sense and extraordinarily enhance the usability.

An outline meets the over three standards will have the capacity to have a decent convenience. Perception, human-PC collaboration and information because systems can successfully improve ease of use. Behind these innovations, gigantic metadata administration needs our unique consideration [46]. So how to accomplish a proficient administration of the enormous metadata in a huge scale capacity framework

will have an import effect on the ease of use of ongoing huge information.

e) Test benchmark of performance

Advantages A critical perspective for enormous information administration is the quality affirmation, particularly for ongoing administration of extensive information as catastrophe created by information mistake will be intense and even limitless. The initial phase in quality affirmation is to do execution testing. There is not yet a test benchmark for the administration of enormous information. Principle challenges confronted by building huge information benchmarks are as followings [47]:

i. *High many-sided quality of framework*

Constant enormous information is exceedingly heterogeneous in information design and in addition equipment and programming and it is hard to model every single huge dat items with a uniform model. Continuous enormous information framework requires high opportuneness which makes it difficult to remove a delegate client conduct progressively. What's more, information size is substantial and information is extremely hard to imitate which both make the test more troublesome.

ii. *Rapid upset of framework*

The customary social database framework design is moderately steady, however the information continuously enormous information preparing is in a consistent condition of development, and there is a sure relationship be tween's the information, which makes the benchmark test comes about got soon not mirror the present framework real execution. Continuously enormous information framework test results are required to be finished inside a brief timeframe delay with high exactness, which in the equipment and programming angles is a genuine test to the test benchmark.

Reconstruct or reuse existing test benchmark Extend and reuse on the current benchmarks will significantly decrease the workload of building another vast information test benchmark. Potential applicant's principles are SWIM (Measurable Workload Injector for Map Reduce) [48], MRBS [49], Hadoop possess Grid Mix [50], TPC-DS [51], YCSB++ [52], and so on. In any case, these benchmarks are no longer relevant continuously enormous information preparing. Presently there are as of now some explores concentrating on the development of huge pieces of information test benchmark, yet there is additionally a view which thinks its untimely to talk about that at present. By following and investigating the heaps of seven items which are connected with Map Reduce innovation, Chen et al [47] [53] think it is difficult to decide ordinary client situations in the period of huge information. When all is said in done, building huge information and constant huge information test benchmark is vital. In any case, the difficulties it will face are a considerable measure, and it is extremely hard to construct a perceived testing models like TPC.

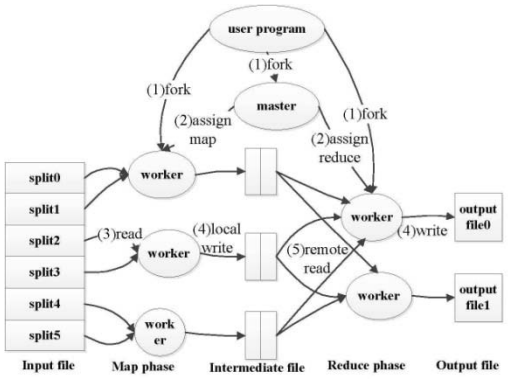

Fig. 1: The framework of the Map Reduce Model

IV. Shortcomings of Cloud Computing Architecture

a) Cloud Computing Overview

Cloud computing is the product of the traditional computer technology and network technology development integration such as grid computing, distributed computing, parallel computing, utility computing, network storage, virtualization, load balancing, etc., which aims at integrating multiple relative low-cost computing entities into one perfect system with powerful computing ability via the network and with the help of the SaaS, PaaS, IaaS, MSP and other advanced business models distributing this powerful computing ability to the hands of the end user.

b) Shortcomings of cloud computing architecture

MapReduce model is simple, and in reality, many problems can be represented with MapReduce model. Thus MapReduce model has a high value as well as many application scenarios. But MapReduces achievement is mainly relying on the Hadoop framework while the data processing method of Hadoop is" Store first post-processing" which is not applicable to real-time large data processing. Though currently there are some improved algorithms able to make Hadoop-based architecture almost real-time, for example, some latest technology like Cloudera Impala is trying to solve problems of processing real-time big data on Hadoop the batch processing of Hadoop and its structural features make Hadoop defective in processing big data in real time. Hadoops defect in real-time big data processing mainly reflects in data processing modes and application deployment. This paper will discuss these two aspects separately in the following.

Big data processing mode can be divided into stream processing and batch processing. The former is store-then-process, and the latter is straight-through-processing. In stream processing, the value of data reduces as time goes by which demanding real-time; in batch processing, data firstly is stored and then can be

processed online and offline [46]. MapReduce is the most representative of the batch processing method.

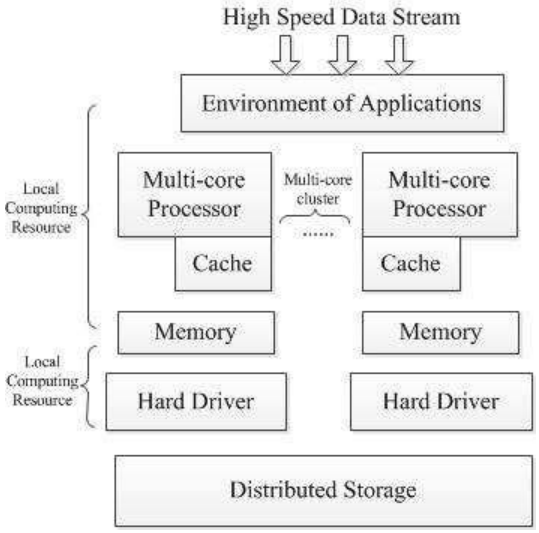

Fig. 2: Supporting Environment

c) Other architectures

Except the Hadoop-based real-time big data processing architecture, researchers already design an architecture to deal with a streaming data based on the way to process the stream-oriented data, noticing that batch processing in Hadoop can't meet the feature of real-time big streaming data. For example, Twitters Storm processing mode, Apaches Spark and LinkedIn In-stream.

Spark, as an advanced version of Hadoop, is a cluster distributed computing system that aims to make super-big data collection analytics fast. As the third generation product of Hadoop, Spark stores the middle results with internal storage instead of HDFS, improving Hadoops performance to some extent with a higher cost. Resilient Distributed Dataset, RDD, is an abstract use of distributed memory as well as the most fundamental abstract of Spark, achieving operating the local collection to operate the abstract of a distributed data set. Spark provides multiple types operations of data set which is called Transformations.

Storm cluster has some similarity with Hadoop. The difference is that its Job in MapReduce running in Hadoop cluster and Topology in Strom. Topology is the highest-level abstract in Storm. Every work process executes a sub-set of a Topology, which consists of multiple Workers running in several machines. But naturally the two frameworks are different. Job in MapReduce is a short-time task and dies with the tasks ending but Topology is a process waiting for a task and it will run all the time as system running unless is killed explicitly. In Storm cluster, it also has Master node and Worker node.

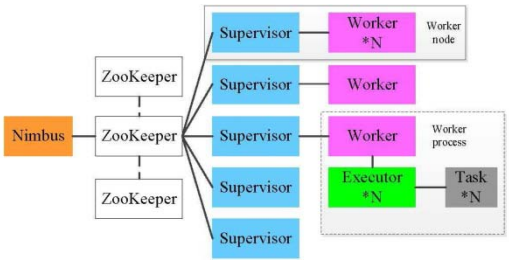

Fig. 3: Physical Architecture of Storm

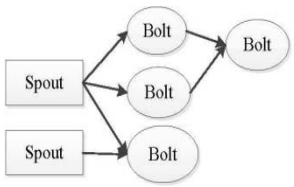

Fig. 4: Physical Architecture of Storm

V. Real-Time Big Data Processing Framework

In addition to powerful computing ability, real-time big data processing system must have strong timeliness which means it must quickly respond to the request from system terminals in a very short time delay. So at first, real-time big data processing system must have powerful computing ability for big data. A traditional method to process big data is to rely on the powerful computing capabilities of the cloud computing platform to achieve, while for the timeliness it must rely on the ability of the rapid data exchange between system's internal and nodes.

RTDP (Real-Time Data Processing) framework into four layers–Data, Analytics, Integration and Decision from a functional level. Shown in Figure 5.

a) Data

This layer mainly charges for data collection and storage, but also including data cleaning and some simple data analysis, preparing data for Analytics. At the terminal of data collection, it needs to manage all terminals. For example, the FPGA commonly used in Data Stream Management System, DSMS; the ASIC used in Complex Event Processing, CEP; and CPU and GPU (Graphic Processing Unit) in batch processing system represented by MapReduce. Data storage module is responsible for the management of large-scale storage systems. Thanks to the heterogeneity of real-time data sources and the large data processing platform, RTDP systems can handle data from various data sources, including Hadoop for unstructured storage, the data warehouse system for structured storage and analysis, SQL databases, and some other data source system.

b) Analytics

This layer is the core of RTDP system and the critical layer to determine the performance of RTDP system. This layer is mainly responsible for data structure modelling, data cleansing and other data analysis processing, preparing data for the algorithm integration layer.

c) Integration

d) Decision

This layer makes decisions with the results of data analysis which is the highest layer of data processing system as well as the ultimate goal of data analysis process. RTDP is a procedure involving numerous tools and systems interact with each other iteratively. At every level, the definition of" Big data" and" Real time" is not immutable. They have their own unique meaning at every level due to the functional association at each level. The four layers will be general process of RTDP in the future as well as the basic framework of the RTDP in this paper. Here we are going to discuss each layer in detail from the functionality, processing methods, related tools and deployment aspects of the system.

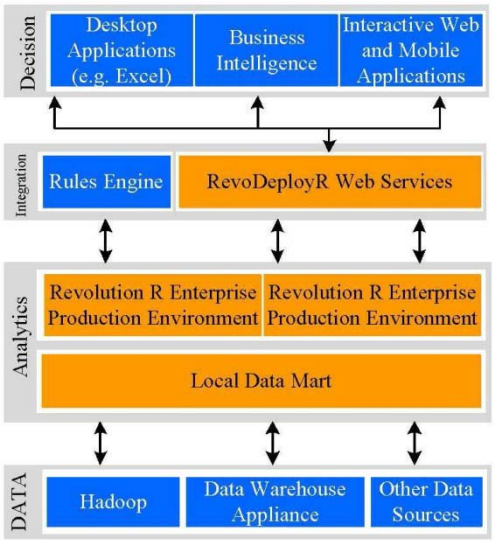

Fig. 5: Architecture of Real-Time Data Processing

e) Data Layer

Since the data collected by sensors is rough and messy, and original data often contain too much useless data, modeling and data analysis for the tremendous difficulties, so the data collection process must be preliminary data analys is and filtering. First need to extract the data features, integrated data sources, extraction points of interest, select the characteristic function to determine the data formats and extract useful information from data marts, and several steps in which the data feature extraction for unstructured text data, etc. of data is very important, therefore, makes the feature extraction for data collection and storage is an important part of the process.

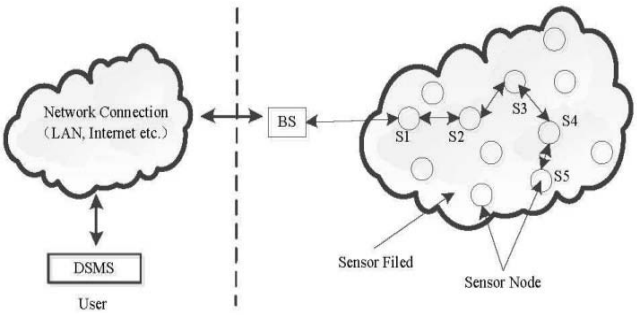

Fig. 6: Adaptive Task Allocation of Data Stream Management System

f) Data Collection and Data Base

RTDP systems heterogeneous platforms and performance makes RTDP system's data source contains a variety of ways, according to the data processing mode can be roughly divided into the CEP, DSMS, DBMS, based on a variety of ways such as MapReduce batch for each treatment have their different data acquisition techniques, such as remote medical field for surgical treatment of complex event processing scenarios for data acquisition ASIC, decoding audio and video coding in an FPGA, etc. Thus, during the data collection and management there are certain rules that must be collected on the side of the device identification, and can be based on different device programming overhead deployment and management nodes.

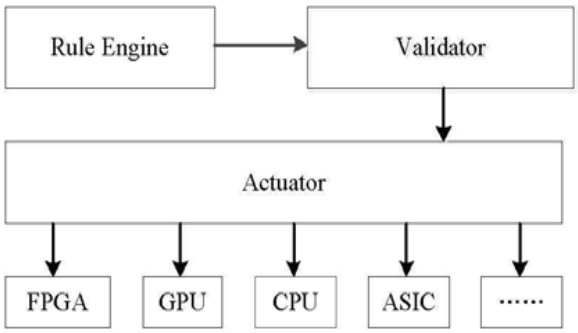

Fig. 7: Structure of Logical Management Adapter

g) Data Storage and Cleaning

In RTDP, data comes from a wide range of sources, unstructured and structured data mixed, so Hadoop and other unstructured storage system in RTDP system has a natural advantage, but Hadoop itself does not achieve full real-time requirements, which determines our in real-time using Hadoop big data storage process Hadoop first need to solve real-time problems in the framework of the proposed RTDP use of multi-level storage architecture to solve the problem, its architecture is shown in Figure 8.

In RTDP multi-level storage system data through a lot of the local server first preliminary processing, and then uploaded to the cloud server for in-depth analysis and processing. Such architectural approach to solve the data filtering is how to determine the relevance of the issue of data is an important means, Since the real time processing of large data nodes need to collect data in the shortest possible time for rapid processing, but also need to filter out unwanted data, but the data collection process, we can confirm the current data be collected for post data key input, for data-dependent judgment is an extremely complex task.

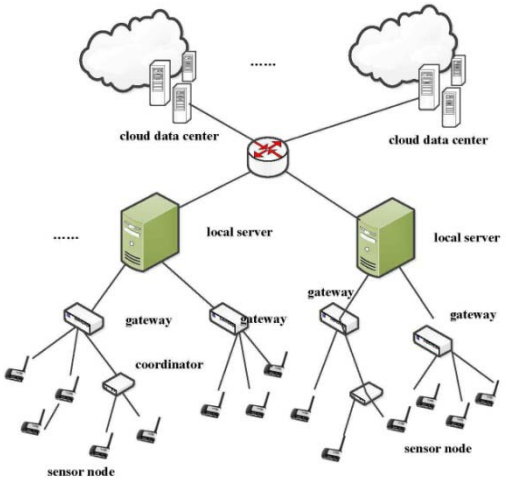

Fig. 8: Multi-Level Data Storage Model

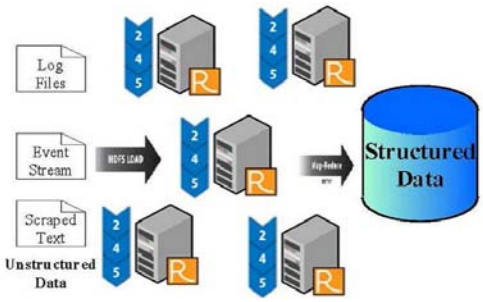

Fig. 9: Structured Data

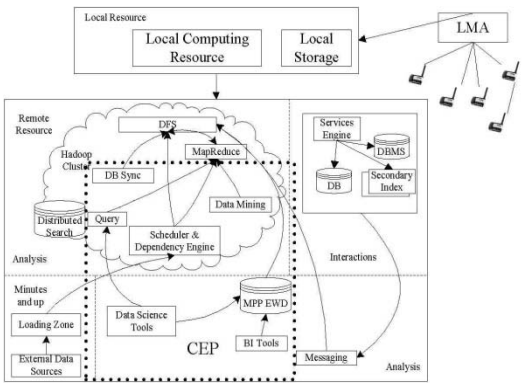

Fig. 10: Schematic of Storage and Analysis

h) Date Processing Algorithm

Along with architectural patterns, algorithm framework also plays an important role in RTDP systems computation results. In recent years, many research has done in big data processing algorithm, but the research about real-time big data has not been taken into account.

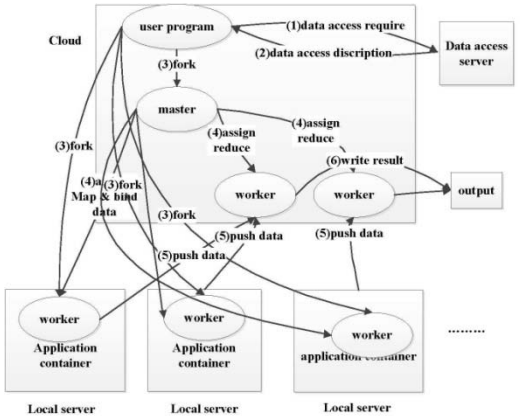

Fig.11: New Big Data Processing Model

i) Calculation Implementation Method

In previous chapters, a two layers' calculation mode has been proposed, first, the local server choose local node management and calculation procedures on the local node management, and simple data cleansing and structured modeling according to LMA. Unstructured data collected by data collector will be transformed into structured data and then uploaded to cloud memory systems and mapped to different management servers. Superstardom makes use of the computing power of cloud terminal to carry through real-time computation and analyze.

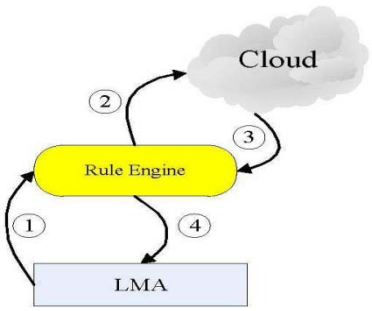

Fig.12: Application Deployment Process

j) Decision Making Layer

Truth be told, basic leadership layer incorporates two sections of ideas, to begin with, the test and refresh the model, the second is to give administrators to basic leadership. RTDP framework amid the procedure of information handling, with the stream of information, the information at various circumstances with a specific changeability, and between information likewise has a specific pertinence. subsequently change with time and profundity

information preparing, information investigation layer information show made may not meet the present needs, so we have to keep the information handling while the refresh information and refresh the information model to adjust to changes in the information then again, choice bolster layer is the most elevated amount of RTDP framework, the reason for existing is to complete information preparing related choices, so the layer must picture the created yield brings about request to give leaders to oversee related basic leadership exercises. Next a capacity diagram will be settled on about model approval and choice support.

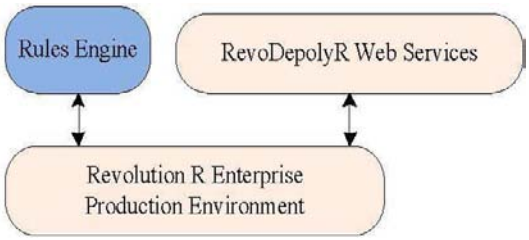

Fig.13: Application Deployment Process

VI. Application Examples

What's more, application area related RTDP explore has likewise got a great deal of consideration, and have made some preparatory research. IBM in 2010 formally proposed "shrewd city" vision and from that point forward a vast continuous information and research to get savvy city boundless consideration as urban framework concentrates canny transportation, brilliant matrix, and urban administrations identified with therapeutic insight has likewise been an incredible improvement. Different applications have their own particular one of a kind quality, many key issues still uncertain, a hefty portion of the current zones most important reviews likewise stay in the research facility stage. A case of RTDP shrewd matrix keen lattice framework in the application review will be made underneath.

Savvy lattice through countless time sensor detecting framework operation state changes, can give quicker element assessment of constant blame conclusion and vitality tracking in covering areas, districts and even across the nation dynamic direction of vitality to accomplish conveyance vitality creation with sensible dispatch, for enhancing vitality effectiveness, enhancing urban foundation assumes an essential part.

The various specialized challenges are all inside the extent of the entire system and information identified with checking and constant computation, so this paper the answer for the present field of keen lattice of the key issues, we should depend RTDP handling structures for expansive scale organize wide continuous information joining the proposed RTDP structure to fabricate another sort of shrewd network design, appeared in Figure 15.

Quick reenactment and demonstrating is the center programming of ADO, including hazard evaluation, self-mending and other propelled control and advancement programming framework for the savvy lattice to offer help and prescient numerical capacity, keeping in mind the end goal to accomplish enhanced network strength, security, dependability and operational productivity. Dispersion quick recreation and demonstrating need to bolster arrange reconfiguration, voltage and responsive power control; blame area, separation and reclamation of power; when the framework topology changes taking after the security re-tuning four self-mending abilities. Above capacity interconnectedness, coming about DFSM turn out to be extremely convoluted, for instance, either a network reproduction requires another hand-off with voltage control or the new celebration program likewise incorporates capacities to reestablish control. DFSM by means of dispersed insightful system specialists to accomplish hierarchical limits crosswise over geological limits and canny control framework keeping in mind the end goal to accomplish self-mending capacities of these keen system operators, ready to gather and trade data and frameworks, (for example, the accompanying such electrical insurance operation) nearby control choices, while as indicated by the framework prerequisites to facilitate these projects.

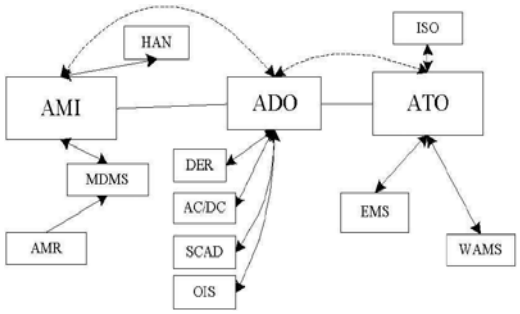

Fig.14: Technical Composition and Functionality of Smart Grid

Shrewd Grid unwavering quality issues contemplated by the AC transmission hardware for blunder infusion approach the gear disappointment mode investigation. Disappointment of the hardware and on its impact and power network unwavering quality were surveyed. writing demonstrates that the present development of brilliant framework mix of utilizations and innovation arrangements, including savvy meters, correspondence systems, metering database administration (MDMS), client premises organize (HAN), client benefit, remote turn on or off, as appeared in Figure 14, keen lattice advancements has achieved a specific level of accessibility and adaptability, however in the force of constant sending, administration, blame location and recuperation, there are still deficiencies.

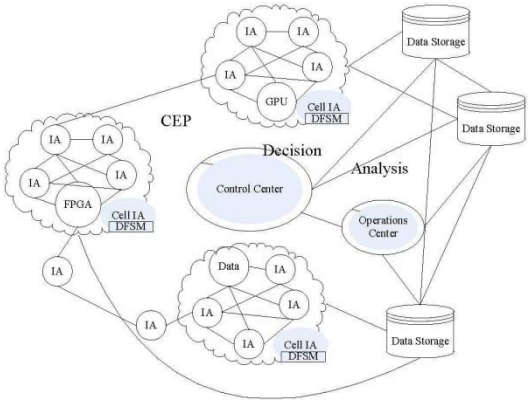

Fig.15: Architecture of Distributed Smart Grid

VII. Conclusion & Future Work

Real-time data processing for large current technology is undoubtedly a huge challenge, there is lack of support for massive real-time processing of large data frame and platform real-time processing of large data processing compared to conventional static data with high data throughput and real-time requirements. cloud computing technology in order to solve massive data processing and developed a series of techniques, however, cloud computing is very suitable for mass static, long-term without the written data has a good effect, but it is difficult to achieve real-time processing.

In this paper, on the basis of cloud computing technology to build a kind of real-time processing of large data frame, the model proposed RTDP four architecture, and hierarchical computing model. RTDP system in order to meet real-time requirements, and to consider different system platforms RTDP structural characteristics, the paper also presents a large data storage for real-time multi-level storage model and the LMA-based application deployment methods of data collection terminal based on the different ways of data processing were used DSMS, CEP, batch-based MapReduce other processing mode, depending on the environment in which the sensor data acquisition and the desired type of difference data collected were used FPGA, GPU, CPU, ASIC technology to achieve data collection and data cleansing and, through a structured process the data structure modeling, uploaded to the cloud server for storage, while the washed structured data on the local server for Reduce, combined with powerful computing capabilities cloud architecture for large-scale real-time computing with MapReduce.

This thesis indicates generally that the basic framework for future RTDP system and basic processing mode, but there are still many issues that need further study. The main point are as follows:

1. How to determine the appropriate mode of calculation in a RTDP system, how to determine the data processing mode and approach is a key factor in determining system performance, so the calculation mode and how to determine the appropriate method of calculating the design of the future core of the work;
2. Calculation models and how to achieve unity between computing technology is currently used mainly batch calculation mode and streaming processing, data computing model in determining how to design the corresponding calculation after the manner and with what kind of hardware implementation is the next big real-time data processing priority;
3. How to ensure the network transmission speed and QoS (Quality of Services); now widely used in a variety of network QoS technology, RTDP not sufficient to ensure a real-time, high reliability requirement. RTDP network QoS issues RTDP with difficulty from the inherent characteristics, so to guarantee QoS of the real-time RTDP sex have a significant impact;
4. How to ensure the system's physical time synchronization. RTDP system involves many interactions between systems and tools, software used for real-time marker approach does not meet the future RTDP high real-time requirements, the interactive how to ensure data during physical time synchronization is the future research directions;
5. How to ensure the correctness of the data processing. Error detection mechanism and automatically repair the computer has long been a difficult area of research, how to handle the data detection and error diagnostic and system repair is a huge project.

RTDP is a complex project involving many disciplines and techniques to be thorough in all aspects of research, pointed out that the article provides an overview of future research directions, and this is our future research subject.

References Références Referencias

1. Zhigao Zheng, Ping Wang, Jing Liu, Appl. Math. Inf. Sci. 9, No. 6, 3169-3190 (2015).
2. Tene O, Polonetsky J. Big data for all: Privacy and user control in the age of analytics[J]. Nw. J. Tech. & Intell. Prop., 2012, 11: xxvii.
3. LohrS.Theage ofbigdata[J].NewYorkTimes, 2012,11.
4. LynchC.Bigdata:How do your data grow ?[J].Nature, 2008, 455*(7209)*:2829
5. Bryant, RE, Katz, RH, Lazowska, ED, Big-Data Computing: Creating revolutionary breakthroughs in commerce, science, and society [R], Ver. 8, (2008), Computing Research Association, Computing Community Consortium.
6. Jonathan T. Overpeck, Gerald A. Meehl, Sandrine Bony, and David R. Easterling. Climate Data

Challenges in the 21st Century [J]. Science, 2011, 331(6018): 700-702
7. Labrinidis, Alexandros and Jagadish, H. V. Challenges and opportunities with big data [J]. Proc. VLDB Endow. 2012, 5(12): 2032-2033.
8. WANG Shan, WANG Hui-Ju, QIN Xiong-Pai, ZHOU Xuan. Architecting Big Data: Challenges, Studies and Forecasts [J]. Chinese Journal of Computer. 2011, 34 (10): 1741-1752.
9. Lu Weixing, Shou Yinbiao, Shi Lianjun. WSCC DISTURBANCE ON AUGUST 10, 1996 IN THE UNITED
10. P. Jeffrey Palermo. The August 14, 2003 blackout and Its importance to China [J]. EAST CHINA ELECERIC POWER. 2004, 32(1): 2-6.
11. Li Cuiping, Wang Minfeng .Excerpts from the Translation of Challenges and Opportunities with Big Data[J].e-Science Technology & Application, 2013, 4(1): 12-18.
12. Dobbie W, Fryer Jr RG. Getting beneath the veil of effective schools: Evidence from New YorkCity[R]. National Bureau of Economic Research,2011.
13. H.V.Jagadish, Johannes Gehrke, Alexandros Labrinidis, Yannis Papakonstantinou, Jignesh M. Patel, Raghu Ramakrishnan, Cyrus Shahabi. Bigdata and its technical challenges[J]. Communications of the ACM,2014, 57(7): 86-94.
14. Flood M, Jagadish H V, Kyle A, et al. Using Data for Systemic Financial Risk Management[C]. Proceedings of The 5th biennial Conference on Innovative Data Systems Research (CIDR 2011). 2011: 144-147.
15. Genovese Y, Prentice S. Pattern-based strategy: getting value from big data[J]. Gartner Special Report G, 2011, 214032: 2011.
16. Albert-Lszl Barabsi. The network takeover. Nature Physics, 2012, 8(1): 14-16.
17. Labrinidis A, Jagadish H V. Challenges and opportunities with big data[J]. Proceedings of the VLDB Endowment, 2012, 5(12): 2032-2033.
18. Lohr S. How big data became so big[J]. New York Times, 2012, 11.
19. Gattiker A, Gebara F H, Hofstee H P, et al. Big Data text-oriented benchmark creation for Hadoop[J]. IBM Journal of Research and Development, 2013, 57(3/4): 10: 1-10: 6.
20. Chen M, Mao S, Liu Y. Big data: A survey[J]. Mobile Networks and Applications, 2014, 19(2): 171- 209.
21. Li Guojie, Cheng Xueqi. Research Status and Scientific Thinking of Big Data [J]. Bulletin of the Chinese Academy of Sciences. 2012, 27(6): 647-657.
22. Yadagiri S, Thalluri P V S. Information technology on surge: information literacy on demand[J]. DESIDOC Journal of Library & Information Technology, 2011, 32(1):64-69.
23. Cohen J, Dolan B, Dunlap M, Hellerstein JM, Welton C. MAD skills: New analysis practices for big data [J]. PVLDB, 2009, 2(2): 14811492.
24. Randal E. Bryant & Joan Digney. Data-Intensive Supercomputing: The case for DISC [R].2007.10: 1-14.
25. John Boyle. Biology must develop its own big-data systems. Nature. 2008, 499(7): 7.
26. Wang Yuan-Zhuo, Jin Xiao-Long, Chen Xue-Qi. Network Big Data: Present and Future [J].Chinese Journal of Computer. 2013, 36(6): 1125-1138.
27. Zou Guowei, Cheng Jianbo. The Application of Big Data Technology to Smart City [J]. POWER SYSTEM TECHNOLOGY, 2013, 4: 25-28.
28. QIN Xiong-Pai, WANG Hui-Ju, DU Xiao-Yong, WANG Shan. Big Data Analysis-Competition and Symbiosis of RDBMS and MapReduce [J]. Journal of Software. 2012, 23(1): 32-45.
29. Tan Xiongpai, Wang Huiju, Li Furong, et al. New Landscape of Data Management Technologies [J]. Journal of Software. 2013, 24(2): 175-197.
30. CHEN Hai-Ming, CUI Li, XIE Kai-Bin. A Comparative Study on Architectures and Implementation Methodologies of Internet of Things [J]. Chinese Journal of Computers. 2013, 36(1): 168-188.
31. Lee E A, Seshia S A. Introduction to embedded systems: A cyber-physical systems approach[M]. Lee & Seshia, 2011.
32. Thusoo A, Sarma JS,Jainn, etal. Hive-Apeta byte scale data ware house using Hadoop [C]. Proc. of ICDE2010. Piscataway, NJ:IEEE, 2010: 996-1005 Piscataway, NJ: IEEE, 2010: 996-1005
33. Abouzied A, Bajda-Pawlikowski K, Huang Jiewen, et al. HadoopDB in action: Building real world applications [C]. Proc. of SIGMOD 2010, New York: ACM, 2010: 1111-1114.
34. Chen Songting. Cheetah: A high performance, custom data warehouse on top of MapReduce [J]. PVLDB, 2010, 3(2): 1459-1468.
35. Agrawal R, Srikant R. Privacy preserving data mining [C]. Proc. of SIGMOD 2000. New York: ACM, 2000: 439-450.
36. Dwork C. Differential privacy [C]. Proc. of ICALP 2006. Berlin: Springer, 2006: 1-12.
37. Norman D A. The Design of Everyday Things [M].New York: Basic Books. 2002.
38. Olston C, Reed B, Srivastava U, et al. Pig Latin: A not-so-foreign language for data processing [C]. Proc of SIGMOD 2008, New York:ACM, 2008: 1099-1110.
39. Pike R, Dorward S, Griesemter R, et al. Interpreting the data: Parallel analysis with Sawzall [J]. Scientific Programming, 2005, 13(4): 277-298.
40. Chaiken R, Jenkins B, Larson P-A, et al. SCOPE: Easy and efficient parallel processing of massive data sets [J]. PVLDB, 2008, 1(2): 1265-1276.

41. Isard M, Yu Y. Distributed data-parallel computing using a high-level programming language [C]. Proc. of SIGMOD 2009. New York: ACM, 2009: 987-994.
42. Fegaras L, Li C, Gupta U, et al. XML query optimization in MapReduce [C]. Proc. of WebDB 2011. New York: ACM, 2011.
43. Morton K, Balazinska M, Grosstman D. Para Timer: A progress indicator for MapReduce DAGs [C]. Proc. of SIGMOD 2010. New York: ACM, 2010: 507-518.
44. Morton K, Friesen A, Balazinka A, et al. KAMD: Estimating the progress of MapReduce pipelines [C]. Proc. of ICDE 2010. Piscataway, NJ: IEEE, 2010: 681-684.
45. Huang Dachuan, Shi Xuanhua, Ibrabim Shadi, et al. MR- scope: A real-time tracing tool for MapReduce [C]. Proc. of HPDC 2010. New York: ACM, 2010: 849-855.
46. Meng Xiaofeng, Ci Xiang. Big Data Management: Concepts, Techniques and Challenges [J]. Journal of Computer Research and Development. 2013, 50(1): 146-169.
47. Chen Y. We dont know enough to make a big data benchmark suite-an academia-industry view[C]. Proc. of WBDB, 2012.
48. Chen Yanpei, Ganapathi A, Griffith R, et al. The case for evaluating MapReduce performance using workload suites [C]. Proc. of MASCOTS 2011. Piscataway, NJ: IEEE, 2011: 390-399.
49. Sangroya A, Serrano D, Bouchenak S. Mrbs: A comprehensive mapreduce benchmark suite[R]. LIG, Grenoble, France, Research Report RR-LIG-024, 2012.
50. Tan J, Kavulya S, Gandhi R, et al. Light-weight black-box failure detection for distributed systems[C]. Proceedings of the 2012 workshop on Management of big data systems. ACM, 2012: 13-18.
51. Zhao J M, Wang W S, Liu X, et al. Big data benchmark- big DS[M]. Advancing Big Data Benchmarks. Springer International Publishing, 2014: 49-57.
52. Patil S, Polte M, Ren K, et al. YCSB++: Benchmarking and performance debugging advance features in scalable table stores [C]. Proc. of SoCC 2011. New York: ACM, 2011.
53. Chen Y, Alspaugh S, Katz R. Interactive query processing in big data systems: A cross-industry study of MapReduce workloads [J]. PVLDB, 2012, 5(12).1802-1813.

Global Journal of Computer Science and Technology: C
Software & Data Engineering
Volume 18 Issue 3 Version 1.0 Year 2018
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

An Extended Linked Clustering Algorithm for Spatial Data Sets

By K. Lakshmaiah, Dr. S Murali Krishna & Dr. B Eswara Reddy
Sir Vishveshwaraiah Institute of Science & Technology

Abstract- Spatial data mining techniques and for the most part conveyed clustering are broadly utilized as a part of the most recent decade since they manage huge and differing datasets which can't be assembled midway. Current disseminated clustering approaches are typically producing universal models by amassing neighborhood outcomes that are acquired on every region. While this approach mines the data collections on their areas the accumulation stage is more perplexing, which may deliver inaccurate, and equivocal all universal clusters and in this manner mistaken learning. In this paper we propose an Extended Linked clustering approach for each huge spatial data collections that are assorted and appropriated. The approach in view of K-means algorithm yet it produces the quantity of all universal clusters progressively. In addition this approach utilizes an explained collection stage. The conglomerations stage is outlined in such way that the general procedure is proficient in time and memory assignment .Preliminary outcomes demonstrate that the proposed approach delivers excellent outcomes and scales up well. We likewise contrasted it with two prominent clustering algorithms and demonstrate that this approach is substantially more proficient.

Keywords: *spatial data, extended linked clustering, distributed data mining, data analysis, k-means, aggregation.*

GJCST-C Classification: *I.5.3*

Strictly as per the compliance and regulations of:

An Extended Linked Clustering Algorithm for Spatial Data Sets

K. Lakshmaiah[α], Dr. S Murali Krishna[σ] & Dr. B Eswara Reddy[ρ]

Abstract- Spatial data mining techniques and for the most part conveyed clustering are broadly utilized as a part of the most recent decade since they manage huge and differing datasets which can't be assembled midway. Current disseminated clustering approaches are typically producing universal models by amassing neighborhood outcomes that are acquired on every region. While this approach mines the data collections on their areas the accumulation stage is more perplexing, which may deliver inaccurate, and equivocal all universal clusters and in this manner mistaken learning. In this paper we propose an Extended Linked clustering approach for each huge spatial data collections that are assorted and appropriated. The approach in view of K-means algorithm yet it produces the quantity of all universal clusters progressively. In addition this approach utilizes an explained collection stage. The conglomerations stage is outlined in such way that the general procedure is proficient in time and memory assignment .Preliminary outcomes demonstrate that the proposed approach delivers excellent outcomes and scales up well. We likewise contrasted it with two prominent clustering algorithms and demonstrate that this approach is substantially more proficient.

Keywords: spatial data, extended linked clustering, distributed data mining, data analysis, k-means, aggregation.

I. Introduction

Over a wide assortment of fields, datasets are being gathered and amassed at a sensational pace and enormous measures of data that are being assembled are put away in various destinations. In this specific situation, data mining (DM) strategies have turned out to be vital for removing valuable learning from the quickly developing substantial and multi-dimensional datasets [1]. Keeping in mind the end goal to adapt to vast volumes of data, analysts have created parallel forms of the consecutive DM algorithms [2]. These parallel renditions may help to speedup serious calculations, yet they present critical correspondence overhead, which make them wasteful. To decrease the correspondence overheads circulated data mining (DDM) approaches that comprise of two principle steps are proposed. As the data is normally circulated the main stage comprises of executing the mining procedure on neighborhood datasets on every node to make nearby outcomes. These neighborhood results will be collected to fabricate all inclusive ones. Along these lines the effectiveness of any DDM calculation depends nearly on the productivity of its collection stage. In this unique situation, appropriated data mining (DDM) systems with proficient total stage have turned out to be fundamental for investigating these expansive and multi-dimensional datasets. In addition, DDM is more proper for expansive scale disseminated stages, for example, Clusters and Grids [3], where datasets are regularly geologically circulated and possessed by various associations. Many DDM techniques, for example, disseminated affiliation governs and circulated characterization [4], [5], [6], [7], [8], [9] have been proposed and created over the most recent couple of years. Be that as it may, just a couple of research concerns disseminated clustering for dissecting vast, diversed and conveyed datasets. Ongoing investigates [10], [11], [12], [13] have proposed conveyed clustering approaches in view of a similar 2-step process: perform halfway examination on nearby data at singular destinations and after that send them to a local region to create all universal models by accumulating the neighborhood comes about. In this paper, we propose a conveyed clustering approach in view of a similar 2-step process, be that as it may, it diminishes fundamentally the measure of data traded amid the total stage, produces consequently the right number of groups, and furthermore it can utilize any clustering algorithm to play out the investigation on nearby datasets. A contextual analysis of a proficient conglomeration stage has been produced on unique datasets and turned out to be extremely effective; the data traded is lessened by over 98% of the first datasets [15].

Whatever remains of this paper is sorted out as takes after: In the following segment we will give a diagram of dispersed data mining and examine the constraints of customary strategies. At that point we will introduce and talk about our approach in Section 3. Area 4 introduces the usage of the approach and we talk about exploratory outcomes in Section 5. At last, we finish up in Section.

II. Spatial Distributed Data Mining

Existing DDM procedures comprise of two principle stages: 1) performing halfway investigation on

Author α: B. Tech, M. Tech, [Ph.D]., MISTE., MIAENG., Assoc., Professor Dept of Computer Science and Engineering Sir Vishveshwaraiah Institute of Science & Technology MADANAPALLE–517325, Chittoor Dist, Andhra Pradesh.
e-mail: klakshmaiah78@gmail.com
Author σ: Professor in CSE Dept, SV College Of Engineering, TIRUPATI. Chittoor Dist, Andhra Pradesh. India.
Author ρ: Professor and Principal, JNTUA College of Engineering, Kalikiri, Chittoor Dist, Andhra Pradesh. India.

nearby data at singular destinations and 2) producing all universal models by amassing the neighborhood comes about. These two stages are not autonomous since credulous ways to deal with neighborhood investigation may deliver erroneous and questionable all inclusive data models. So as to exploit mined data at various areas, DDM ought to have a perspective of the learning that encourages their reconciliation as well as limits the impact of the nearby outcomes on the general models. Quickly, a productive administration of appropriated learning is one of the key variables influencing the yields of these procedures.

Additionally, the data that will be gathered in various areas utilizing diverse instruments may have distinctive arrangements, highlights, and quality. Conventional incorporated data mining procedures don't consider every one of the issues of data driven applications, for example, adaptability in both reaction time and exactness of arrangements, appropriation and heterogeneity [8], [16].

Some DDM approaches depend on outfit realizing, which utilizes different procedures to total the outcomes [11], among the most refered to in the writing: greater part voting, weighted voting, and stacking [17], [18]. A few methodologies are appropriate to be performed on disseminated stages. For example, the incremental calculations for finding spatio-transient examples by breaking down the hunt space into a progressive structure, tending to its application to multi-granular spatial data can be effectively streamlined on various leveled disseminated framework topology. From the writing, two classifications of methods are utilized: parallel procedures that frequently require devoted machines and instruments for correspondence between parallel procedures which are exceptionally costly, and systems in light of conglomeration, which continue with an absolutely conveyed, either on the data construct models or in light of the execution stages [7], [12]. Nonetheless, the measure of data keeps on expanding as of late, in that capacity, the larger part of existing data mining strategies are not performing admirably as they experiences the versatility issue. This turns into an exceptionally basic issue as of late. Numerous arrangements have been proposed up until this point. They are for the most part in view of little changes to fit a specific data close by.

Clustering is one of the major strategies in data mining. It Clusters data objects in view of data found in the data that portrays the articles and their connections. The objective is to streamline closeness measure inside a bunch and the dissimilarities between groups with a specific end goal to distinguish fascinating structures/designs/models in the data [12]. The two principle classes of bunching are parceling and various leveled. Diverse expounded scientific classifications of existing grouping calculations are given in the writing and numerous appropriated bunching variants in light of these calculations have been proposed in [12], [20]–[25], and so forth. Parallel bunching calculations are grouped into two sub-classifications. The principal comprises of techniques requiring various rounds of message passing. They require a lot of synchronization. The second sub-class comprises of techniques that manufacture nearby bunching models and send them to a focal site to construct all inclusive models[15].In [20] and [24], message-passing versions of the widely used k-means algorithm were proposed. In [21] and [25], the authors dealt with the parallelization of the DBSCAN density based clustering algorithm. In [22] a parallel message passing version of the BIRCH algorithm was presented. A parallel version of a hierarchical clustering algorithm, called MPC for Message Passing Clustering, which is especially dedicated to Microarray data, was introduced in [23]. Most of the parallel approaches need either multiple synchronization constraints between processes or a universal view of the dataset, or both [12].

Both dividing and various leveled classes have a few shortcomings. For the parceling class, the k-means algorithm requires the quantity of clusters to be settled ahead of time, while in the lion's share of cases K isn't known, moreover various leveled clustering algorithms have beaten this restriction, however they should characterize the halting conditions for clustering deterioration, which are not direct. limitation, but they must define the stopping conditions for clustering decomposition, which are not straightforward.

III. Extended Linked Spatial Distributed Clustering

The proposed circulated approach takes after the regular two-advance system; 1) it initially creates neighborhood clusters on each sub-dataset that is alloted to a given preparing node, 2) these nearby groups are accumulated to frame all universal ones. This approach is produced for clustering spatial datasets. The nearby clustering algorithm can be any clustering algorithm. For purpose of clearness it is been K-Means executed with guaranteed (Ki) which can be diverse for every node (see Figure 1). Ki ought to be been sufficiently huge to recognize all clusters in the nearby data sets.

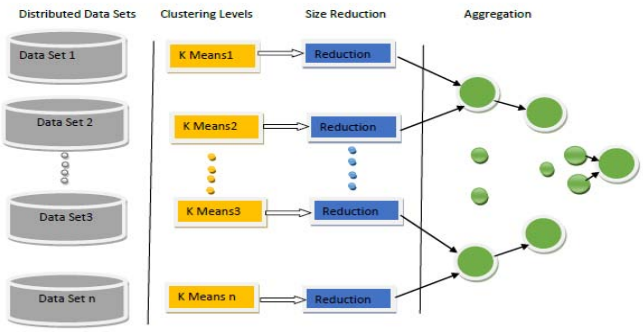

Fig. 1: The Frame work of the Proposed Approach

In the wake of producing nearby outcomes, every node contrasts its neighborhood Clusters and its neighbors' groups. A portion of the nodes, called pioneer, will be chosen to combine neighborhood Clusters to frame bigger groups utilizing the overlay method. These pioneers are chosen by a few conditions, for example, their ability, their handling fake, and so on. The way toward blending groups will proceed until the point that we achieve the root node. The root node will contain the widespread Clusters (models).

Amid the second stage, imparting the neighborhood groups to the pioneers may produce gigantic overhead. Thusly, the goal is to limit the data correspondence and computational time, while getting precise general outcomes. Actually our approach limits the overheads because of the data trade. Thusly as opposed to trading the entire data (entire Clusters) between nodes (neighborhood nodes and pioneers), we initially continue by lessening the data that speak to a group. The span of this new data group is significantly littler that the underlying one. This procedure is done on every nearby node.

There are the number of data diminishment methods proposed in the literature. A significant number of them are centering just in dataset measure i.e., they endeavor to decrease the capacity of the data without focusing on the learning behind this data. In [26], a proficient decrease method has been proposed; it depends on density based clustering algorithm. Each cluster comprises of its agents. Notwithstanding, choosing agents is as yet a test regarding quality and size. We can pick, for instance, medoids points, core points, or even specific core points [10] as representatives [15].

We revolve around the outline and the density of the clustering. The condition of a gathering is addressed by its farthest point centers (called frame) (see Fig 2). Various computations for removing the breaking points from a group can be found in the literature work [27], [28], [29], [30], [31]. We used the figuring proposed in [32] which relies upon Triangulation to make as far as possible. It is a successful figuring for creating non-angled points of confinement. The computation can definitely portray the condition of a broad assortment of dissimilar point flows and densities with a sensible diserse nature of O(n log n).

The limits of the Clusters speak to the new dataset, and they are substantially littler than the first datasets. So the limits of the Clusters will turn into the nearby outcomes at every node in the system. These neighborhood comes about are sent to the pioneers following a tree topology. The general outcomes will be situated at the foundation of the tree.

IV. Implemented Approach

a) Extended Linked Distributed Clustering Algorithm (ELDCA)

In the main stage, called the parallel stage, the neighborhood grouping is performed utilizing the K-means calculation. Every node (di) executes K-means on its nearby dataset to create Ki neighborhood Clusters. When all the nearby groups are resolved, we ascertain their forms. These shapes will be utilized as delegates of their comparing groups. The second period of the method comprises of trading the forms of every node with its neighborhood nodes. This will enable us to check whether there are any covering shapes (Clusters).

In the third step every pioneer endeavors to consolidate covering shapes of its gathering. The pioneers are chosen among nodes of each gathering. In this way, every pioneer produces new shapes (new Clusters). We rehash the second and third steps till we achieve root node. The sub-groups collection is finished after a tree structure and the all inclusive outcomes are situated in the best level of the tree (root node).

As in all Cluster calculations, the normal huge inconstancy in groups shapes and densities is an issue. Be that as it may, as we will appear in the following segment, the calculation utilized for producing the group's form is proficient to distinguish all around isolated clusters with any shapes. In addition ELDCA decides likewise progressively the quantity of the clusters without from the earlier data about the data or an estimation procedure of the quantity of the groups. In the accompanying we will depict the principle highlights and the necessities of the calculation and its condition. The nodes of the distributed computing system are organised following a tree topology.

A. Each node is dispensed a dataset speaking to a part of the scene or of the general dataset.
B. Each leaf node (ni) executes the K-means algorithm with Ki parameter on its neighborhood information.
C. Neighbouring nodes must share their groups to shape much bigger clusters utilizing the overlay system. The results must reside in the father node (called ancestor).
D. Repeat C and D until reaching the root node.

In the following we give a pseudo-code of the algorithm:

Algorithm 1: Extended Linked Distributed Clustering Algorithm (ELDCA)

Input : D_i: Dataset Fragment, K_i: Number of sub-clusters for $Node_i$, T: tree degree.

Output: Ku: Universal Clusters (universal results)
level = treeheight;
I. K-means(Di. K_i);
// *$Node_i$ executes K-Means algorithm locally.*

II. Contour(K_i);
// Node-i executes Contour algorithm to create the limit of each cluster produced locally.
III. *Node*$_i$ joins a group G of T elements;
// Node$_i$ *joins his neighbourhood.*
IV. Compare cluster of Node$_i$ to other Node's clusters in the same group;
// search for covering between Clusters
V. j= Elect leader Node();
// choose a node which will combine the covering Clusters

```
if (i <> j) then
Send(contour i, j);
else
      if( level > 0) then
       level - - ;
       Repeat III, IV, and V until level=1;
     else
        return (K_u: Node_i's selected clusters);
```

b) Example of execution

We suppose that the system contains five Nodes (N=5), and every Node executes K-Means algorithmic rule with totally different Ki, because it is shown in Fig 2. Node1 executes the K-Means with K=40, Node2 with K=80, Node3 with K=120, Node4 with k=180, and Node5 with K= 220. so every node within the system generates its native clusters. future step consists of merging overlapping clusters at intervals the neighborhood. As we are able to see, though we have a tendency to started with totally different values of K, we have a tendency to generated solely 5 clusters results (See Fig 2).

V. Experimental Results

In this segment, we examine the execution of ELDCA Algorithm and show its viability contrasted with BIRCH and CURE calculations:

BIRCH: We utilized the execution of BIRCH gave by the creators in [33]. It plays out a pre-grouping and after that uses a centroid-based various leveled bunching calculation. Note that the time and space many-sided quality of this approach is quadratic to the quantity of focuses after pre-grouping. We set parameters to the default esteems recommended in [33].

CURE: We utilized the usage of CURE gave by the creators in [34]. The calculation utilizes agent focuses with contracting towards the mean. As portrayed in [34], when two groups are converged in each progression of the calculation, agent focuses for the new blended group are chosen from the ones of the two unique clusters as opposed to every one of the focuses in the consolidated clusters.

ELDCA: Our calculation is portrayed in Section IV. The key point in our approach is to pick Ki greater than the right number of groups. As portrayed toward the finish of Section IV, when two groups are converged in each progression of the calculation, delegate purposes of the new consolidated bunch are the association of the shapes of the two unique groups instead of all focuses in the new group. This paces up the execution time without unfavorably affecting on the nature of the created groups. Also, our system utilizes the tree topology, store information structures and Agglomerative various leveled grouping. Accordingly, this additionally enhances the many-sided quality of the calculation.

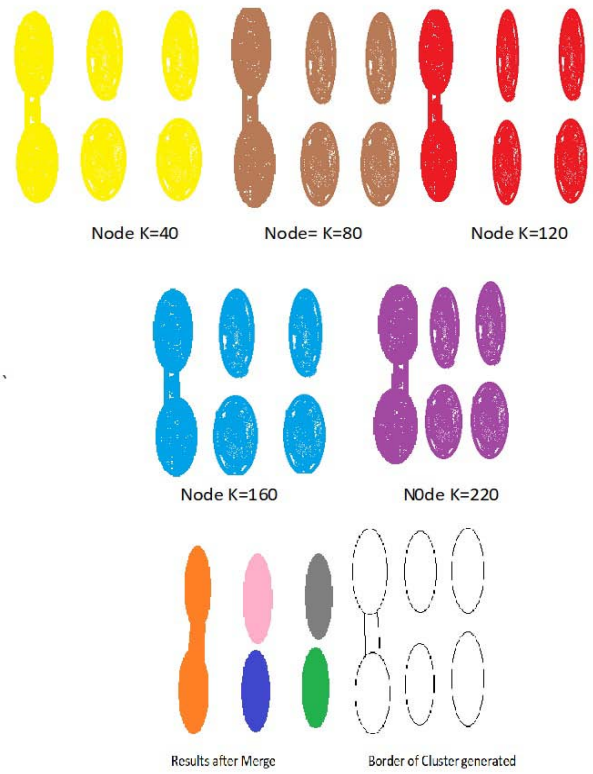

Fig. 2: Extended Linked Distributed Clustering Algorithm (ELDCA)

a) Data sets

We run experiments with different datasets. In this paper we use three types of datasets. These are summarised in Table 1. The number of points and clusters in each dataset is also given in Table 1. We show that ELDCA algorithm not only correctly clusters the datasets, but also its execution time is much quicker than BIRCH and CURE.

b) The Obtained Quality of Clustering

We run the three algorithms on the three datasets to compare them with respect to the quality of clusters generated and their response time. Fig 3, Fig 4 and Fig 5 show the clusters found by the three algorithms for the three datasets (dataset1, dataset2 and dataset3). We use different colours to show the clusters returned by each algorithm.

Fig 3 shows the clusters generated from the dataset1. As expected, since BIRCH uses a centroid-based hierarchical clustering algorithm for clustering the pre-clustered points, it could not find all the clusters correctly. It splits the larger cluster while merging the others. In contrast, the CURE algorithm succeeds to generate the majority of clusters but it still fails to discover all the correct clusters. Our distributed clustering algorithm successfully generates all the clusters with the default parameter settings described in section IV. As it is shown in Fig 3, after merging the local clusters, we generated five final clusters.

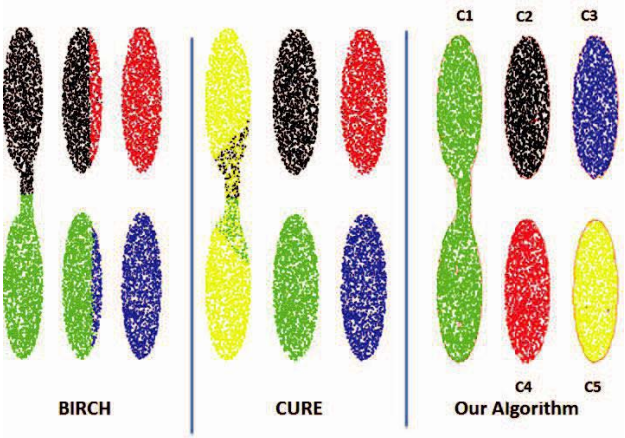

Fig. 3: Clusters generated from dataset 1.

Table 1: Datasets

Data Sets	Numbers of points	Shape of Clusters	Number of Clusters
Data set 1	16000	Big Oval (Egg Shape)	Five
Data set 2	41350	2 Small Circles,1 Big Circle and 2 Ovals Linked	Four
Data set 3	19080	4 Circles and 2 Circles Linked	Five

Fig 4 shows the outcomes found by the three algorithms for the dataset 2. Once more, BIRCH and CURE neglected to create every one of the clusters, while our algorithm effectively produced the four right clusters.

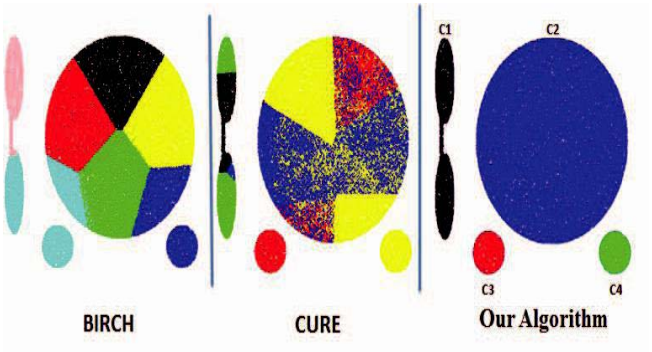

Fig. 4: Clusters generated from dataset 2.

Fig 5 Represents the clusters we found from the dataset 3. As should be obvious BIRCH still neglects to discover every one of the clusters effectively. Interestingly CURE found the 5 clusters, yet not flawlessly. For example, we can see some red focuses in the blue cluster and some blue focuses in the green cluster. Our Algorithm produced the five clusters effectively and impeccably.

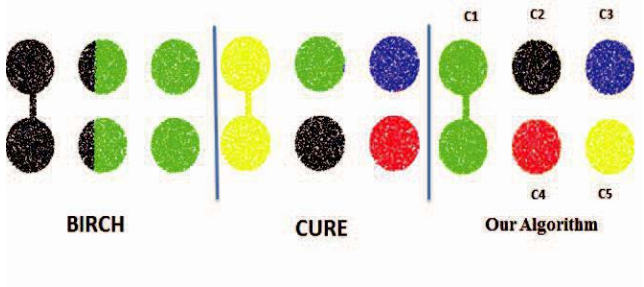

Fig. 5: Clusters generated from dataset 3.

c) Observations

As should be obvious, our method effectively produced the last groups for the three datasets. This is because of the way that:

At the point when two groups are combined, the new bunch is spoken to by the association of the two shapes of the two unique bunches. This paces up the execution times without affecting the nature of groups generated. The number of all inclusive bunches is dynamic.

d) Comparison of ELDCA's Execution Time to BIRCH and CURE

The goal here is to demonstrate the impact of using the combination of parallel and distributed architecture to deal with the limited capacity of a node in the system and tree topology to accelerate the speed of computation.

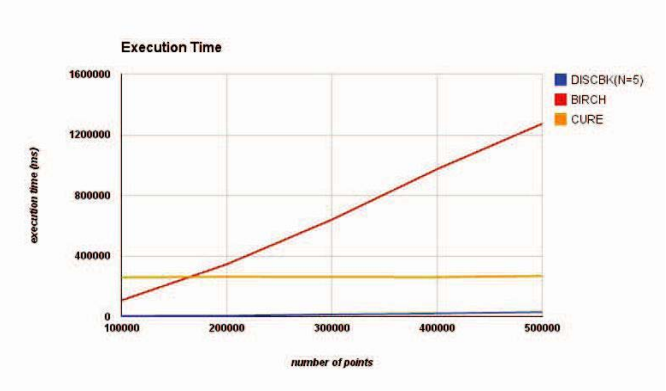

Fig. 6: Comparison to BIRCH and CURE.

Fig. 6 shows the execution of our algorithm contrasted with BIRCH and CURE as the quantity of information directs increments from 100,000 toward 500,000 the quantity of groups and their shapes are not adjusted. In this manner, for our calculation we think about the quantity of nodes in the system: N=5. The execution times do exclude the ideal opportunity for showing the clusters since these are the same for the three algorithms.

As can be found in Fig 6, ELDCA's execution time is much lower than CURE's and BIRCH's execution times. Moreover, as the quantity of focuses expands, our execution time is about even, while, the executions time of BIRCH increments quickly with the dataset estimate. This is on the grounds that BIRCH sweeps the whole database and uses every one of the focuses for pre-clustering. At long last as the quantity of focuses expands the CURE's execution time is about even, since CURE utilizes a testing method, where the span of this example remains the same and the main extra cost brought about by CURE is simply the inspecting strategy.

The above outcomes affirm that our disseminated clustering algorithm is extremely proficient contrasted with both BIRCH and CURE either in nature of the clusters created and in the computational time.

e) Scalability

The objective of the adaptability tests is to decide the impacts of the quantity of nodes in the framework on the execution times. The dataset contains 1000,000 focuses. Fig 7 demonstrates the execution time against the quantity of nodes in the framework. Our calculation took just a couple of moments to group 1000,000 focuses in a conveyed framework that contains more than 100 nodes. In this way, the calculation can serenely deal with high-dimensional data in view of its low multifaceted nature.

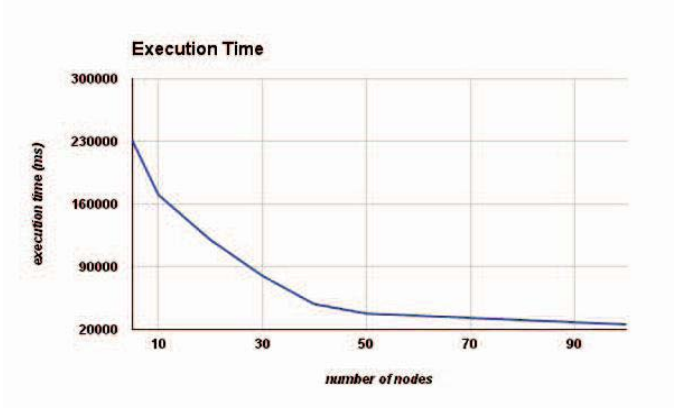

Fig. 7: Scalability Experiments.

VI. Conclusions

In this paper, we propose another and imaginative Extended Linked DCA, to manage spatial datasets. This approach misuses the preparing intensity of the appropriated stage by augmenting the parallelism and limiting the interchanges and for the most part the measure of the information that is traded between the hubs in the framework. Nearby models are created by executing a grouping calculation in every hub, and afterward the neighborhood comes about are converged to construct the all inclusive clusters. The nearby models are spoken to with the goal that their sizes are sufficiently little to be traded through the system.

Trial comes about are likewise displayed and talked about. They likewise demonstrate the viability of ELDCA either on amount of the clustering produced or the execution time contrasting with BIRCH and CURE calculations. Besides, they show that the calculation outflanks existing calculations as well as scales well for extensive databases without giving up the grouping quality. ELDCA is not quite the same as present dispersed grouping models introduced in the writing, it describes by the dynamic number of clusters created and its proficient information decrease stage.

A more broad assessment is continuous. We will plan to run tries different things with different neighborhood algorithms and investigate the conceivable outcomes of stretching out the strategies to different sorts of expansive and appropriated datasets.

References Références Referencias

1. U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "Knowledge discovery and data mining: Towards a unifying framework," in *Proc. KDD-96*, 1996, pp. 82–88.
2. A. A. Freitas and S. H. Lavington, *Mining very large databases with parallel processing*. 1st edition, Springer; 2000 edition, 30 November 2007.

3. I. Foster and C. Kesselman, *The Grid: Blueprint for a New Computing Infrastructure*. Morgan Kaufmann Publishers Inc. San Francisco, CA, USA, 1999.
4. T. G. Dietterich, "An experimental comparison of three methods for Constructing ensembles of decision trees: Bagging, boosting and randomization," *Machine Learning*, vol. 40, pp. 139–157, 2000.
5. H. Kargupta and P. Chan, *Advances in distributed and Parallel Knowledge Discovery*, USA MIT Press Cambridge, MA, October 2000.
6. R. Agrawal and J. C. Shafer, "Parallel mining of association rules," *in proc*. *IEEE Transactions on Knowledge and Data Engineering*, vol. 8, pp. 962–969, 1996.
7. L-M. Aouad, N-A. Le-Khac, and M-T. Kechadi, "Performance study of a distributed apriori-like frequent itemsets mining technique," *Knowledge Data Systems*, vol. 23, pp. 55-72, Apr. 2010.
8. L-M. Aouad, N-A. Le-Khac, and M-T. Kechadi, "Grid-based approaches for distributed data mining applications*," algorithms Computational Technology*, vol. 3, pp. 517–534, 10 Dec. 2009.
9. N-A. Le-Khac, L-M. Aouad, and M-T. Kechadi, "Toward a distributed knowledge discovery on grid systems, "*in Emergent Web Intelligence: Advanced Semantic Technologies*, London. Springer, April 2010, pp 213-243.
10. E. Januzaj, H-P. Kriegel, and M. Pfeifle, "DBDC: Density-based distributed clustering," *in Advances in Database Technology,* vol. 2992, Greece, March 14-18, 2004, pp. 88-105.
11. N-A. Le-Khac, L-M. Aouad, and M-T. Kechadi, "A new approach for distributed density based clustering on grid platform*." In Data Management. Data, Data Everywhere,* Volume 4587, Springer-Verlag Berlin, Heidelberg, 2007, pp. 247–258.
12. L-M. Aouad, N-A. Le-Khac, and M-T. Kechadi, "Lightweight clustering Technique for distributed data mining applications," *in Advances in Data Mining. Theoretical Aspects and Applications*, Germany. Springer Berlin Heidelberg, 2007, pp. 120–134.
13. L-M. Aouad, N-A. Le-Khac, and M-T. Kechadi, *Advances in Data Mining. Theoretical Aspects and Applications*. Ed Berlin Heidelberg, Germany Springer 14-18 July 2007.
14. J. Han, M. Kamber, J. Pei, *Data Mining Concept and Techniques*, 2nd edition. Morgan Kaufmann, 6 April 2006.
15. J.F. Laloux, N-A. Le-Khac, and M-T. Kechadi, "Efficient distributed approach for density-based clustering*," Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE), 20th IEEE International Workshop*s, pp. 145 – 150, 27-29, June 2011.
16. M. Bertolotto, S. Di Martino, F.Ferrucci, and M-T. Kechadi, "Towards a framework for mining and analysing spatio-temporal datasets," *International Journal of Geographical Data Science – Geovisual Analytics for Spatial Decision Support*, vol. 21, pp. 895-906, January 2007.
17. P. Chan and S. J. Stolfo, "A comparative evaluation of voting and meta-learning on partitioned data*," in In Proceedings of the Twelfth International Conference on Machine Learning*. Morgan Kaufmann, 1995, pp. 90–98.
18. C. R. Reeves, *Modern heuristic techniques for combinatorial problems*. 1st edition, John Wiley & Sons, Inc. New York, NY, USA, May 1993.
19. L-M. Aouad, N-A. Le-Khac, and M-T. Kechadi, "Performance study of a distributed apriori-like frequent itemsets mining technique," *Knowledge Data Systems,* Springer-Verlag, vol. 23, pp 55-72, 2009.
20. I. S. Dhillon and D. S. Modha, "A data-clustering algorithm on distributed memory multiprocessor*," in large-Scale Parallel Data Mining, Workshop on Large-Scale Parallel KDD Systems, SIGKDD*. Springer-Verlag London, UK, 1999, pp. 245–260.
21. M. Ester, H. P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise." *In proc. KDD96*, 1996, pp. 226–231.
22. A. Garg, A. Mangla, V. Bhatnagar, and N. Gupta, "PBirch: A scalable parallel clustering algorithm for incremental data," *in proc*. *Database Engineering and Applications Symposium. IDEAS' 06. 10th International, Delhi*, 2006, pp. 315-316.
23. H. Geng, and X. Deng, "A new clustering algorithm using message passing and its applications in analyzing microarray data*," in proc. ICMLA '05 Proceedings of the Fourth International Conference on Machine Learning and Applications. IEEE*, 15-17 December 2005, pp. 145–150.
24. I. D. Dhillon and D. S. Modha, "A data-clustering algorithm on distributed memory multiprocessors," *in proc. Large-Scale Parallel Data Mining. Springer Berlin Heidelberg*, 2000, pp. 245–260.
25. X. Xu, J. Jager, and H. P. Kriegel, "A fast parallel clustering algorithm for large spatial databases," *in Data Mining and Knowledge Discovery archive*, vol. 3, September 1999, pp. 263 – 290.
26. N-A. L-Khac, M. Bue, and M. Whelan, "A knowledge based data reduction for very large spatio-temporal datasets*," in proc. International Conference on Advanced Data Mining and Applications (ADMA'2010).* Springer Verlag LNCS/LNAI, Chongquing, China, November 19-21,
27. J. M. Fadili and M. Melkemi and A. ElMoataz,"Non-convex onion-peeling using a shape hull algorithm," *Pattern Recognition Letters,* vol. 25, pp. 1577 – 1585, 14-15 October 2004.

28. A. R. Chaudhuri and B. B. Chaudhuri and S. K. Parui, "A novel approach to computation of the shape of a dot pattern and extraction of its perceptual border," *Computer vision and Image Understanding*, vol. 68, pp. 257- 275 , 03 December 1997.
29. M. Melkemi and M. Djebali, "Computing the shape of a planar points set," *Pattern Recognition,* vol. 33, pp. 1423–1436, 9 September 2000.
30. H. Edelsbrunner, D. Kirkpatrick, and R. Seidel, "On the shape of a set of points in the plane," *IEEE Transaction on Data Theory*, vol. 29,pp. 551 – 559, July 1983.
31. A. Moreira and M. Y. Santos, "Concave hull: A k-nearest neighbours approach for the computation of the region occupied by a set of points," *in proc. proceedings of the International Conference on Computer Graphics Theory and Applications*, March 2007.
32. M. Duckham, L. Kulik, and M. Worboys, "Efficient generation of simple polygons for characterizing the shape of a set of points in the plane," *Pattern Recognition,* vol. 41, pp. 3224–3236, 15 March 2008.
33. T. Zhang, and R. Ramakrishnan and M. Livny, "Birch: An efficient data clustering method for very large databases*," in proc. SIGMOD '96 Proceedings of the 1996 ACM SIGMOD international conference on Management of data*, vol. 25. ACM New York, NY, USA, 1996, pp. 103–114.
34. S. Guha and R. Rastogi and K. Shim, "Cure: An efficient clustering algorithm for large databases," *Data Systems,* vol. 26, pp. 35– 58, Nov. 2001.

Global Journals Guidelines Handbook 2018

www.GlobalJournals.org

FELLOW OF ASSOCIATION OF RESEARCH SOCIETY IN COMPUTING (FARSC)

Global Journals Incorporate (USA) is accredited by Open Association of Research Society (OARS), U.S.A and in turn, awards "FARSC" title to individuals. The 'FARSC' title is accorded to a selected professional after the approval of the Editor-in-Chief/Editorial Board Members/Dean.

- The "FARSC" is a dignified title which is accorded to a person's name viz. Dr. John E. Hall, Ph.D., FARSC or William Walldroff, M.S., FARSC.

FARSC accrediting is an honor. It authenticates your research activities. After recognition as FARSC, you can add 'FARSC' title with your name as you use this recognition as additional suffix to your status. This will definitely enhance and add more value and repute to your name. You may use it on your professional Counseling Materials such as CV, Resume, and Visiting Card etc.

The following benefits can be availed by you only for next three years from the date of certification:

FARSC designated members are entitled to avail a 40% discount while publishing their research papers (of a single author) with Global Journals Incorporation (USA), if the same is accepted by Editorial Board/Peer Reviewers. If you are a main author or co-author in case of multiple authors, you will be entitled to avail discount of 10%.

Once FARSC title is accorded, the Fellow is authorized to organize a symposium/seminar/conference on behalf of Global Journal Incorporation (USA).The Fellow can also participate in conference/seminar/symposium organized by another institution as representative of Global Journal. In both the cases, it is mandatory for him to discuss with us and obtain our consent.

You may join as member of the Editorial Board of Global Journals Incorporation (USA) after successful completion of three years as Fellow and as Peer Reviewer. In addition, it is also desirable that you should organize seminar/symposium/conference at least once.

We shall provide you intimation regarding launching of e-version of journal of your stream time to time. This may be utilized in your library for the enrichment of knowledge of your students as well as it can also be helpful for the concerned faculty members.

The FARSS can go through standards of OARS. You can also play vital role if you have any suggestions so that proper amendment can take place to improve the same for the benefit of entire research community.

As FARSS, you will be given a renowned, secure and free professional email address with 100 GB of space e.g. johnhall@globaljournals.org. This will include Webmail, Spam Assassin, Email Forwarders,Auto-Responders, Email Delivery Route tracing, etc.

The FARSS will be eligible for a free application of standardization of their researches. Standardization of research will be subject to acceptability within stipulated norms as the next step after publishing in a journal. We shall depute a team of specialized research professionals who will render their services for elevating your researches to next higher level, which is worldwide open standardization.

The FARSS member can apply for grading and certification of standards of their educational and Institutional Degrees to Open Association of Research, Society U.S.A.
Once you are designated as FARSS, you may send us a scanned copy of all of your credentials. OARS will verify, grade and certify them. This will be based on your academic records, quality of research papers published by you, and some more criteria. After certification of all your credentials by OARS, they will be published on your Fellow Profile link on website https://associationofresearch.org which will be helpful to upgrade the dignity.

The FARSS members can avail the benefits of free research podcasting in Global Research Radio with their research documents. After publishing the work, (including published elsewhere worldwide with proper authorization) you can upload your research paper with your recorded voice or you can utilize chargeable services of our professional RJs to record your paper in their voice on request.

The FARSS member also entitled to get the benefits of free research podcasting of their research documents through video clips. We can also streamline your conference videos and display your slides/ online slides and online research video clips at reasonable charges, on request.

The FARSS is eligible to earn from sales proceeds of his/her researches/reference/review Books or literature, while publishing with Global Journals. The FARSS can decide whether he/she would like to publish his/her research in a closed manner. In this case, whenever readers purchase that individual research paper for reading, maximum 60% of its profit earned as royalty by Global Journals, will be credited to his/her bank account. The entire entitled amount will be credited to his/her bank account exceeding limit of minimum fixed balance. There is no minimum time limit for collection. The FARSS member can decide its price and we can help in making the right decision.

The FARSS member is eligible to join as a paid peer reviewer at Global Journals Incorporation (USA) and can get remuneration of 15% of author fees, taken from the author of a respective paper. After reviewing 5 or more papers you can request to transfer the amount to your bank account.

MEMBER OF ASSOCIATION OF RESEARCH SOCIETY IN SCIENCE (MARSS)

The ' MARSS ' title is accorded to a selected professional after the approval of the Editor-in-Chief / Editorial Board Members/Dean.

The “MARSS” is a dignified ornament which is accorded to a person’s name viz. Dr. John E. Hall, Ph.D., MARSS or William Walldroff, M.S., MARSS.

MARSS accrediting is an honor. It authenticates your research activities. After becoming MARSS, you can add 'MARSS' title with your name as you use this recognition as additional suffix to your status. This will definitely enhance and add more value and repute to your name. You may use it on your professional Counseling Materials such as CV, Resume, Visiting Card and Name Plate etc.

The following benefitscan be availed by you only for next three years from the date of certification.

MARSS designated members are entitled to avail a 25% discount while publishing their research papers (of a single author) in Global Journals Inc., if the same is accepted by our Editorial Board and Peer Reviewers. If you are a main author or co-author of a group of authors, you will get discount of 10%.

As MARSS, you will be given a renowned, secure and free professional email address with 30 GB of space e.g. johnhall@globaljournals.org. This will include Webmail, Spam Assassin, Email Forwarders,Auto-Responders, Email Delivery Route tracing, etc.

We shall provide you intimation regarding launching of e-version of journal of your stream time to time. This may be utilized in your library for the enrichment of knowledge of your students as well as it can also be helpful for the concerned faculty members.

The MARSC member can apply for approval, grading and certification of standards of their educational and Institutional Degrees to Open Association of Research, Society U.S.A.

Once you are designated as MARSC, you may send us a scanned copy of all of your credentials. OARS will verify, grade and certify them. This will be based on your academic records, quality of research papers published by you, and some more criteria.

It is mandatory to read all terms and conditions carefully.

Auxiliary Memberships

Institutional Fellow of Open Association of Research Society (USA)-OARS (USA)

Global Journals Incorporation (USA) is accredited by Open Association of Research Society, U.S.A (OARS) and in turn, affiliates research institutions as "Institutional Fellow of Open Association of Research Society" (IFOARS).

The "FARSC" is a dignified title which is accorded to a person's name viz. Dr. John E. Hall, Ph.D., FARSC or William Walldroff, M.S., FARSC.

The IFOARS institution is entitled to form a Board comprised of one Chairperson and three to five board members preferably from different streams. The Board will be recognized as "Institutional Board of Open Association of Research Society"-(IBOARS).

The Institute will be entitled to following benefits:

The IBOARS can initially review research papers of their institute and recommend them to publish with respective journal of Global Journals. It can also review the papers of other institutions after obtaining our consent. The second review will be done by peer reviewer of Global Journals Incorporation (USA) The Board is at liberty to appoint a peer reviewer with the approval of chairperson after consulting us.

The author fees of such paper may be waived off up to 40%.

The Global Journals Incorporation (USA) at its discretion can also refer double blind peer reviewed paper at their end to the board for the verification and to get recommendation for final stage of acceptance of publication.

The IBOARS can organize symposium/seminar/conference in their country on behalf of Global Journals Incorporation (USA)-OARS (USA). The terms and conditions can be discussed separately.

The Board can also play vital role by exploring and giving valuable suggestions regarding the Standards of "Open Association of Research Society, U.S.A (OARS)" so that proper amendment can take place for the benefit of entire research community. We shall provide details of particular standard only on receipt of request from the Board.

The board members can also join us as Individual Fellow with 40% discount on total fees applicable to Individual Fellow. They will be entitled to avail all the benefits as declared. Please visit Individual Fellow-sub menu of GlobalJournals.org to have more relevant details.

We shall provide you intimation regarding launching of e-version of journal of your stream time to time. This may be utilized in your library for the enrichment of knowledge of your students as well as it can also be helpful for the concerned faculty members.

After nomination of your institution as "Institutional Fellow" and constantly functioning successfully for one year, we can consider giving recognition to your institute to function as Regional/Zonal office on our behalf.
The board can also take up the additional allied activities for betterment after our consultation.

The following entitlements are applicable to individual Fellows:

Open Association of Research Society, U.S.A (OARS) By-laws states that an individual Fellow may use the designations as applicable, or the corresponding initials. The Credentials of individual Fellow and Associate designations signify that the individual has gained knowledge of the fundamental concepts. One is magnanimous and proficient in an expertise course covering the professional code of conduct, and follows recognized standards of practice.

Open Association of Research Society (US)/ Global Journals Incorporation (USA), as described in Corporate Statements, are educational, research publishing and professional membership organizations. Achieving our individual Fellow or Associate status is based mainly on meeting stated educational research requirements.

Disbursement of 40% Royalty earned through Global Journals : Researcher = 50%, Peer Reviewer = 37.50%, Institution = 12.50% E.g. Out of 40%, the 20% benefit should be passed on to researcher, 15 % benefit towards remuneration should be given to a reviewer and remaining 5% is to be retained by the institution.

We shall provide print version of 12 issues of any three journals [as per your requirement] out of our 38 journals worth $ 2376 USD.

Other:

The individual Fellow and Associate designations accredited by Open Association of Research Society (US) credentials signify guarantees following achievements:

- The professional accredited with Fellow honor, is entitled to various benefits viz. name, fame, honor, regular flow of income, secured bright future, social status etc.

- In addition to above, if one is single author, then entitled to 40% discount on publishing research paper and can get 10%discount if one is co-author or main author among group of authors.
- The Fellow can organize symposium/seminar/conference on behalf of Global Journals Incorporation (USA) and he/she can also attend the same organized by other institutes on behalf of Global Journals.
- The Fellow can become member of Editorial Board Member after completing 3yrs.
- The Fellow can earn 60% of sales proceeds from the sale of reference/review books/literature/publishing of research paper.
- Fellow can also join as paid peer reviewer and earn 15% remuneration of author charges and can also get an opportunity to join as member of the Editorial Board of Global Journals Incorporation (USA)
- • This individual has learned the basic methods of applying those concepts and techniques to common challenging situations. This individual has further demonstrated an in–depth understanding of the application of suitable techniques to a particular area of research practice.

Note :

"

- In future, if the board feels the necessity to change any board member, the same can be done with the consent of the chairperson along with anyone board member without our approval.
- In case, the chairperson needs to be replaced then consent of 2/3rd board members are required and they are also required to jointly pass the resolution copy of which should be sent to us. In such case, it will be compulsory to obtain our approval before replacement.
- In case of "Difference of Opinion [if any]" among the Board members, our decision will be final and binding to everyone.

"

Preferred Author Guidelines

We accept the manuscript submissions in any standard (generic) format.

We typeset manuscripts using advanced typesetting tools like Adobe In Design, CorelDraw, TeXnicCenter, and TeXStudio. We usually recommend authors submit their research using any standard format they are comfortable with, and let Global Journals do the rest.

Alternatively, you can download our basic template from https://globaljournals.org/Template.zip

Authors should submit their complete paper/article, including text illustrations, graphics, conclusions, artwork, and tables. Authors who are not able to submit manuscript using the form above can email the manuscript department at submit@globaljournals.org or get in touch with chiefeditor@globaljournals.org if they wish to send the abstract before submission.

Before and during Submission

Authors must ensure the information provided during the submission of a paper is authentic. Please go through the following checklist before submitting:

1. Authors must go through the complete author guideline and understand and *agree to Global Journals' ethics and code of conduct,* along with author responsibilities.
2. Authors must accept the privacy policy, terms, and conditions of Global Journals.
3. Ensure corresponding author's email address and postal address are accurate and reachable.
4. Manuscript to be submitted must include keywords, an abstract, a paper title, co-author(s') names and details (email address, name, phone number, and institution), figures and illustrations in vector format including appropriate captions, tables, including titles and footnotes, a conclusion, results, acknowledgments and references.
5. Authors should submit paper in a ZIP archive if any supplementary files are required along with the paper.
6. Proper permissions must be acquired for the use of any copyrighted material.
7. Manuscript submitted *must not have been submitted or published elsewhere* and all authors must be aware of the submission.

Declaration of Conflicts of Interest

It is required for authors to declare all financial, institutional, and personal relationships with other individuals and organizations that could influence (bias) their research.

Policy on Plagiarism

Plagiarism is not acceptable in Global Journals submissions at all.

Plagiarized content will not be considered for publication. We reserve the right to inform authors' institutions about plagiarism detected either before or after publication. If plagiarism is identified, we will follow COPE guidelines:

Authors are solely responsible for all the plagiarism that is found. The author must not fabricate, falsify or plagiarize existing research data. The following, if copied, will be considered plagiarism:

- Words (language)
- Ideas
- Findings
- Writings
- Diagrams
- Graphs
- Illustrations
- Lectures

- Printed material
- Graphic representations
- Computer programs
- Electronic material
- Any other original work

Authorship Policies

Global Journals follows the definition of authorship set up by the Open Association of Research Society, USA. According to its guidelines, authorship criteria must be based on:

1. Substantial contributions to the conception and acquisition of data, analysis, and interpretation of findings.
2. Drafting the paper and revising it critically regarding important academic content.
3. Final approval of the version of the paper to be published.

Changes in Authorship

The corresponding author should mention the name and complete details of all co-authors during submission and in manuscript. We support addition, rearrangement, manipulation, and deletions in authors list till the early view publication of the journal. We expect that corresponding author will notify all co-authors of submission. We follow COPE guidelines for changes in authorship.

Copyright

During submission of the manuscript, the author is confirming an exclusive license agreement with Global Journals which gives Global Journals the authority to reproduce, reuse, and republish authors' research. We also believe in flexible copyright terms where copyright may remain with authors/employers/institutions as well. Contact your editor after acceptance to choose your copyright policy. You may follow this form for copyright transfers.

Appealing Decisions

Unless specified in the notification, the Editorial Board's decision on publication of the paper is final and cannot be appealed before making the major change in the manuscript.

Acknowledgments

Contributors to the research other than authors credited should be mentioned in Acknowledgments. The source of funding for the research can be included. Suppliers of resources may be mentioned along with their addresses.

Declaration of funding sources

Global Journals is in partnership with various universities, laboratories, and other institutions worldwide in the research domain. Authors are requested to disclose their source of funding during every stage of their research, such as making analysis, performing laboratory operations, computing data, and using institutional resources, from writing an article to its submission. This will also help authors to get reimbursements by requesting an open access publication letter from Global Journals and submitting to the respective funding source.

Preparing your Manuscript

Authors can submit papers and articles in an acceptable file format: MS Word (doc, docx), LaTeX (.tex, .zip or .rar including all of your files), Adobe PDF (.pdf), rich text format (.rtf), simple text document (.txt), Open Document Text (.odt), and Apple Pages (.pages). Our professional layout editors will format the entire paper according to our official guidelines. This is one of the highlights of publishing with Global Journals—authors should not be concerned about the formatting of their paper. Global Journals accepts articles and manuscripts in every major language, be it Spanish, Chinese, Japanese, Portuguese, Russian, French, German, Dutch, Italian, Greek, or any other national language, but the title, subtitle, and abstract should be in English. This will facilitate indexing and the pre-peer review process.

The following is the official style and template developed for publication of a research paper. Authors are not required to follow this style during the submission of the paper. It is just for reference purposes.

Manuscript Style Instruction (Optional)

- Microsoft Word Document Setting Instructions.
- Font type of all text should be Swis721 Lt BT.
- Page size: 8.27" x 11'", left margin: 0.65, right margin: 0.65, bottom margin: 0.75.
- Paper title should be in one column of font size 24.
- Author name in font size of 11 in one column.
- Abstract: font size 9 with the word "Abstract" in bold italics.
- Main text: font size 10 with two justified columns.
- Two columns with equal column width of 3.38 and spacing of 0.2.
- First character must be three lines drop-capped.
- The paragraph before spacing of 1 pt and after of 0 pt.
- Line spacing of 1 pt.
- Large images must be in one column.
- The names of first main headings (Heading 1) must be in Roman font, capital letters, and font size of 10.
- The names of second main headings (Heading 2) must not include numbers and must be in italics with a font size of 10.

Structure and Format of Manuscript

The recommended size of an original research paper is under 15,000 words and review papers under 7,000 words. Research articles should be less than 10,000 words. Research papers are usually longer than review papers. Review papers are reports of significant research (typically less than 7,000 words, including tables, figures, and references)

A research paper must include:

a) A title which should be relevant to the theme of the paper.
b) A summary, known as an abstract (less than 150 words), containing the major results and conclusions.
c) Up to 10 keywords that precisely identify the paper's subject, purpose, and focus.
d) An introduction, giving fundamental background objectives.
e) Resources and techniques with sufficient complete experimental details (wherever possible by reference) to permit repetition, sources of information must be given, and numerical methods must be specified by reference.
f) Results which should be presented concisely by well-designed tables and figures.
g) Suitable statistical data should also be given.
h) All data must have been gathered with attention to numerical detail in the planning stage.

Design has been recognized to be essential to experiments for a considerable time, and the editor has decided that any paper that appears not to have adequate numerical treatments of the data will be returned unrefereed.

i) Discussion should cover implications and consequences and not just recapitulate the results; conclusions should also be summarized.
j) There should be brief acknowledgments.
k) There ought to be references in the conventional format. Global Journals recommends APA format.

Authors should carefully consider the preparation of papers to ensure that they communicate effectively. Papers are much more likely to be accepted if they are carefully designed and laid out, contain few or no errors, are summarizing, and follow instructions. They will also be published with much fewer delays than those that require much technical and editorial correction.

The Editorial Board reserves the right to make literary corrections and suggestions to improve brevity.

Format Structure

It is necessary that authors take care in submitting a manuscript that is written in simple language and adheres to published guidelines.

All manuscripts submitted to Global Journals should include:

Title

The title page must carry an informative title that reflects the content, a running title (less than 45 characters together with spaces), names of the authors and co-authors, and the place(s) where the work was carried out.

Author details

The full postal address of any related author(s) must be specified.

Abstract

The abstract is the foundation of the research paper. It should be clear and concise and must contain the objective of the paper and inferences drawn. It is advised to not include big mathematical equations or complicated jargon.

Many researchers searching for information online will use search engines such as Google, Yahoo or others. By optimizing your paper for search engines, you will amplify the chance of someone finding it. In turn, this will make it more likely to be viewed and cited in further works. Global Journals has compiled these guidelines to facilitate you to maximize the web-friendliness of the most public part of your paper.

Keywords

A major lynchpin of research work for the writing of research papers is the keyword search, which one will employ to find both library and internet resources. Up to eleven keywords or very brief phrases have to be given to help data retrieval, mining, and indexing.

One must be persistent and creative in using keywords. An effective keyword search requires a strategy: planning of a list of possible keywords and phrases to try.

Choice of the main keywords is the first tool of writing a research paper. Research paper writing is an art. Keyword search should be as strategic as possible.

One should start brainstorming lists of potential keywords before even beginning searching. Think about the most important concepts related to research work. Ask, “What words would a source have to include to be truly valuable in a research paper?” Then consider synonyms for the important words.

It may take the discovery of only one important paper to steer in the right keyword direction because, in most databases, the keywords under which a research paper is abstracted are listed with the paper.

Numerical Methods

Numerical methods used should be transparent and, where appropriate, supported by references.

Abbreviations

Authors must list all the abbreviations used in the paper at the end of the paper or in a separate table before using them.

Formulas and equations

Authors are advised to submit any mathematical equation using either MathJax, KaTeX, or LaTeX, or in a very high-quality image.

Tables, Figures, and Figure Legends

Tables: Tables should be cautiously designed, uncrowned, and include only essential data. Each must have an Arabic number, e.g., Table 4, a self-explanatory caption, and be on a separate sheet. Authors must submit tables in an editable format and not as images. References to these tables (if any) must be mentioned accurately.

Figures

Figures are supposed to be submitted as separate files. Always include a citation in the text for each figure using Arabic numbers, e.g., Fig. 4. Artwork must be submitted online in vector electronic form or by emailing it.

Preparation of Eletronic Figures for Publication

Although low-quality images are sufficient for review purposes, print publication requires high-quality images to prevent the final product being blurred or fuzzy. Submit (possibly by e-mail) EPS (line art) or TIFF (halftone/ photographs) files only. MS PowerPoint and Word Graphics are unsuitable for printed pictures. Avoid using pixel-oriented software. Scans (TIFF only) should have a resolution of at least 350 dpi (halftone) or 700 to 1100 dpi (line drawings). Please give the data for figures in black and white or submit a Color Work Agreement form. EPS files must be saved with fonts embedded (and with a TIFF preview, if possible).

For scanned images, the scanning resolution at final image size ought to be as follows to ensure good reproduction: line art: >650 dpi; halftones (including gel photographs): >350 dpi; figures containing both halftone and line images: >650 dpi.

Color charges: Authors are advised to pay the full cost for the reproduction of their color artwork. Hence, please note that if there is color artwork in your manuscript when it is accepted for publication, we would require you to complete and return a Color Work Agreement form before your paper can be published. Also, you can email your editor to remove the color fee after acceptance of the paper.

Tips for writing a good quality Computer Science Research Paper

Techniques for writing a good quality computer science research paper:

1. Choosing the topic: In most cases, the topic is selected by the interests of the author, but it can also be suggested by the guides. You can have several topics, and then judge which you are most comfortable with. This may be done by asking several questions of yourself, like "Will I be able to carry out a search in this area? Will I find all necessary resources to accomplish the search? Will I be able to find all information in this field area?" If the answer to this type of question is "yes," then you ought to choose that topic. In most cases, you may have to conduct surveys and visit several places. Also, you might have to do a lot of work to find all the rises and falls of the various data on that subject. Sometimes, detailed information plays a vital role, instead of short information. Evaluators are human: The first thing to remember is that evaluators are also human beings. They are not only meant for rejecting a paper. They are here to evaluate your paper. So present your best aspect.

2. Think like evaluators: If you are in confusion or getting demotivated because your paper may not be accepted by the evaluators, then think, and try to evaluate your paper like an evaluator. Try to understand what an evaluator wants in your research paper, and you will automatically have your answer. Make blueprints of paper: The outline is the plan or framework that will help you to arrange your thoughts. It will make your paper logical. But remember that all points of your outline must be related to the topic you have chosen.

3. Ask your guides: If you are having any difficulty with your research, then do not hesitate to share your difficulty with your guide (if you have one). They will surely help you out and resolve your doubts. If you can't clarify what exactly you require for your work, then ask your supervisor to help you with an alternative. He or she might also provide you with a list of essential readings.

4. Use of computer is recommended: As you are doing research in the field of computer science then this point is quite obvious. Use right software: Always use good quality software packages. If you are not capable of judging good software, then you can lose the quality of your paper unknowingly. There are various programs available to help you which you can get through the internet.

5. Use the internet for help: An excellent start for your paper is using Google. It is a wondrous search engine, where you can have your doubts resolved. You may also read some answers for the frequent question of how to write your research paper or find a model research paper. You can download books from the internet. If you have all the required books, place importance on reading, selecting, and analyzing the specified information. Then sketch out your research paper. Use big pictures: You may use encyclopedias like Wikipedia to get pictures with the best resolution. At Global Journals, you should strictly follow here.

6. Bookmarks are useful: When you read any book or magazine, you generally use bookmarks, right? It is a good habit which helps to not lose your continuity. You should always use bookmarks while searching on the internet also, which will make your search easier.

7. Revise what you wrote: When you write anything, always read it, summarize it, and then finalize it.

8. Make every effort: Make every effort to mention what you are going to write in your paper. That means always have a good start. Try to mention everything in the introduction—what is the need for a particular research paper. Polish your work with good writing skills and always give an evaluator what he wants. Make backups: When you are going to do any important thing like making a research paper, you should always have backup copies of it either on your computer or on paper. This protects you from losing any portion of your important data.

9. Produce good diagrams of your own: Always try to include good charts or diagrams in your paper to improve quality. Using several unnecessary diagrams will degrade the quality of your paper by creating a hodgepodge. So always try to include diagrams which were made by you to improve the readability of your paper. Use of direct quotes: When you do research relevant to literature, history, or current affairs, then use of quotes becomes essential, but if the study is relevant to science, use of quotes is not preferable.

10.Use proper verb tense: Use proper verb tenses in your paper. Use past tense to present those events that have happened. Use present tense to indicate events that are going on. Use future tense to indicate events that will happen in the future. Use of wrong tenses will confuse the evaluator. Avoid sentences that are incomplete.

11. Pick a good study spot: Always try to pick a spot for your research which is quiet. Not every spot is good for studying.

12. Know what you know: Always try to know what you know by making objectives, otherwise you will be confused and unable to achieve your target.

13. Use good grammar: Always use good grammar and words that will have a positive impact on the evaluator; use of good vocabulary does not mean using tough words which the evaluator has to find in a dictionary. Do not fragment sentences. Eliminate one-word sentences. Do not ever use a big word when a smaller one would suffice.

Verbs have to be in agreement with their subjects. In a research paper, do not start sentences with conjunctions or finish them with prepositions. When writing formally, it is advisable to never split an infinitive because someone will (wrongly) complain. Avoid clichés like a disease. Always shun irritating alliteration. Use language which is simple and straightforward. Put together a neat summary.

14. Arrangement of information: Each section of the main body should start with an opening sentence, and there should be a changeover at the end of the section. Give only valid and powerful arguments for your topic. You may also maintain your arguments with records.

15. Never start at the last minute: Always allow enough time for research work. Leaving everything to the last minute will degrade your paper and spoil your work.

16. Multitasking in research is not good: Doing several things at the same time is a bad habit in the case of research activity. Research is an area where everything has a particular time slot. Divide your research work into parts, and do a particular part in a particular time slot.

17. Never copy others' work: Never copy others' work and give it your name because if the evaluator has seen it anywhere, you will be in trouble. Take proper rest and food: No matter how many hours you spend on your research activity, if you are not taking care of your health, then all your efforts will have been in vain. For quality research, take proper rest and food.

18. Go to seminars: Attend seminars if the topic is relevant to your research area. Utilize all your resources.

19. Refresh your mind after intervals: Try to give your mind a rest by listening to soft music or sleeping in intervals. This will also improve your memory. Acquire colleagues: Always try to acquire colleagues. No matter how sharp you are, if you acquire colleagues, they can give you ideas which will be helpful to your research.

20. Think technically: Always think technically. If anything happens, search for its reasons, benefits, and demerits. Think and then print: When you go to print your paper, check that tables are not split, headings are not detached from their descriptions, and page sequence is maintained.

21. Adding unnecessary information: Do not add unnecessary information like "I have used MS Excel to draw graphs." Irrelevant and inappropriate material is superfluous. Foreign terminology and phrases are not apropos. One should never take a broad view. Analogy is like feathers on a snake. Use words properly, regardless of how others use them. Remove quotations. Puns are for kids, not grunt readers. Never oversimplify: When adding material to your research paper, never go for oversimplification; this will definitely irritate the evaluator. Be specific. Never use rhythmic redundancies. Contractions shouldn't be used in a research paper. Comparisons are as terrible as clichés. Give up ampersands, abbreviations, and so on. Remove commas that are not necessary. Parenthetical words should be between brackets or commas. Understatement is always the best way to put forward earth-shaking thoughts. Give a detailed literary review.

22. Report concluded results: Use concluded results. From raw data, filter the results, and then conclude your studies based on measurements and observations taken. An appropriate number of decimal places should be used. Parenthetical remarks are prohibited here. Proofread carefully at the final stage. At the end, give an outline to your arguments. Spot perspectives of further study of the subject. Justify your conclusion at the bottom sufficiently, which will probably include examples.

23. Upon conclusion: Once you have concluded your research, the next most important step is to present your findings. Presentation is extremely important as it is the definite medium though which your research is going to be in print for the rest of the crowd. Care should be taken to categorize your thoughts well and present them in a logical and neat manner. A good quality research paper format is essential because it serves to highlight your research paper and bring to light all necessary aspects of your research.

Informal Guidelines of Research Paper Writing

Key points to remember:

- Submit all work in its final form.
- Write your paper in the form which is presented in the guidelines using the template.
- Please note the criteria peer reviewers will use for grading the final paper.

Final points:

One purpose of organizing a research paper is to let people interpret your efforts selectively. The journal requires the following sections, submitted in the order listed, with each section starting on a new page:

The introduction: This will be compiled from reference matter and reflect the design processes or outline of basis that directed you to make a study. As you carry out the process of study, the method and process section will be constructed like that. The results segment will show related statistics in nearly sequential order and direct reviewers to similar intellectual paths throughout the data that you gathered to carry out your study.

The discussion section:

This will provide understanding of the data and projections as to the implications of the results. The use of good quality references throughout the paper will give the effort trustworthiness by representing an alertness to prior workings.

Writing a research paper is not an easy job, no matter how trouble-free the actual research or concept. Practice, excellent preparation, and controlled record-keeping are the only means to make straightforward progression.

General style:

Specific editorial column necessities for compliance of a manuscript will always take over from directions in these general guidelines.

To make a paper clear: Adhere to recommended page limits.

Mistakes to avoid:

- Insertion of a title at the foot of a page with subsequent text on the next page.
- Separating a table, chart, or figure—confine each to a single page.
- Submitting a manuscript with pages out of sequence.
- In every section of your document, use standard writing style, including articles ("a" and "the").
- Keep paying attention to the topic of the paper.
- Use paragraphs to split each significant point (excluding the abstract).
- Align the primary line of each section.
- Present your points in sound order.
- Use present tense to report well-accepted matters.
- Use past tense to describe specific results.
- Do not use familiar wording; don't address the reviewer directly. Don't use slang or superlatives.
- Avoid use of extra pictures—include only those figures essential to presenting results.

Title page:

Choose a revealing title. It should be short and include the name(s) and address(es) of all authors. It should not have acronyms or abbreviations or exceed two printed lines.

Abstract: This summary should be two hundred words or less. It should clearly and briefly explain the key findings reported in the manuscript and must have precise statistics. It should not have acronyms or abbreviations. It should be logical in itself. Do not cite references at this point.

An abstract is a brief, distinct paragraph summary of finished work or work in development. In a minute or less, a reviewer can be taught the foundation behind the study, common approaches to the problem, relevant results, and significant conclusions or new questions.

Write your summary when your paper is completed because how can you write the summary of anything which is not yet written? Wealth of terminology is very essential in abstract. Use comprehensive sentences, and do not sacrifice readability for brevity; you can maintain it succinctly by phrasing sentences so that they provide more than a lone rationale. The author can at this moment go straight to shortening the outcome. Sum up the study with the subsequent elements in any summary. Try to limit the initial two items to no more than one line each.

Reason for writing the article—theory, overall issue, purpose.

- Fundamental goal.
- To-the-point depiction of the research.
- Consequences, including definite statistics—if the consequences are quantitative in nature, account for this; results of any numerical analysis should be reported. Significant conclusions or questions that emerge from the research.

Approach:

- Single section and succinct.
- An outline of the job done is always written in past tense.
- Concentrate on shortening results—limit background information to a verdict or two.
- Exact spelling, clarity of sentences and phrases, and appropriate reporting of quantities (proper units, important statistics) are just as significant in an abstract as they are anywhere else.

Introduction:

The introduction should "introduce" the manuscript. The reviewer should be presented with sufficient background information to be capable of comprehending and calculating the purpose of your study without having to refer to other works. The basis for the study should be offered. Give the most important references, but avoid making a comprehensive appraisal of the topic. Describe the problem visibly. If the problem is not acknowledged in a logical, reasonable way, the reviewer will give no attention to your results. Speak in common terms about techniques used to explain the problem, if needed, but do not present any particulars about the protocols here.

The following approach can create a valuable beginning:

- Explain the value (significance) of the study.
- Defend the model—why did you employ this particular system or method? What is its compensation? Remark upon its appropriateness from an abstract point of view as well as pointing out sensible reasons for using it.
- Present a justification. State your particular theory(-ies) or aim(s), and describe the logic that led you to choose them.
- Briefly explain the study's tentative purpose and how it meets the declared objectives.

Approach:

Use past tense except for when referring to recognized facts. After all, the manuscript will be submitted after the entire job is done. Sort out your thoughts; manufacture one key point for every section. If you make the four points listed above, you will need at least four paragraphs. Present surrounding information only when it is necessary to support a situation. The reviewer does not desire to read everything you know about a topic. Shape the theory specifically—do not take a broad view.

As always, give awareness to spelling, simplicity, and correctness of sentences and phrases.

Procedures (methods and materials):

This part is supposed to be the easiest to carve if you have good skills. A soundly written procedures segment allows a capable scientist to replicate your results. Present precise information about your supplies. The suppliers and clarity of reagents can be helpful bits of information. Present methods in sequential order, but linked methodologies can be grouped as a segment. Be concise when relating the protocols. Attempt to give the least amount of information that would permit another capable scientist to replicate your outcome, but be cautious that vital information is integrated. The use of subheadings is suggested and ought to be synchronized with the results section.

When a technique is used that has been well-described in another section, mention the specific item describing the way, but draw the basic principle while stating the situation. The purpose is to show all particular resources and broad procedures so that another person may use some or all of the methods in one more study or referee the scientific value of your work. It is not to be a step-by-step report of the whole thing you did, nor is a methods section a set of orders.

Materials:

Materials may be reported in part of a section or else they may be recognized along with your measures.

Methods:

- Report the method and not the particulars of each process that engaged the same methodology.
- Describe the method entirely.
- To be succinct, present methods under headings dedicated to specific dealings or groups of measures.
- Simplify—detail how procedures were completed, not how they were performed on a particular day.
- If well-known procedures were used, account for the procedure by name, possibly with a reference, and that's all.

Approach:

It is embarrassing to use vigorous voice when documenting methods without using first person, which would focus the reviewer's interest on the researcher rather than the job. As a result, when writing up the methods, most authors use third person passive voice.

Use standard style in this and every other part of the paper—avoid familiar lists, and use full sentences.

What to keep away from:

- Resources and methods are not a set of information.
- Skip all descriptive information and surroundings—save it for the argument.
- Leave out information that is immaterial to a third party.

Results:

The principle of a results segment is to present and demonstrate your conclusion. Create this part as entirely objective details of the outcome, and save all understanding for the discussion.

The page length of this segment is set by the sum and types of data to be reported. Use statistics and tables, if suitable, to present consequences most efficiently.

You must clearly differentiate material which would usually be incorporated in a study editorial from any unprocessed data or additional appendix matter that would not be available. In fact, such matters should not be submitted at all except if requested by the instructor.

Content:

- Sum up your conclusions in text and demonstrate them, if suitable, with figures and tables.
- In the manuscript, explain each of your consequences, and point the reader to remarks that are most appropriate.
- Present a background, such as by describing the question that was addressed by creation of an exacting study.
- Explain results of control experiments and give remarks that are not accessible in a prescribed figure or table, if appropriate.
- Examine your data, then prepare the analyzed (transformed) data in the form of a figure (graph), table, or manuscript.

What to stay away from:

- Do not discuss or infer your outcome, report surrounding information, or try to explain anything.
- Do not include raw data or intermediate calculations in a research manuscript.
- Do not present similar data more than once.
- A manuscript should complement any figures or tables, not duplicate information.
- Never confuse figures with tables—there is a difference.

Approach:

As always, use past tense when you submit your results, and put the whole thing in a reasonable order.

Put figures and tables, appropriately numbered, in order at the end of the report.

If you desire, you may place your figures and tables properly within the text of your results section.

Figures and tables:

If you put figures and tables at the end of some details, make certain that they are visibly distinguished from any attached appendix materials, such as raw facts. Whatever the position, each table must be titled, numbered one after the other, and include a heading. All figures and tables must be divided from the text.

Discussion:

The discussion is expected to be the trickiest segment to write. A lot of papers submitted to the journal are discarded based on problems with the discussion. There is no rule for how long an argument should be.

Position your understanding of the outcome visibly to lead the reviewer through your conclusions, and then finish the paper with a summing up of the implications of the study. The purpose here is to offer an understanding of your results and support all of your conclusions, using facts from your research and generally accepted information, if suitable. The implication of results should be fully described.

Infer your data in the conversation in suitable depth. This means that when you clarify an observable fact, you must explain mechanisms that may account for the observation. If your results vary from your prospect, make clear why that may have happened. If your results agree, then explain the theory that the proof supported. It is never suitable to just state that the data approved the prospect, and let it drop at that. Make a decision as to whether each premise is supported or discarded or if you cannot make a conclusion with assurance. Do not just dismiss a study or part of a study as "uncertain."

Research papers are not acknowledged if the work is imperfect. Draw what conclusions you can based upon the results that you have, and take care of the study as a finished work.

- You may propose future guidelines, such as how an experiment might be personalized to accomplish a new idea.
- Give details of all of your remarks as much as possible, focusing on mechanisms.
- Make a decision as to whether the tentative design sufficiently addressed the theory and whether or not it was correctly restricted. Try to present substitute explanations if they are sensible alternatives.
- One piece of research will not counter an overall question, so maintain the large picture in mind. Where do you go next? The best studies unlock new avenues of study. What questions remain?
- Recommendations for detailed papers will offer supplementary suggestions.

Approach:

When you refer to information, differentiate data generated by your own studies from other available information. Present work done by specific persons (including you) in past tense.

Describe generally acknowledged facts and main beliefs in present tense.

The Administration Rules

Administration Rules to Be Strictly Followed before Submitting Your Research Paper to Global Journals Inc.

Please read the following rules and regulations carefully before submitting your research paper to Global Journals Inc. to avoid rejection.

Segment draft and final research paper: You have to strictly follow the template of a research paper, failing which your paper may get rejected. You are expected to write each part of the paper wholly on your own. The peer reviewers need to identify your own perspective of the concepts in your own terms. Please do not extract straight from any other source, and do not rephrase someone else's analysis. Do not allow anyone else to proofread your manuscript.

Written material: You may discuss this with your guides and key sources. Do not copy anyone else's paper, even if this is only imitation, otherwise it will be rejected on the grounds of plagiarism, which is illegal. Various methods to avoid plagiarism are strictly applied by us to every paper, and, if found guilty, you may be blacklisted, which could affect your career adversely. To guard yourself and others from possible illegal use, please do not permit anyone to use or even read your paper and file.

Criterion for Grading a Research Paper (Compilation)
by Global Journals Inc. (US)

Please note that following table is only a Grading of "Paper Compilation" and not on "Performed/Stated Research" whose grading solely depends on Individual Assigned Peer Reviewer and Editorial Board Member. These can be available only on request and after decision of Paper. This report will be the property of Global Journals Inc. (US).

Topics	Grades		
	A-B	**C-D**	**E-F**
Abstract	Clear and concise with appropriate content, Correct format. 200 words or below	Unclear summary and no specific data, Incorrect form Above 200 words	No specific data with ambiguous information Above 250 words
Introduction	Containing all background details with clear goal and appropriate details, flow specification, no grammar and spelling mistake, well organized sentence and paragraph, reference cited	Unclear and confusing data, appropriate format, grammar and spelling errors with unorganized matter	Out of place depth and content, hazy format
Methods and Procedures	Clear and to the point with well arranged paragraph, precision and accuracy of facts and figures, well organized subheads	Difficult to comprehend with embarrassed text, too much explanation but completed	Incorrect and unorganized structure with hazy meaning
Result	Well organized, Clear and specific, Correct units with precision, correct data, well structuring of paragraph, no grammar and spelling mistake	Complete and embarrassed text, difficult to comprehend	Irregular format with wrong facts and figures
Discussion	Well organized, meaningful specification, sound conclusion, logical and concise explanation, highly structured paragraph reference cited	Wordy, unclear conclusion, spurious	Conclusion is not cited, unorganized, difficult to comprehend
References	Complete and correct format, well organized	Beside the point, Incomplete	Wrong format and structuring

Index

save our planet

Global Journal of Computer Science and Technology

Visit us on the Web at www.GlobalJournals.org | www.ComputerResearch.org
or email us at helpdesk@globaljournals.org

ISSN 9754350