



Research on Big Data Analytics

By Saloni Jain

Lecturer in RG PG College, India

Abstract - This paper gives an insight in the scope of Big Data in the field of Geoscience and which scripting language is acceptable for Big Data. Big Data transforms traditional information to customary conviction of how data should be aggregated, processed, analysed and stored. Big data is rudimentary transfiguring world of science. The large volume of data is posing a great menace on scientists. As data volumes is increasing with time there is complication in transferring Big Data. Thus, it is vital to reinforce sustainable infrastructure, correct analysis of data and reduction in data around Geoscience Big Data. The prominence of the growth on sharing of data has led to new inventions and it's variations.

Keywords: big data, geoscience, python.

GJCST-C Classification : H.2.8



Strictly as per the compliance and regulations of:



Research on Big Data Analytics

Saloni Jain

Abstract- This paper gives an insight in the scope of Big Data in the field of Geoscience and which scripting language is acceptable for Big Data. Big Data transforms traditional information to customary conviction of how data should be aggregated, processed, analysed and stored. Big data is rudimentary transfiguring world of science. The large volume of data is posing a great menace on scientists. As data volumes is increasing with time there is complication in transferring Big Data. Thus, it is vital to reinforce sustainable infrastructure, correct analysis of data and reduction in data around Geoscience Big Data. The prominence of the growth on sharing of data has led to new inventions and its variations.

Keywords: big data, geoscience, python.

I. INTRODUCTION

Big Data or Data Integration is basically related with interoperability of data. Big Data deals with divergent fields such as:

1. Substantial data movement
2. Replication of data
3. Synchrony of data
4. Transmutation of data

Geoscience is the application and exploration of Earth's minerals, soil, water and energy resources. The variability in Earth sciences in any area can be shown in both spatial and temporal variations.

II. ANALYSIS OF BIG DATA

Prior to 2012 U.S was the largest single contributor to global data.

The emerging markets are showing the largest increases in data growth. In 2012, the amount of information stored worldwide exceeded 2.8 zettabytes. By 2020, the total amount of data stored is expected to be 50 times greater than today.

What good is all of this data? Data is raw, unrecognized facts that is in and of itself worthless. Information is potentially valuable concepts based on data. Knowledge is what we understand based on information. Wisdom is effective use of knowledge in decision making.



Figure 1 : Analysis of Big Data

III. LITERATURE REVIEW

There are many studies wherein many scientists have studied Big Data by inventing customized tools have been developed using various scripting languages. An overview of such studies is discussed in this section. Azza Abouzeid et al devised a paper entitled "Ha doop DB: An Architectural Hybrid of Map Reduce and DBMS Technologies for Analytical Workloads" This paper elaborates on how Hadoop DB is able to approach the performance of parallel data systems and how Hadoop works in heterogeneous environments.

Jerome Boulon et al have discussed about "Chukwa: A large-scale monitoring system" used for monitoring and analysing large distributed systems.

Jeffrey Dean et al have elaborated on "MapReduce: Simplified Data Processing on Large Clusters" which is a programming model and is used processing and generating large data sets. Two functions are used: map function and reduce function.

Tom Narock and Pascal Hitzler discussed about "Crowd sourcing Semantics for Big Data in Geosciences Applications" i.e. how semantic algorithms have been used for achieving accurate data.

Sanjay Ghemawat et al discussed on "The Google File System" which is a scalable distributed file system for large distributed data-intensive applications. It enhances the performance while analysing large clusters of data and provides great performance when dealing with large number of clients.

a) Technique used

Big Data can be coded in many different languages such as C, C++, Python. However, most suitable language considered for coding is Python. Python is said to be multi-model programming language. It authorizes programmers to acquire various methodology of programming: object-oriented and structured programming which is fully sustained by Python. Python offers diverse language characteristics which stimulates functional programming and aspect-oriented programming.

There are many factors that favour Python as a language to code for Big Data. In modern times plenty of API's and libraries have been advanced for Python. In research also Python has a lot to implement ranging from networking to GUI development. Thus the interaction among systems has been highly enriched even though it remains a formidable task in many programming languages.

Libraries in Python which are used for Big Data coding are PyDoop and SciPy. PyDoop offers an API for writing Hadoop programs in Python and is used for Map Reduce also.

Pydoop recommends diverse features which are usually not found in other Python libraries for Hadoop like MapReduce library which enables users to combine and partition data sets, easily installed library and can be used freely.

SciPy is an open source library that is offered by Python for all the users aiming to do scientific computations. This library furnish various modules such as ODE(Ordinary Differential Equations), FFT(Fast Fourier Transformation), optimization which finds application in the field of science and engineering.

b) Snapshots of Coding

```

File Edit Format Run Options Windows Help
bigdata1.py - D:/article for rg/bigdata1.py
from itertools import groupby
from operator import itemgetter
def _pick_last(it):
    for t in it:
        yield t[-1]
def mapreduce(data, mapf, redf):
    buf = []
    for line in data.splitlines():
        for ik, iv in mapf("foo", line):
            buf.append((ik, iv))
    buf.sort()
    for ik, values in groupby(buf, itemgetter(0)):
        for ok, ov in redf(ik, _pick_last(values)):
            print ok, ov

```

Figure : Snapshot of Program 1

```

File Edit Format Run Options Windows Help
bigdata2.py - D:/article for rg/bigdata2.py
from bigdata1 import mapreduce
DATA = """ Computer Programming is fun"""
def map_(k, v):
    for w in v.split():
        yield w, 1
def reduce_(k, values):
    yield k, sum(v for v in values)
if __name__ == "__main__":
    mapreduce(DATA, map_, reduce_)

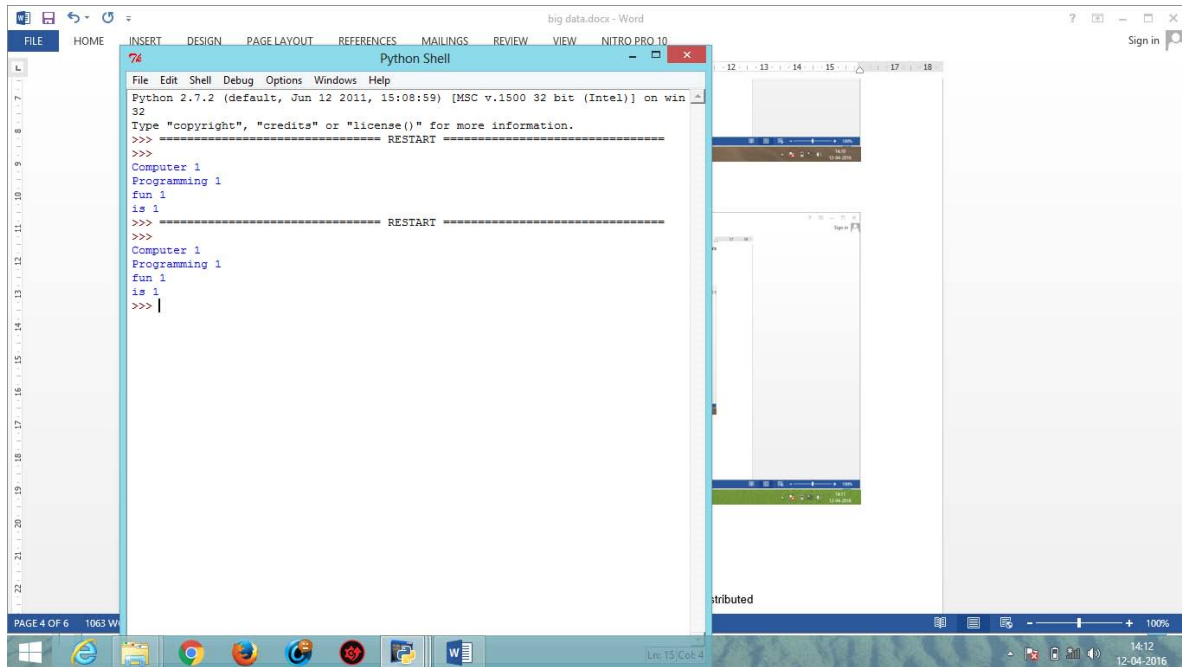
```

Figure : Snapshot of Program 2

c) Conclusion And Future Scope

Big Data analyses large clusters of data and provides users with accurate and refined search. In this

case whatever text the user enters the tool will count the words and displays the output for the user as shown in the snapshot.



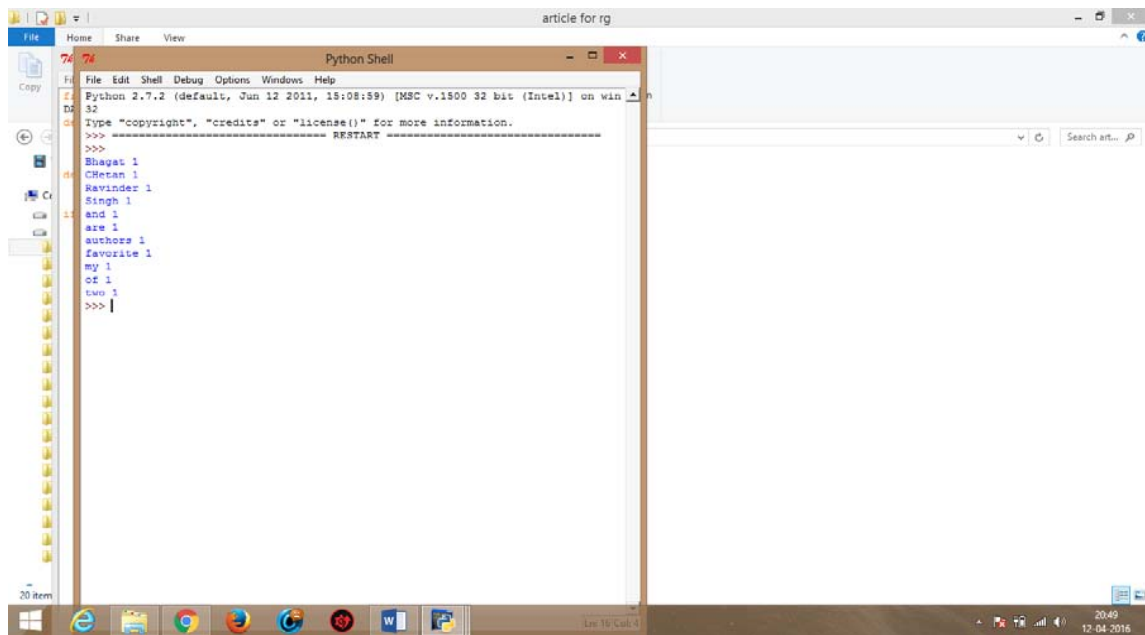
```

Python Shell
File Edit Shell Debug Options Windows Help
Python 2.7.2 (default, Jun 12 2011, 15:08:59) [MSC v.1500 32 bit (Intel)] on win
32
Type "copyright", "credits" or "license()" for more information.
>>> ----- RESTART -----
>>>
Computer 1
Programming 1
fun 1
is 1
>>> ----- RESTART -----
>>>
Computer 1
Programming 1
fun 1
is 1
>>> |

```

Figure : Snapshot of Execution 1

If the user manipulates the text then the output will be modified accordingly as indicated.



```

Python Shell
File Edit Shell Debug Options Windows Help
Python 2.7.2 (default, Jun 12 2011, 15:08:59) [MSC v.1500 32 bit (Intel)] on win
32
Type "copyright", "credits" or "license()" for more information.
>>> ----- RESTART -----
>>>
Bhagat 1
Chetan 1
Ravinder 1
Singh 1
and 1
are 1
authors 1
favorite 1
my 1
or 1
two 1
>>> |

```

Figure : Snapshot of Execution 2

The MapReduce provides a framework where large volumes of data can be analysed. The tool can be extended further by increasing the volume of data supplied as well as some other scripting language can be adopted by the scientists to enhance the power of Big Data and thus make new discoveries in this

discipline. Big Data is emerging as a powerful technique in recent years and provides solutions to the challenges of merging data thus making a mark in manifold fields like banking, health care, education which will involve whole world at large.



REFERENCES RÉFÉRENCES REFERENCIAS

1. Lynch, C., (2008), Big data: How do your data grow?, *Nature*, 455, 28-29, doi:10.1038/455028a.
2. R. Chaiken, B. Jenkins, P.-A. Larson, B. Ramsey, D. Shakib, S. Weaver, and J. Zhou. Scope: Easy and efficient parallel processing of massive data sets. In *Proc. of VLDB*, 2008.
3. G. Czajkowski. Sorting 1pb with mapreduce. googleblog.blogspot.com/2008/11/sorting-1pb-with-mapreduce.html.
4. J. Dean and S. Ghemawat. MapReduce: Simplified Data Processing on Large Clusters. In *OSDI*, 2004.
5. Jeffrey Dean and Sanjay Ghemawat. "MapReduce: Simplified Data Processing on Large Clusters." In *Communications of the ACM*, Volume 51, Issue 1, pp. 107-113, 2008.
6. Matthew L. Massie, Brent N. Chun, and David E. Culler. "The Ganglia Distributed Monitoring System: Design, Implementation, and Experience". In *Parallel Computing* Volume 30, Issue 7, pp 817-840, 2004.
7. Luiz A. Barroso, Jeffrey Dean, and Urs Holzle. Web search for a planet: The Google cluster architecture. *IEEE Micro*, 23(2):22-28, April 2003.
8. Miller, H. G., and P. Mork, (2013), From Data to Decisions: A Value Chain for Big Data, *IT Professional*, vol. 15, no. 1, pp. 57-59, doi:10.1109/MITP.2013.11

