Global Journal of Computer Science and Technology: H Information & Technology
Volume 16 Issue 4 Version 1.0 Year 2016
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Big Data Analysis of Salary Dataset using Hive

By Ishan Fafadia
California State University Los Angeles

Abstract- One way to understand how a city government works is by looking at who it employs and how its employees are compensated. This data contains the names, job title, and compensation for San Francisco city employees on an annual basis from 2011 to 2014. The analyzed data will be shown in the form of various charts and graphs with respect to 1. Yearly Mean Pay, 2. Mean Pay by Job Type, 3. Pay based on Base Pay, Overtime Pay, Other Pay and Benefits. As the Salary seeking population grows, the data also grows in size. This becomes a challenge for the traditional RDBMS to manage the huge volumes of data. Hence Salary data Analysis can be made using Hive and Map Reduce algorithms to eliminate the challenges faced by the traditional RDBMS.

GJCST-H Classification: C.2.1,C.2.3

Strictly as per the compliance and regulations of:

Big Data Analysis of Salary Dataset using Hive

Ishan Fafadia

Abstract- One way to understand how a city government works is by looking at who it employs and how its employees are compensated. This data contains the names, job title, and compensation for San Francisco city employees on an annual basis from 2011 to 2014. The analyzed data will be shown in the form of various charts and graphs with respect to 1. Yearly Mean Pay, 2. Mean Pay by Job Type, 3. Pay based on Base Pay, Overtime Pay, Other Pay and Benefits. As the Salary seeking population grows, the data also grows in size. This becomes a challenge for the traditional RDBMS to manage the huge volumes of data. Hence Salary data Analysis can be made using Hive and Map Reduce algorithms to eliminate the challenges faced by the traditional RDBMS.

- We have observed the drop of budget allocation of salaries in San Francisco.
- There were some departments which didn't provide any benefits to their employers.
- For some departments, even if the employer had worked overtime they were not paid for their extra work.
- Good thing that we observed is that there was no gender discrimination among the department.

I. Introduction

A standout amongst the most well-known datasets urban areas ordinarily discharge is their compensation structure.

Thus we grabbed the dataset of San Francisco as it is the most essential city for any graduate understudy. What's more, we discovered some conceivably intriguing edges of investigation:

1. How has pay rates changed after sometime between various Departments of individuals?
2. How are base pay, extra minutes pay, and advantages apportioned between various gatherings?
3. Is there any proof of pay separation taking into account sexual orientation in this dataset?
4. How spending plan is distributed in light of various Department and obligations?
5. And In this project we have focused on the payment structure of the considerable number of divisions and attempt to give the answer for low paying office.

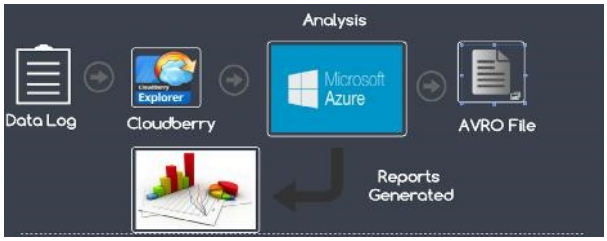
Hadoop is an open source, Java-based programming structure that backings the handling and capacity of to a great degree substantial information sets in a disseminated figuring environment. Hadoop makes it conceivable to run applications on frameworks with a huge number of product equipment hubs, and to handle a large number of terabytes of information.

Author: e-mail: ifafadi@calstatela.edu

Apache Hive is an information distribution center framework based on top of Hadoop for giving information synopsis, query, and analysis. Hive gives a SQL-like interface to inquiry information put away in different databases and document frameworks that incorporate with Hadoop.

II. Work Flow

Initially a data set with Employee_Id, EmployeeName, JobTitle, BasePay, OvertimePay, Other Pay,Benefits, TotalPay, TotalPayBenefits, Year, Notes, Agency,Status is taken from an authentic source. As a next step, this comma separated file has to be uploaded to the cloud. This is done with the help of cloud berry explorer. And data is converted to Avro format.



a) Data Storage

We changed over our information in Avro Format and we utilize that same information we put away in cloudberry explorer and we are utilizing avro in light of the fact that, Avro is one of the favored information serialization frameworks in view of its language lack of bias.

Because of absence of language versatility in Hadoop writable classes, Avro turns into a characteristic decision as a result of its capacity to handle various information designs which can be further prepared by different languages.

b) Conversion to Avro Format

To change over csv information to Avro information utilizing Hive we have to take after the progressions beneath:

1. Make a Hive table put away as content document and indicate your csv delimiter too.
2. Load csv document to above table utilizing "load data" command.
3. Make another Hive table utilizing AvroSerDe.
4. Embed information from previous table to new Avro Hive table utilizing "insert overwrite" command.

c) Data Representation

In this Project, we have considered four main parameters to Analyze the data.

1. Change of Maximum Total Pay in 4 Years.
2. Change of Mean Pay Yearly
3. Mean Pay of each Department
4. Benefits given in each Year.
5. Payment Structure of each Department.

These data were obtained by writing suitable queries in Hive QL.

1. Select Year, max (TotalPay) from avro_table group by Year order by 1
2. Select year, percentile (cast (Totalpay as bigint), 0.5), count (*) Records from avro_table group by year order by 1;
3. Select Job Type, percentile (cast (Totalpay as bigint), 0.5), count (*) as Records from avro_table; JobType group by JobType;
4. Select Year, Sum (Benefits) from avro_table group by year order by 1
5. select JobType, cast(avg(Basepay) as bigint), cast (avg(Overtimepay) as bigint),cast(avg(Otherpay) as bigint), cast(avg(Benefits) as bigint) from avro_table JobType group by JobType;

III. Data Analysis

a) Change of Maximum Total Pay in 4 Years

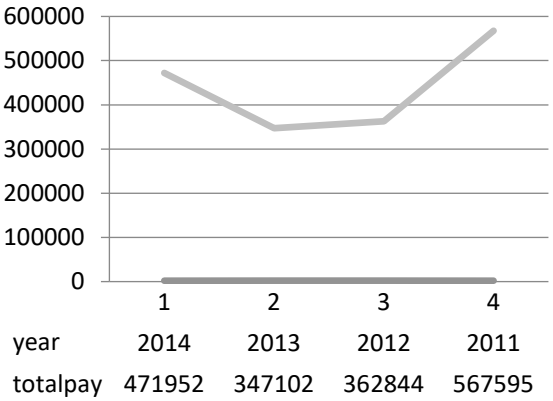


Figure 1

From Figure 1, it's clearly shown that there was a drastic changes of Maximum Total pay from the year 2011 to 2012 which continues till 2013 but it was risen in the year 2014 .From this we can inferred that in year 2012 and 2013 there was slight recession and because of that payment structure of employees were not increase.

b) Change of Mean Pay Yearly

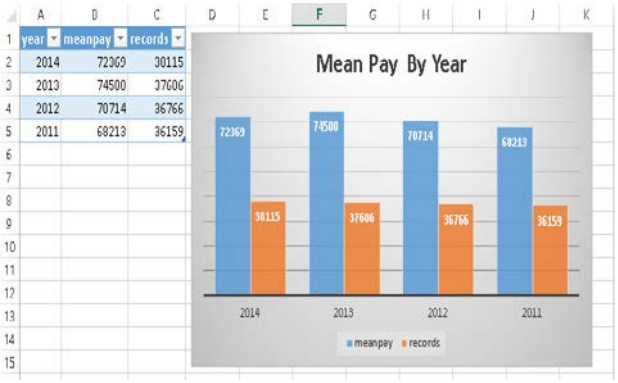


Figure 2

From Figure 2 we come to know that the mean pay was increased leaving the fact that maximum total pay was decreased and this could be possible because the whole budget was well distributed among the employees and number of employees was also increased so San Francisco hired more employees from 2011 to 2014 so because of that more people were employed and benefited And there was drop of budget allocation from 2011 to 2014.

c) Mean Pay of each Department

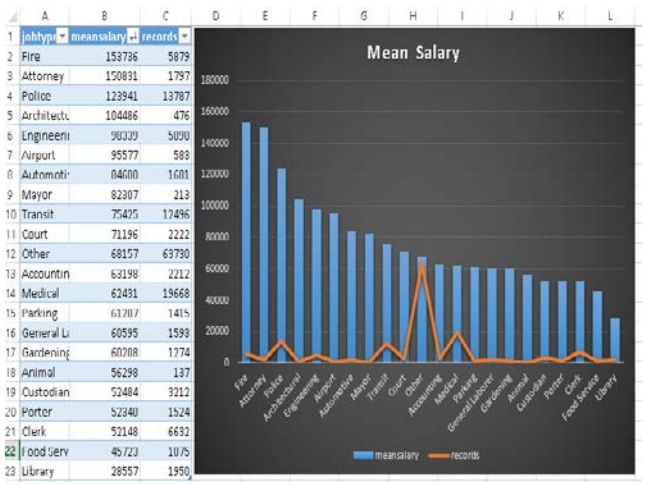


Figure 3

In Figure 3, we have shown Job Type, Mean pay & records.

And to find the best paying department we need to find the Mean pay for each department and the number of records in it.So here we found that fire department has the highest Mean pay.

And we know that in any city Fire department is the most important group of professionals as they serve for day and night at any situation and library department is the least paying department. And one thing we found surprising that medical department is also not a good pay master and because of this less people are interested in taking medical as their career

d) Benefits given in each Year

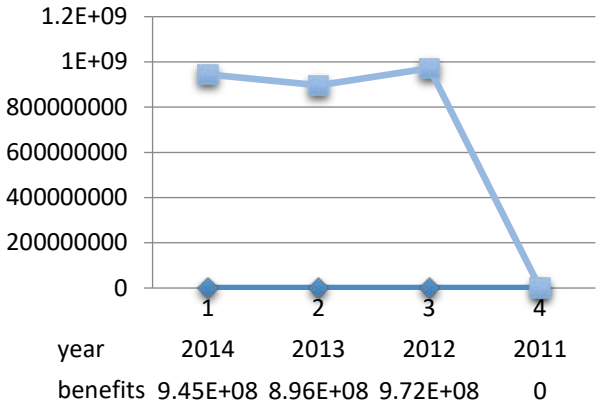


Figure 4

In Figure 4, we try to find the benefits that was given to employees in all the year and we find that in the year 2011 no benefits were given by any department and this could be reason of least mean pay in the that year

But by the year 2012 there was the added pay in the name of benefits to the payment structure of each department to lure more employees and provide better living standard to the people.

Payment Structure of each Department

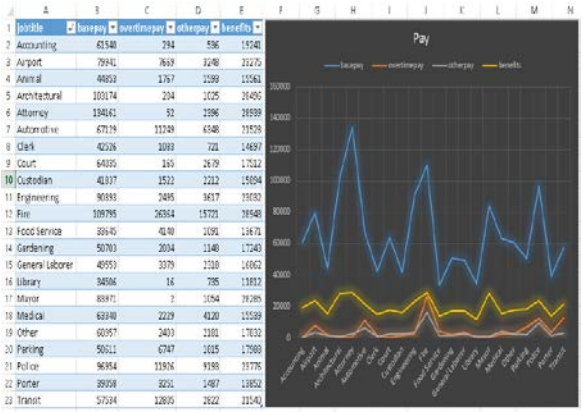


Figure 5

Figure 5 shows the basepay, overtimepay, otherpay & benefits.

From this we come to know that Attorney has got the highest base pay and Food Service got the lowest. But there was not so much difference in Overtime,Otherpay and Benefits.

IV. Business Solution

All the departments should have better structure for overtime pay & other pays.

We can reallocate the budget to the smaller departments.

We can reduce the payment structure of Attorney, Police and Fire.

And to encourage employees we should provide the equality among all people regarding their post and stature Example:

As we all the Engineering department is also very nowadays because of growth of computers in every field so to increase the mean pay of that department we can increase the overtime pay and other pay which can make them in top 4 earning department of San Francisco.

And medical department is deep low in the mean pay so to uphill their department we can increase the base pay which is very low in terms of their dedication and risk in their works and by doing that we can encourage more people to join the medical line in future.

For some departments, even if the employer had worked overtime they were not paid for their extra work. And from this we come to know that main reason of fire department having best mean pay is there they have the best structure of Overtime pay, Benefits, Other pay.

V. Conclusion

We have observed the drop of budget allocation of salaries in San Francisco.

There were some departments which didn't provide any benefits to their employers.

For some departments, even if the employer had worked overtime they were not paid for their extra work.

Good thing that we observed is that there was no gender discrimination among the department.

And Fire Department is the best in the San Francisco area.

VI. Future Work

With the dataset we had, we analyzed the salaries based on different departments. But if we had bigger data i.e. if we had data of last 15-20 years we would have more precisely provided results about departmental salaries.

And with more precise data we could have shown some better solution for the employees working in their respective departments and for the departments as well And seeing the future prospect of our analysis we can say that San Francisco government can use this to decide all the future payment structure of all a departments to provide better life and better living standards for people And seeing the future prospect of our analysis we can say that San Francisco government can use this to decide all the future payment structure of all a departments to provide better life and better living standards for people

Github Code:
https://github.com/saket18/sfsalariesanalysis

Dataset URL:
https://www.kaggle.com/kaggle/sf-salaries

References Références Referencias

1. https://hadoop.apache.org/ https://hadoop.apache.org/
2. https://hive.apache.org/
3. http://hortonworks.com/hadoop-tutorial/how-to-processdata-with-apache-hive/
4. http://blog.cloudera.com/wp-content/uploads/2010/01/6IntroToHive.pdf
5. http://blog.cloudera.com/blog/2013/04/demoanalyzing-data-with-hue-and-hive/
6. https://azure.microsoft.com/enus/documentation/articles/hdinsight-hadoop-tutorial-getstarted-windows/
7. https://azure.microsoft.com/enus/documentation/articles/hdinsight-connect-excel-hiveodbc-driver/
8. https://azure.microsoft.com/en-us/documentation/articles/hdinsight-connect-excel-powerquery/
9. https://hadoop.apache.org/docs/r1.2.1/mapred_tutorial. html
10. http://www.infoworld.com/article/2683729/hadoop/10ways-to-query-hadoop-with-sql.html

Global Journals Inc. (US) Guidelines Handbook 2016

www.GlobalJournals.org