Global Journals LaTeX JournalKaleidoscope™
Artificial Intelligence formulated this projection for compatibility purposes from the original article published at Global Journals. However, this technology is currently in beta. *Therefore, kindly ignore odd layouts, missed formulae, text, tables, or figures.*

Big Data Analysis of Salary Dataset using Hive

Ishan Fafadia[1]

[1] California State University Los Angeles

Received: 12 December 2015 Accepted: 4 January 2016 Published: 15 January 2016

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.

Index terms—

1 Information & Technology

Global Journal of Computer Science and Technology: H 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.

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.

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.

2 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.

3 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.

To change over csv information to Avro information utilizing Hive we have to take after the progressions beneath: 1. Make a Hive 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

From Figure **??** 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.

4 d) Benefits given in each Year

Figure **??** In Figure **??**, 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.

5 Payment Structure of each Department

Figure **??** Figure **??** 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.

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.

6 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.

7 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 [1]

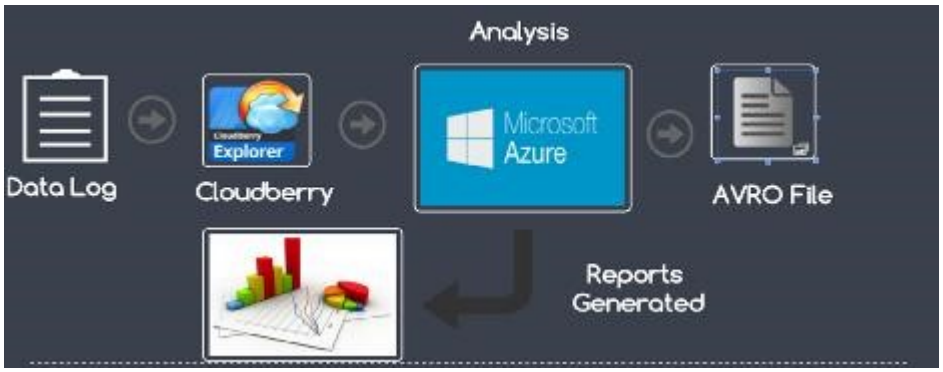
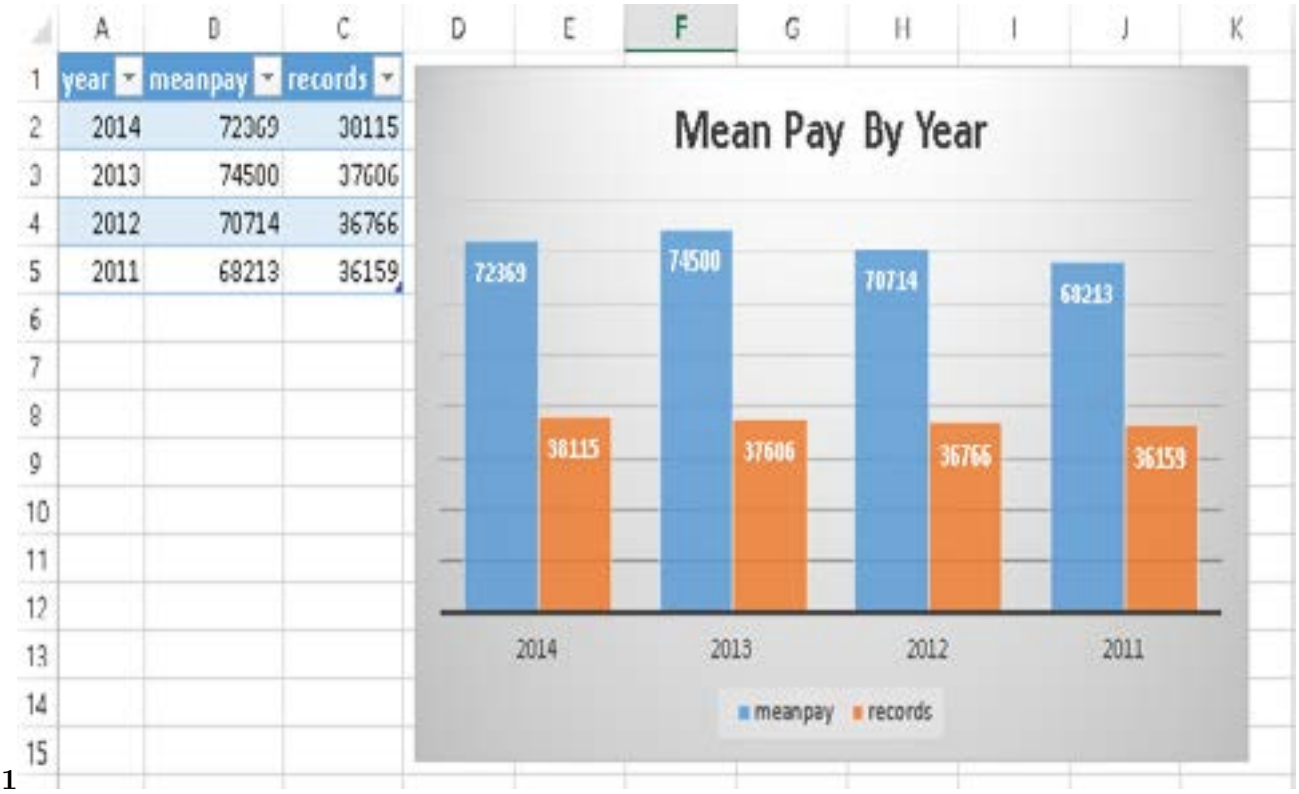


Figure 1:



1

Figure 2: Figure 1 From

[1] © 2016 Global Journals Inc. (US) 1

	A	B	C
1	jobtyp	meansalary	records
2	Fire	153736	5879
3	Attorney	150831	1797
4	Police	123941	13787
5	Architectu	104486	476
6	Engineeri	98339	5090
7	Airport	95577	583
8	Automoti	84600	1601
9	Mayor	82307	213
10	Transit	75425	12496
11	Court	71196	2222
12	Other	68157	63730
13	Accountin	63198	2212
14	Medical	62431	19668
15	Parking	61207	1415
16	General L	60595	1593
17	Gardening	60208	1274
18	Animal	56298	137
19	Custodian	52484	3212
20	Porter	52340	1524
21	Clerk	52148	6632
22	Food Serv	45723	1075
23	Library	28557	1950

Figure 3:

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.

[Note: 4. Embed information from previous table to new AvroHive table utilizing ”insert overwrite” command.c) Data RepresentationIn this Project, we have considered four main parameters to Analyze the data.]

Figure 4: table put