Effective Visual Big Data Processing with Machine Learning Methodologies

B Kranthi Kiran

Abstract: ***Development of high web utilization made business procedures in a difficult manner. In request to dissect the online business un-organized and gigantic measure of information is unimaginable with the Traditional frameworks. Recent innovations propel the strategies for examination are made to break down a lot of the information utilizing the Big Data Techniques, and to improve the adaptability and the precision of investigating the business methodologies, it has actualized on Hadoop with parallel preparing. This paper presents the experimental study on IBM real time data of one lakh records for demonstrating the efficiency of proposed Hadoop based distributed query processing technique.***

Keywords: Big data techniques, query processing, Hadoop, distributed processing

I. INTRODCUTION

Data collection plays an important role in any research. The data which is used for any research can be segregate. The techniques of data gathering or somebody already existing data can be use, only it will provide the purpose of your research documents. Before you start data analysis [1] the main priority is given to the valid data. So, the data which is collected needs to verify whether the data collected is correct or not. Analysis of wrong data may lead to wrong predictions [2]. After data collection, accurate analysis needs to be performed on the data collected.

For any research to be performed, data analysis plays a vital role since it explains various theories, concepts, methods [3] and frame works which are used. Data analysis gradually provides a solution in arriving at the conclusions and it also helps in proving the hypothesis which was predicted.

Data analysis is the way or process to inspect, clean, transform and model the data with the idea of discovering outcome for a given purpose or getting the useful information. It also helps in taking decisions for business more scientifically, profit or loss, help them work more effectively and new strategies can be planned based on the proper data analysis.

Data analysis is two types: Qualitative and Quantitative. Analysis can be used on the data is determined based on the type of data.

Qualitative Analysis [4]: In this type, the analysis is done on non-numerical data like text or individual words, etc. Quantitative Analysis [5]: This analysis mainly focuses on data measurement and it uses mathematical science to unveil conclusions and results.

The output can be numeric. In a few scenarios, both outcome of analysis is used with one in one, for eg: For Qualitative conclusions we use quantitative analysis. The existing uses the Hadoop [6] and Hadoop’s environment [7] for the data generation and respective query processing and It is shown in Fig. 1.

This persistent system [8] uses the Hadoop and Hadoop environment and it does the great work with large datasets and produces the high accuracy in analyzing the data.

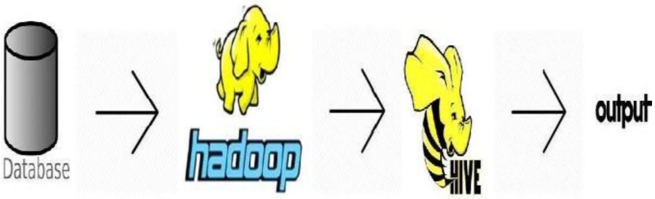


Fig. 1. Query Processing in Hadoop Environment

However, there is one drawback of this model is that it cannot use any of the visualization tool for proper analysis of the data. Thus, this paper is Focused on this problem, and it upgrade the existing system with a new visualization method. View of proposed system is shown in Fig. 2.

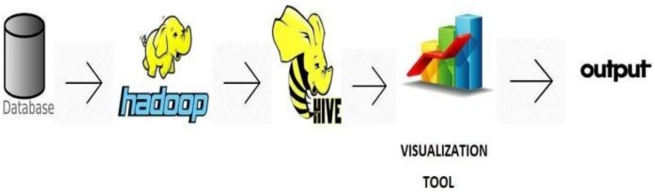


Fig. 2: Proposed Query Processing Visualized System

Remaining part of the paper is organized as follows: Section 2 describes the related work, Section 3 presents the proposed work, and Section 4 illustrates the experimental study, Section 5 conclusion and future scope of the work.

Revised Manuscript Received on October 22, 2019.

B Kranthi Kiran, Associate Professor in department of CSE at JNTUH University, Kukatpally Hyderabad.

Retrieval Number: K131610812S19/2019©BEIESP
DOI: 10.35940/ijitee.K1316.10812S19

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

II. RELATED WORK

Hadoop Big Data Techniques [9] and Spark [10] permits the scattering of measurements and the representations of execution and the factual Reports and assessment in close to continuous information.

The paper likewise triggers more extensive exchanges concerning future research difficulties and openings in principle and practice. Generally speaking, the finishes of the investigation integrate various big data analytics (BDA) [11] considerations (e.g., clarification of huge information, types, nature, corporate worth, and suitable hypotheses) that convey further experiences alongside the cross-cutting examination applications in online business.

We show that the internet business area can offer quite a few parts for fruitful information mining and guarantee that it is the executioner territory for information mining. We assign a coordinated engineering, grounded as far as we can tell at Blue Martini Software [12], for supporting this combination. The design can dramatically decrease the pre-preparing, spring-cleaning, and information thoughtful exertion frequently recorded to take 70% of the time in data revelation ventures. We accentuate the requirement for information accumulation at the application server layer (not the webserver) to sustenance logging of data and metadata that is imperative to the revelation method. We assign the information change scaffolds required from the exchange allotment frameworks and client occasion streams (e.g., clickstreams) to the information distribution center. We detail the mining work surface, which solicitations to convey various assessments of the data through news coverage, information mining calculations, representation, and OLAP [13]. We achieve a lot of difficulties.

With regards to information investigation, you can discover answers to pretty much every business cross examination. From how much income did we make this quarter to where are a large portion of our traffic originating from, examining your information can be a savvy methodology. It's tied in with realizing what the data is, however progressively about expressive how to manage it.

2.1 Using Hadoop eco system for the storage of the large amount of the data [14].

Here we are using the Hadoop ecosystems for the analysis of the ecommerce data which will helps for the improvement of the business and the prediction of the business analysis to the escalation issues, Hadoop helps in the processing of the unstructured large amount of the data for the preprocessing and let the data analysis to improve the immediate results over there in this part it used the MAPPER and the reducers for the pre-processing of the data and send it to the hive metadata for the storage of the index content over there. Using the open source tools like hive and the Hadoop components we are going to perform the data analysis in this prediction business analysis.

2.2 Pre-processing of the data, unstructured data to structured data [15]

For the analysis of the data we need to take the data which is taken from the online website is un-structured data it cannot be processed over the hive meta data tables and it will be available in JSON Format, where Hadoop accepts different types of the data formats which will helps to a readable formats, here it will takes the two types of the classifications in the comparison of the data and the processing of the data where the data can be supervised and un-supervised data , where supervised data is having the data sets already built in and the un-supervised data sets does not have the preprocessed data sets which will be available as a corpus data sets area not available in this datasets.

2.3 Analyzing the data of the pre-processed datasets and moved to HIVE [16]

After the preprocessing of the unstructured data sets into structured data sets then the data can be sent to the hive meta data tables and the bucketing and the indexing will be performed in this data sets, where these will increase the improvement of the performance and help in the reduction of the time taking in this data retrieving analyzing, here for the AD-hoc querying the data can be performed and then the immediate reports can be generated in this hive tables, which will helps in the retrieving of the data from the hive meta data tables on immediate basis, When the reports generating needs, hive works as AD-HOC querying to generate the reports as same as the traditional database like MYSQL , but when we compare the hive with the traditional data bases , here the large amount of the data cannot be loaded in the traditional databases, where these works for the limited amount of the data can be loaded in this.

III. PROPOSED WORK

After the data is analyzed by performing hive queries Hive installed on Hadoop environment, it is imported into the visualization tool. The proposed system uses a visualization tool to resolve the issue of visualizing the analyzed data. Analyzed data can be visualized in any of the charts like Column, Line, Pie, Doughnut, Area, XY (Scatter), Stock, Radar, Combo, Bar, Bubble, Surface, etc.

The common arrangement of prerequisites characterized in any working framework or programming application is the physical PC assets, which is known as equipment, an equipment necessities rundown is as often as possible joined by an equipment similarity list, particularly on account of working frameworks. A HCL records tried, appropriate, and some of the time unsuited equipment gadgets for an exact working framework or application.

Hadoop can process any types of unstructured and structure information, giving clients more straightforwardness for collecting, preparing and examining the information than some other information stockrooms and social information product houses gave. It contains a few parts that permit the capacity and handling of enormous information limits in a grouped situation.

Using this visualization tools we can showcase the data into different formats and provides best charts. Using this hive it can connect directly to the twitter account and then connect using the API calls, which can retrieve the data easily using the flume and other kafka streaming tools, using hive we can connect it with the mysql meta data SQL speculation to incorporate SQL-like queries (Hive QL) into the first Java without the essential to actualize questions in the low-level Java API. Since most information warehousing applications work with SQL-based questioning dialects, Hive helps transportability of SQL-based applications to Hadoop. After obtaining the examined data from the above units it is imported to imagining tool for the purpose of visualizing. Here, the visualization tool we used is Microsoft Excel 2007. You can be displaying your data analysis reports in various ways in Excel. However, if your data analysis results can be visualized as graphs that highlight the notable points in the data, your audience can rapidly grasp what you want to project in your data. It also leaves a good effect on presentation style.

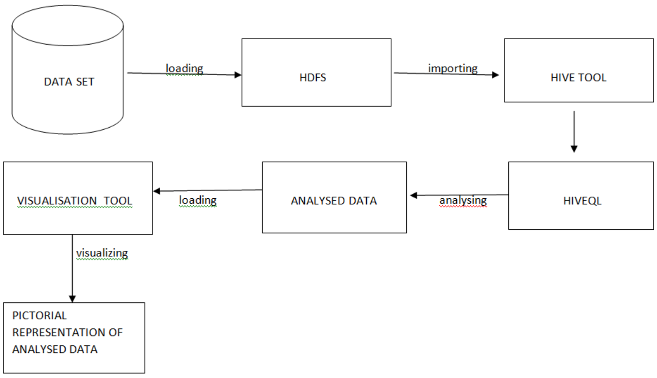


Fig. 3. Proposed Framework

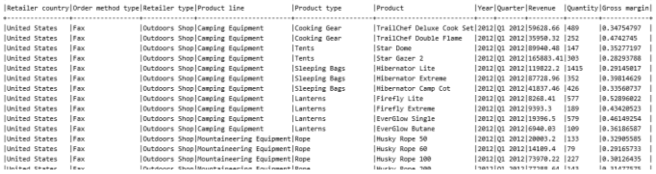
Huge synthetic dataset is generated and stored into the databse and processed in HDFS file system in compressed format of the data. Distributed HIVE queries are processed for faster execution of proposed system and the datasets are anlyzed with the interface of some visualization tools for accessing the required information for processing systems and it is neatly mentioned in the proposed framework, which shown in Fig. 3.

IV. EXPERIMENTAL RESULTS

Proposed system is implemeted within the distributed system using the framework of Hadoop, Hive and Visualization tool, which discussed in the earlier section.

Hadoop permits distributed fashion of programming and its open source structure is executed by Java language; Data and its structure is created by HIVE queries and this proposed system has the capacity of taking huge information and run the big applications on a Hadoop bunched frameworks. It is at the in the midst of a developing environment of huge information innovations that are chiefly used to arrangement progressed investigation devices, including all prescient examination, AI and information mining applications.

Hive give a bucketing system and provides various data processing approaches which can deals with the analysis of the data and using the hive we can connect with the number of visualization tools like tableau, Qlikview. Using this visualization tools we can showcase the data into different formats and provides best charts. Using this hive it can connect directly to the twitter account and then connect using the API calls, which can retrieve the data easily using the flume and other Kafka streaming tools, using hive we can connect it with the MySQL meta data SQL speculation to incorporate SQL-like queries (Hive QL) into the first Java without the essential to actualize questions in the low-level Java API.



```
|Retailer country|Order method type|Retailer type|Product line            |Product type    |Product                 |Year|Quarter|Revenue  |Quantity|Gross margin|
+----------------+-----------------+-------------+------------------------+----------------+------------------------+----+-------+---------+--------+------------+
|United States   |Fax              |Outdoors Shop|Camping Equipment       |Cooking Gear    |TrailChef Deluxe Cook Set|2012|Q1 2012|59628.66 |489     |0.34754797  |
|United States   |Fax              |Outdoors Shop|Camping Equipment       |Cooking Gear    |TrailChef Double Flame  |2012|Q1 2012|35950.32 |252     |0.4742745   |
|United States   |Fax              |Outdoors Shop|Camping Equipment       |Tents           |Star Dome               |2012|Q1 2012|89940.48 |147     |0.35277197  |
|United States   |Fax              |Outdoors Shop|Camping Equipment       |Tents           |Star Gazer 2            |2012|Q1 2012|165883.41|303     |0.28293788  |
|United States   |Fax              |Outdoors Shop|Camping Equipment       |Sleeping Bags   |Hibernator Lite         |2012|Q1 2012|119822.2 |1415    |0.29145017  |
|United States   |Fax              |Outdoors Shop|Camping Equipment       |Sleeping Bags   |Hibernator Extreme      |2012|Q1 2012|87728.96 |352     |0.39814629  |
|United States   |Fax              |Outdoors Shop|Camping Equipment       |Sleeping Bags   |Hibernator Camp Cot     |2012|Q1 2012|41837.46 |426     |0.33560737  |
|United States   |Fax              |Outdoors Shop|Camping Equipment       |Lanterns        |Firefly Lite            |2012|Q1 2012|8268.41  |577     |0.52896022  |
|United States   |Fax              |Outdoors Shop|Camping Equipment       |Lanterns        |Firefly Extreme         |2012|Q1 2012|9393.3   |189     |0.43420523  |
|United States   |Fax              |Outdoors Shop|Camping Equipment       |Lanterns        |EverGlow Single         |2012|Q1 2012|19396.5  |579     |0.46149254  |
|United States   |Fax              |Outdoors Shop|Camping Equipment       |Lanterns        |EverGlow Butane         |2012|Q1 2012|6940.03  |109     |0.36186587  |
|United States   |Fax              |Outdoors Shop|Mountaineering Equipment|Rope            |Husky Rope 50           |2012|Q1 2012|20003.2  |133     |0.32905585  |
|United States   |Fax              |Outdoors Shop|Mountaineering Equipment|Rope            |Husky Rope 60           |2012|Q1 2012|14109.4  |79      |0.29165733  |
|United States   |Fax              |Outdoors Shop|Mountaineering Equipment|Rope            |Husky Rope 100          |2012|Q1 2012|73970.22 |227     |0.30126435  |
|United States   |Fax              |Outdoors Shop|Mountaineering Equipment|Rope            |Husky Rope 200          |2012|Q1 2012|77288.64 |143     |0.31477575  |
```

Fig. 4. Queary processing of 1 lakh IBM real time records

After setting up the environment and loading data into database, we can now execute queries according to user’s requirement. For the process of execution, we need a dummy data, this dummy data is extracted from the IBM real time data. The data consists around 1 lakh records. We are processing and analyzing the data of 1 Lakh. This data is processed using different queries and Spark. The results are shown in Fig. 4

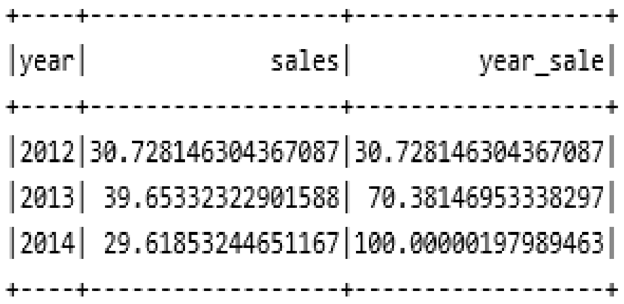
Fig.4 shows all the retailer information. As we can figure out that the query results consist of 11 columns. Each Column consists of its related data like the Retailer Country Column which describes the country of the retailer, Order Method Type Column which describes the mode through which order is placed, Retailer Type Column which describes the type of retailer where the product is being sold. Product Line Column which describes the category product comes in. Product Type Column defines what the type of product which is on sale is. Product Column gives the name of the product; Year Column defines the year in which sales need to be analyzed. Quarter Column also describes Quarterly based analysis; Gross margin describes the margin of the gross product.

Retrieval Number: K131610812S19/2019©BEIESP
DOI: 10.35940/ijitee.K1316.10812S19

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

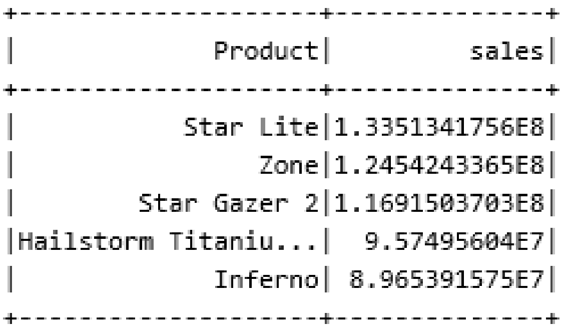
Next we are getting the distinct years from the dummy data and calculating analysis in the given years. We have large amount of data in the all the years which is quite difficult to calculate so we are getting the distinct so it will be easy to calculate the product analysis in particular years' profits and losses, ups and downs, reasons everything will be cleared if we have distinct years. Accordingly, we are getting all the details and then importing all the functions. We are going to get the total revenue. Dummy data is uploaded to csv file and the data is retrieved from the csv file. Based on yearly sales the total revenue is calculated. These gives the total revenue is calculated. This is all calculated by entering the query in the spark and then the total revenue is done.

Here we are calculating, that how much revenue is collected by selling all the products in a year So that business people will get an idea that how much revenue is collected in a year so that there can be idea whether there is a profit or loss in the complete sales.



```
+----+-----------------+-----------------+
|year|            sales|        year_sale|
+----+-----------------+-----------------+
|2012|30.728146304367087|30.728146304367087|
|2013| 39.65332322901588| 70.38146953338297|
|2014| 29.61853244651167|100.00000197989463|
+----+-----------------+-----------------+
```

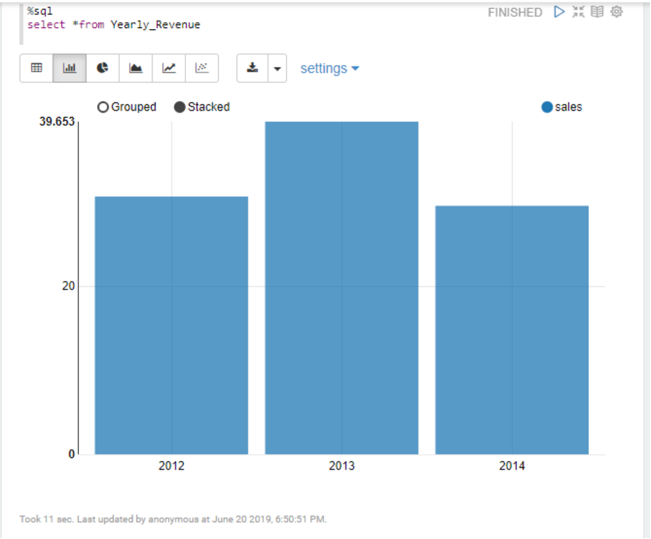
Fig. 5 Sales Information Using Proposed System



```
+--------------------+-------------+
|             Product|        sales|
+--------------------+-------------+
|           Star Lite|1.3351341756E8|
|                Zone|1.2454243365E8|
|        Star Gazer 2|1.1691503703E8|
|Hailstorm Titaniu...|  9.57495604E7|
|             Inferno| 8.965391575E7|
+--------------------+-------------+
```

Fig. 6 Products and Sales Information Using Proposed System

Accessing of products and sales from the big IBM data is effectively performed and respective results are presented in Fig. 5 and Fig. 6. Revenue visialization analysis and sales analysis in our proposed system are shown in Fig.7 and Fig. 8 respectively



```
%sql
select *from Yearly_Revenue
```

Took 11 sec. Last updated by anonymous at June 20 2019, 6:50:51 PM.

Fig. 7. Year wise Revenue Analysis in Proposed Visualization Model

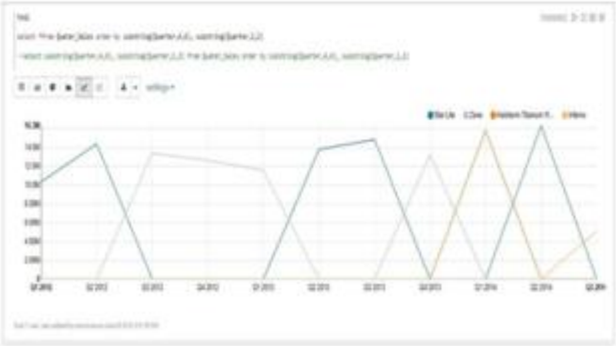


Fig. 8: Effective Quarterly Sales

We are using spark in our proposed system for faster execution queries and visualization of results analysis.

V. CONCLUSION AND FUTURE SCOPE

Present big data systems are effectively works for query processing of large amount of data. The present systems are used the Hadoop framework for the supporting of distributed processing. However, it has the limitation is that it can process the queries in a manner of faster execution and it needs to be visualization for effective big data analysis. This paper presents the proposed visualized big data framework, which supports both the effective distributed query processing as well as visual analytics to an end user. Future scope of the work is to extend proposed system for improving the scalability feature while to run millions of high-dimensional big data.

REFERENCES

1. Liang Wang, Christopher Leckie, Kotagiri Ramamohana rao, James Bezdek, "Automatically determining the number of clusters in unlabeled data sets", IEEE Transactions on knowledge and Data Engineering, Vol.21, Issue. 3,
2009: 335-350

2. Bolshakova N, Azuaje F (2003) Cluster validiation techniques for genome expression data. Sig Process 83:825–833
3. Jacalyn M Huband, James C Bezdek, Richard J Hathaway," big VAT: Visual assessment of cluster tendency for large data sets", Pattern Recognition, Vol.38, Issue.11, 2005
4. Jain AK, Murthi MN, Flynn PJ (1999) Data Clustering: Review. ACM Comput Surv 31(3):266–320
5. Lovasz L, Plummer M (1986) Matching theory. Akade´miai Kiado´, Budapest Nguyen DT (2012) Clustering with multi-viewpoint based similarity measure. IEEE Trans Knowl Data Eng 24(6):988–1001
6. M Suleman Basha, S K Mouleeswaran, K Rajendra Prasad, "Cluster Tendency Methods for Visualizing the Data Partitions" International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-X, Issue-X, July 2019
7. G. Gautam and D. Yadav, "Sentiment analysis of twitter data using machine learning approaches and semantic analysis," *2014 Seventh International Conference on Contemporary Computing (IC3)*, Noida, 2014, pp. 437-442.
8. Y. Singh, P. K. Bhatia, and O.P. Sangwan, "A Review of Studies on Machine Learning Techniques," International Journal of Computer Science and Security, Volume (1) : Issue (1), pp. 70-84, 2007
9. A. Sunita. B, Ather, and Anita, R. Kulkarni "Hadoop MapReduce: A programming Model for large Data Processing", American Journal of computer Science and Engineering Survey, AJCS, pp. 001-010, 2014
10. D. Jinto.T, Pawan. M, "Efficient Resource Utilization in Hadoop on Virtual Machine", International Journal of computer science and Mobile computing, vol.4, 2, pg. 965-569, 2015
11. P. Merla and Y. Liang, "Data analysis using hadoop MapReduce environment," *2017 IEEE International Conference on Big Data (Big Data)*, Boston, MA, 2017, pp. 4783-4785.
12. Kiran kumara Reddi & Dnvsl Indira "Different Technique to Transfer Big Data : survey" IEEE Transactions on 52(8) (Aug.2013) 2348 { 2355}
13. Bernice Purcell "The emergence of "big data" technology and analytics" Journal of Technology Research 2013
14. Ahmed Eldawy, Mohamed F. Mokbel "A Demonstration of Spatial Hadoop: An Efficient MapReduce Framework for Spatial Data" Proceedings of the VLDB Endowment, 2013
15. Thashmee Karunaratne ; Henrik Bostrom ; Ulf Norinder, "Pre-Processing Structured Data for Standard Machine Learning Algorithms by Supervised Graph Propositionalization - A Case Study with Medicinal Chemistry Datasets", Ninth International Conference on Machine Learning and Applications, 2010
16. D.Surekha, G. Swamy, S Venkatramaphani Kumar, "Real time streaming data storage and processing using storm and analytics with Hive", International Conference on Advanced Communication Control and Computing Technologies (ICACCCT), 2016.

AUTHOR PROFILE

Dr. B. Kranthi Kiran received B.Tech in Computer science and Engineering, M.Tech in Computer science and Engineering at Osmania University and PhD in Data Mining from JNTUH University. He is currently working as Associate Professor in department of CSE at JNTUH University, Kukatpally Hyderabad. His research includes Databases, Data Mining, Machine Learning and Algorithms.