

Big data Analytics: Boon to OSS/BSS

Prof. Esmita Gupta

*Department of Information Technology
Birla College of Arts, Science and Commerce*

Abstract-Virtualization and Big data shows an essential paradigm shift in industrial technology enabling companies to gain a step into new business opportunities. In today's world there are enormous diversity of data sources containing datasets different by the structure, format, origin (forecasts, estimations, measurements etc.), access protocol and actuality. All these data should be accessed according to its nature and semantic significance. Hence data-intensive scientific discovery is one of the most important issues. There are a lot of data sources which can be considered as a subject to scientific discovery: data produced by simulation, obtained from various sensors, collected during observation, crawling etc. Nowadays such data forms large volume which can be mined to obtain information and knowledge.

• **Management of Big Data is important because in today's world almost all the organizations have a large amount of data which needs to be managed and controlled so that the data of these does not crash. Thus smooth and proper functioning of Big Data requires the management of Big Data.**

This research will help to virtualizes the data so that, a service provider has the flexibility to scale them up and down on the basis of what each application requires, in the current phase of its cycle. This can be done irrespective of the hardware boundaries of heterogeneous data applications. The research work also has considered the Big Data analytics as a solution which can be used in combination with evolutionary computing to extend data-intensive scientific discovery for basic classification and analysis of possible system evaluation scenarios.

Keywords-Big data, big data analytics, Business Intelligence, Sentiment Analysis, virtualization

I. INTRODUCTION

Big data is mostly generated from social media websites, sensors, devices, video/audio, networks, log files and web, and much of it is generated in real time and on a very large scale. Big data analytics is a method of examining huge amount of different data with different data types and constraints, or big data, in an effort to discover unseen patterns, anonymous correlations and other valuable information / knowledge.

All the data is a vital asset for enterprises and is touted to become even more important in the times to come. Organizations require to draw insights for business support, but the capabilities of existing systems need to be augmented to accommodate the changing nature of data. Owing to globalization, digitization, and a rapidly changing

economic climate, its volume and complexity are growing at astounding rate. This research work will help the organizations to effectively and efficiently manage data and transform it into valuable information in a timely manner with the concept of virtualization, – to meet the evolving business requirements with greater agility, and at affordable costs.

II. ADVANTAGES OF BIG DATA ANALYTICS

Big data analysis allows market analysts, researchers and business users to develop profound understandings from the available data, resulting in numerous business benefits / gains. Business users are able to make a precise analysis of the data and the key early indicators from this analysis can mean fortunes for the business. Some of the exemplary use cases are as follows:

• Whenever users browse travel portals, shopping sites, search flights, hotels or add a particular item into their cart, then Advertisement Targeting organizations can scrutinize this extensive variety of data and activities and can provide better recommendations to the user regarding offers, discounts and deals based on the user browsing history and product history.

In the telecommunications space, if customers are moving from one service provider to another service provider, then by analyzing huge call data records of the various issues faced by the customers can be uncovered. Issues could be as comprehensive as a significant upsurge in the call drops or some network congestion problems. Based on analyzing these issues, it can be identified if a telecom company needs to place a new tower in a particular urban area or if they need to revive the marketing strategy for a particular region as a new player has come up there. That way customer churn can be proactively minimized.

III. PROPOSED SYSTEM

Big data analysis involves making "sense" out of large volumes of varied data that in its raw form lacks a data model

to define what each element means in the context of the others. In order to explain this concept the system has considered a use case of social media websites and the concept of sentiment analysis.

The types of analytics that may embody by distinguishing selected keywords in social media posts given by the users, watching for posts that sit down with specific product or company brands, following client sentiment supported positive or negative references to an organization, and detecting problems that could pose a threat to a company's reputation and revenues.

The analysis is processed with following steps using parallel HDFS and Map Reduce functions.

- First, social networking Big Data is gathered from some SNS services.
- Second, the necessary data is extracted from gathered data.
- Third, the extracted data is processed to load into the HDFS.
- Fourth, the processed data is loaded into the parallel HDFS.
- Fifth, trend analysis is processed via the MapReduce functions using dictionaries for trend analysis.

There are two proposed dictionaries used in the MapReduce functions. They are positive word dictionary, a negative word dictionary. These dictionaries are used to find out trend about product. In our project, we mainly focused on the speed of performing analysis which can be achieved by using the various modules which are as follows:

- Determine Reviews :

As our project is Trend analysis using hadoop which shows trends about different products. To show this trend, project takes input data as user’s reviews which mentioned by the users on different social sites like amazon.com or carwale.com.

- Partitioning Reviews/Comments:

The Partitioning means separating the whole comment so to treat each word of comment separately. Partitioning is mainly done on basis of “Space’s” and put a single word of

comment into an array so matching each word with dictionary becomes easy.

- Matching Partitioned Word with Dictionary:

We are using dictionary of Positive words and Negative words to check whether each Partitioned word of comment is positive or negative. And after that, differentiate them as positive or negative. This is done for each single comment.

- Counting Positive or Negative Words :

After differentiating each word of single comment as positive or negative, calculating the final count of positive as well as negative is get performed. And then, comparing positive count with negative count to finalize whether the comment is positive, negative or neutral is decide.

- Mapping Output to Graph :

Once we get final count of positive ,negative and neutral it get map onto graph to show trends about particular product or model. So it become easy for user as well as organization to understand the trends.

IV. FLOW CHART

The proposed system accept the input from the files present in the social media websites, loads the keywords related to sentiments in HDFS and then does the separation of those keywords by removing all the non-required words and then displays the result analysis related to each product and each model that has been used.

The following flowchart display the way analysis of data was carried out by accepting the data from the various types of users on different products:

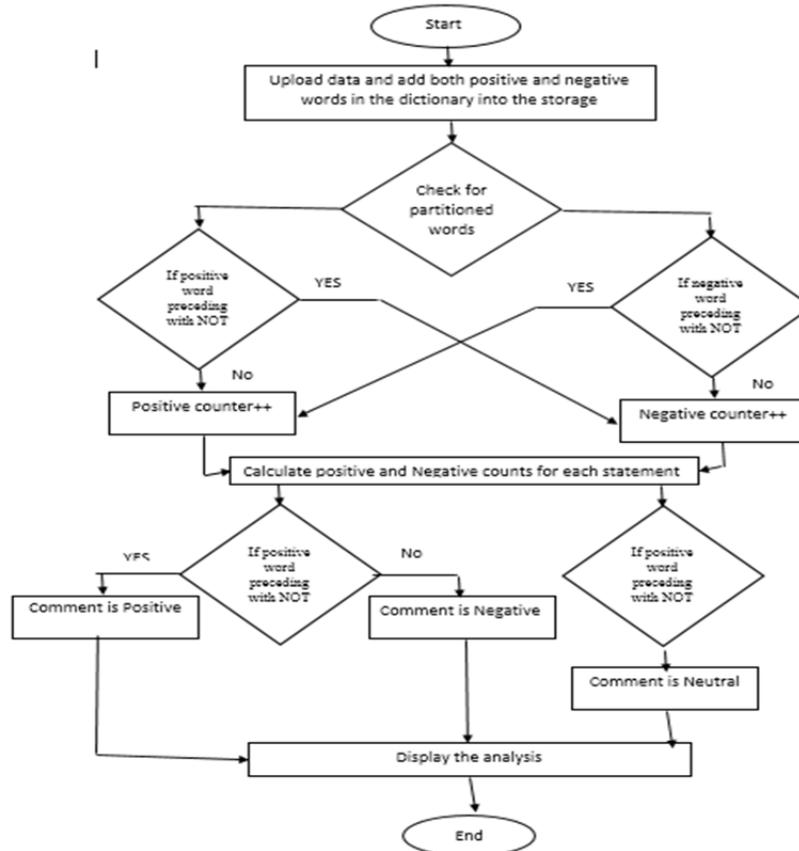


Fig 1.2 Flow chart

V. CONCLUSION

This study aims towards the implementation of a new high-level way of building scientific analytics and data-intensive discovery tasks, where all the infrastructure specifics are hidden behind the automatic performed procedures. Data Virtualization process will help to augment – not replace – the traditional data integration systems. Data virtualization will bring significant benefits to an organization, by providing the organization with the required information, the quest for agility, integration challenges with new and emerging data sources, and advancement in technologies, data virtualization will be a key component in the overall data integration collection of an enterprise.

REFERENCES

- [1] IEEE International Congress on “Big Data Analysis of Technology Trends Based on Big Data” Aviv Segev, Chihoon Jung, Sukhwan Jung
- [2] “Research on Public Opinion Based on Big Data” Songtao Shang, Minyong Shi, Wenqian Shang, Zhiguo Hong 1,2,3,4 School of Computer Science, Communication University of China Beijing, China
- [3] IEEE 17th International Conference on Computational Science and Engineering Big Data Density Analytics using Parallel Coordinate Visualization.
- [4] H. Goto, Y. Hasegawa, and M. Tanaka, “Efficient Scheduling Focusing on the Duality of MPL Representatives,” Proc. IEEE Symp. Computational Intelligence in Scheduling (SCIS 07), IEEE Press, Dec. 2007, pp. 57-64, doi:10.1109/SCIS.2007.357670.
- [5] A Survey-based Research Perspective: Fulfilling the Promise of Data Lake, Ventana Research.
- [6] Data Virtualization in Business Intelligence and Analytics, Nov 7, 2016 / by Julian ereth / in cutting-edge analytics.
- [7] Van der Lans, Rick F. (2012): Data Virtualization for Business Intelligence Systems.
- [8] White, C. (2005): Data Integration: Using ETL, EAI, and EII, Tools to Create an Integrated Enterprise. Business Intelligence Journal, 10(1).
- [9] J. H. Saltzer and M. D. Schroeder. The protection of information in computer systems. In Proc. IEEE, 63(9):1278–1308.
- [10] Hey, T., Tansley, S., & Tolle, K. (2009). The fourth paradigm: data-intensive scientific discovery.