Value Proposition and ETL in Big Data Environment

Asmita Sharma

Drashya Shah

Falguni Mangwani

Computer (COMP)

Miss Rachana Mudholkar, Computer (COMP)

Dr. D.Y Patil Institute of Engineering, Management & Research, Akurdi.

Savitribai Phule Pune University India.

Abstract: Value proposition of the system is done to specify the need of the project. Once the project is approved, spark programs, also called spark jobs, are written to extract data from the databases and stored on Hadoop clusters. The data is then filtered according to the business needs using spark jobs and then this filtered data is placed on the file system again using spark in a format that can be used by SparkML. Spark speeds up the operation by parallelizing the operations. The data stored on Hadoop distributed file system is then used as a test data to be used by SparkML to train. Once the training is done, the machine learning algorithm is then used on production data to forecast the quantity of item required at a particular store at a given date.

Index Terms -ELK stack, ETL, HDFS

I. INTRODUCTION

For any e-commerce company which runs to serve people, having out-of-stock condition is both loss of profits and bad name for the company. To not be able to provide the customer with the items they need, causes decline in sales and makes the customer hesitant to come back. Any e-commerce giant, cannot afford to lose customers and the potential sales this way, so this project was needed to solve this issue. A new method which would help the customers as well as the company. The best way to do it is to predict what quantity of what item is needed at what store. Existing system predicts this number by manually looking at the data and uses a formula to obtain the number. This process is slow and does not work at high accuracy. To solve this, this project was started. This project aims at speeding up the existing system by storing all the data on distributed system that can be used with machine learning algorithm to predict the item quantity based on past sales.This project is made to handle large volumes of data stored on system databases. The extraction process is in line with the volume of data. The process is run in parallel so as to perform at a high speed. The proposed system stores the data on clustered storage so that the storage space can be increased if needed.

II. LITERATURE SURVEY

1. Paper Name: The challenges of Extract, Transform and Loading (ETL) system implementation for near real-time environment
Author: Adilah Sabtu, Nurulhuda Firdaus, Mohd Azmi, Nilam Nur Amir Sjarif
Paper Explanation:

Organization with considerable investment into data warehousing, the influx of various data types and forms requires certain ways of prepping data and staging platform that support fast, efficient and volatile data to reach its targeted audiences or users of different business needs. Extract, Transform and Load (ETL) system proved to be a choice standard for managing and sustaining the movement and transactional process of the valued big data assets. However, traditional ETL system can no longer accommodate and effectively handle streaming or near real-time data and stimulating environment which demands high availability, low latency and horizontal scalability features for functionality. This paper identifies the challenges of implementing ETL system for streaming or near real-time data which needs to evolve and streamline itself with the different requirements. Current efforts and solution approaches to address the challenges are presented. The classification of ETL system challenges are prepared based on near real-time environment features and ETL stages to encourage different perspectives for future research.

2.Paper Name: A big data perspective of current ETL technique
Author: K. V. Phanikanth, Sithu D. Sudarshan
Paper Explanation:

Dynamic data stream processing using real time ETL techniques is currently a high concern as the amount of data generated is increasing day by day with the emergence of Internet of Things, Big Data and Cloud. Data streams are characterized by huge volume that can arrive with a high velocity and in different formats from multiple sources. Therefore, real time ETL techniques should be capable of processing the data to extract value out of it by addressing the issues related to these characteristics that are associated with data streams. In this work, we asses and analyze the capability of existing ETL techniques to handle dynamic data streams and we present whether the existing techniques are relevant in the present situation.

3.Paper Name: Value proposition discovery in big data enabled business model innovation
Author: De-ning Teng, Peng-yu Lu
Paper Explanation:

Business innovation is unavoidable in the intensely competitive marketplace. Under resources constrained condition, the option of innovation perspectives is particularly significant. Business model is the foundation of business goal and business innovation which directly facilitates the innovation of products and business processes. Meanwhile, value proposition is the most significant element in business model innovation. Under such background, an innovative approach to achieve reasonable value proposition from knowledge discovery of enterprise information system is proposed in this paper. The paper also defines major business data sources for information system and presents appropriate big data processing approaches of knowledge discovery. To effectively reveal business value of knowledge discovery, popular data interpretation techniques are presented. Furthermore, in order to facilitate firm leader to prompt value proposition on a better degree, the paper proposes a value proposition generation model based on customers, competitors and profit which called CCP to associate business value discovered from information system with value proposition creation. The pattern of value proposition generation this paper proposed will strengthen competitiveness, and provide a better approach to achieve business goal of enterprises.

4.Paper Name: Forecasting consumer behavior with innovative value proposition for organizations using big data analytics
Author: Ankur Balar, Nikita Malviya, Swadesh Pradesh, Ajinkya Gangurde
Paper Explanation:

The term `Big Data' is used to represent collection of such a huge amount of data that it becomes impossible to manage and process data using conventional database management tools. Big Data is defined by

three important parameters `Volume' - Size of Data, `Velocity' - Speed of increase of data and `Variety' - Type of Data. Big data analytics is the process of analyzing this ever growing Big Data. The goal of every organization is to maximize its value for its stakeholders. The paper aims to demonstrate that Big data analytics can be used as a catalyst for generating and increasing value for organizations by improving various business parameters. Furthermore, by utilizing case studies the paper also aims to establish that big data analytics supports creation, enhancement and improvement of various business services to significantly improve customer experience as well as value creation for organizations.

III. EXISTING SYSTEM

The existing system manually decides the number of items to be shipped to a particular distribution center on a particular day so that they do not run out of stuck. The item and store data are collected and stored in a spreadsheet file which is analyzed by people responsible manually and they come up with a number. This process is slow and not accurate. The proposed system solves this problem by moving the entire operation on a big data environment so that all the data is available at one place and required attributes are filtered out. This filtered data is used as a training set for a machine learning algorithm which will predict the final item number for a particular distribution center. This makes the process much faster and accurate.

3.1 Disadvantages of Existing System

1. Requires Manual Workpower.
2. Chances of errors.
3. The speed is slow.

IV. PROPOSED SYSTEM

This project is made to handle large volumes of data stored on system databases. The extraction process is in line with the volume of data. The process is run in parallel so as to perform at a high speed. The proposed system stores the data on clustered storage so that the storage space can be increased if needed. The end users of the system are the higher management people who look at the value proposition and decide if the project is necessary. They also decide what data to choose and what attributes are needed. Further along the system, they decide on the biases of each item for the machine learning algorithm and then use the final output to determine the amount of item that is needed to be shipped.

4.1 Advantages of Proposed System

1. Fast data retrieval process.
2. Helps to realize ongoing trend.
3. Helps to plan business strategy.

V. SYSTEM ARCHITECTURE

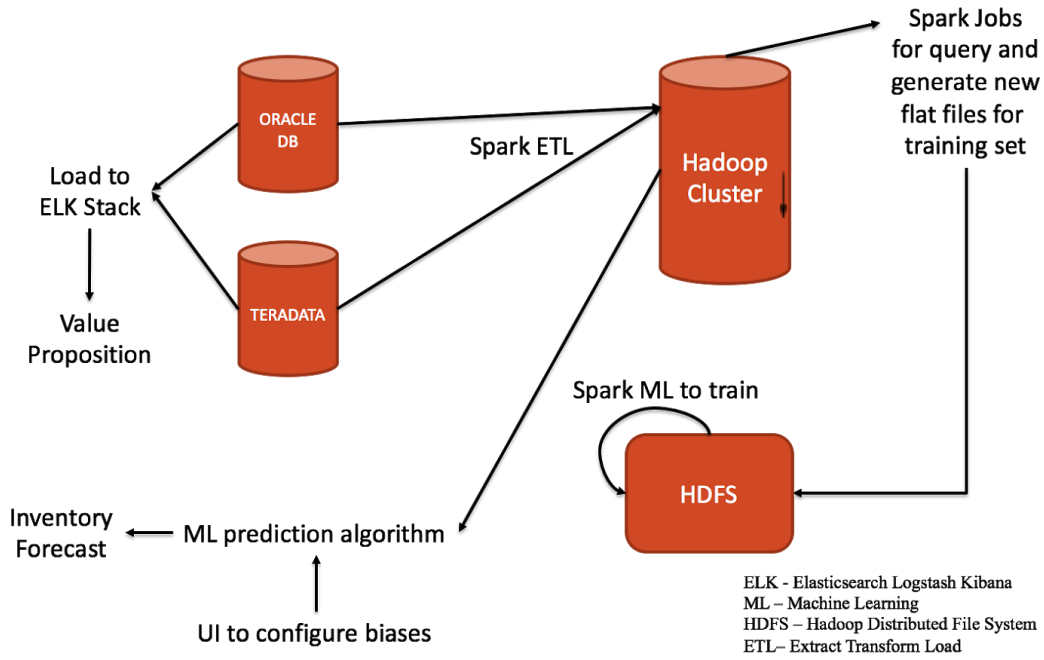

Figure 1. Proposed System Architecture

VI. CONCLUSION

Knowing when to ship a particular item to a particular DC, avoids the out-of-stock condition which would have, otherwise, caused loss of profits to the company. But the data to be worked on is huge and thus needs parallelized operations to speed up the process. The existing system manually reads the data through the files and uses some derived formulae to predict the actual quantity of items. This method is time consuming and does not offer good accuracy. The proposed method loads up all the relevant data on distributed architecture and then uses machine learning to predict the quantity of items to ship. This should speed up the process and provide better accuracy.

REFERENCES

1) M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker, I. Stoica, "Spark: Cluster computing with working sets", 2nd USENIX Workshop on Hot Topics in Cloud Computing 2017.

2) Reynold S. Xin, Josh Rosen, Matei Zaharia, Michael J. Franklin, Scott Shenker, Ion Stoica, "Shark: SQL and rich analytics at scale" in Proceedings of the 2017 ACM SIGMOD International Conference on Management of Data (SIGMOD '17), New York, NY, USA:ACM, pp. 13-24, 2017.

3) Haoyuan Li, Ali Ghodsi, Matei Zaharia, Scott Shenker, Ion Stoica, "Tachyon: Reliable Memory Speed Storage for Cluster Computing Frameworks" in Proceedings of the ACM Symposium on Cloud Computing (SOCC '16), New York, NY, USA:ACM, pp. 15, 2016.

4) Viktor Leis, Andrey Gubichev, Atanas Mirchev, Peter Boncz, Alfons Kemper, Thomas Neumapp, "How good are query optimizers really?", Proc. VLDB Endow, vol. 9, no. 3, pp. 204-215, November 2016.

5) M. Bala, O. Boussaid, Z. Alimazighi, "Big-etl:extracting-transforming-loading approach for big data", Int'l Conf. Par. and Dist. Proc. Tech. and Appl., vol. 8, no. 4, pp. 50-69, Oct. 2016.

6) K. Shvachko, H. Kuang, S. Radia, R. Chansler, "The hadoop distributed file system", 2017 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST), pp. 1-10, May 2017.

7) Adilah Sabtu, Nurulhuda Firdaus Mohd Azmi, Nilam Nur Amir Sjarif, Saiful Adli Ismail, "The Challenges of Extract, Transform and Loading (ETL) System Implementation For Near Real-Time Environment", IEEE Computer Society, 2017.