International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 04 | Apr 2019 www.irjet.net p-ISSN: 2395-0072

Business Intelligence using Hadoop

Shalaka Wikhe[1], Rasika Yelpale[2], Apurva Jagtap[3], Divyani Sirsat[4], Prof. Nutan Deshmukh[5]

[1,2,3,4,5]*Dept. of Computer Engineering, M.K.S.S.S Cummins College of Engineering for Women, Pune, India*

Abstract - *In the big data era, the Information Technology industry is continuously coming up with new models and distributed architecture to handle the exponentially increasing amount of data. Size of data sets being collected and analyzed in industry for Business Intelligence is growing rapidly, making traditional warehouse solutions expensive. Volume and complexity of data collected in data warehouse systems is growing rapidly. These pose as challenges to traditional data warehouse platforms. To achieve Business intelligence it requires proper tools to be selected. The most commonly used Business Intelligence technologies are OLAP and reporting tools for analyzing the data and to make tactical decisions for the better performance of organizations. Hadoop gives new opportunities for implementing data warehouse platforms overcoming these challenges. Hadoop stands out as a well-known open-source framework for big-data analytics. It is designed to work seamlessly with a stack of open-source tools to enable the storage and processing of a significant amount of data using clusters of commodity hardware. In this paper, the proposed system makes use of Hadoop to prevent problems faced by Data warehouse platforms and OLAP. It deals with OLAP on Hadoop using Apache Kylin which creates OLAP cubes using data present on hive. Analysis of cubes are done by firing SQL like queries. Reports and dashboards are further generated using Business Intelligence tool, Qlikview providing insights of the organizational data. It is extremely beneficial to business users in order to take accurate and precise decisions.*

Key Words: Business Intelligence, Data Warehouse, Big data, Hadoop, Sqoop, Kylin, Qlikview, OLAP, Dashboard.

1. INTRODUCTION

With the advent of internet and various technologies, the amount of data being produced is increasing tremendously day by day in all fields. Managing and analysing this data is becoming more difficult and all organizations ranging from startups to big established companies have realized that they want to take full advantage of this data to manage their businesses efficiently. This big data is identified by its characteristics which are – Volume, Velocity, Variety, Veracity and Value. To handle this huge and complex data we need a platform that can increase the efficiency of dealing with this data. Hadoop platform consists of various components like HDFS (Hadoop Distributed File System), Hive, Map Reduce which contribute in handling big data.

1.1 LITERATURE SURVEY

This literature survey is an analysis of Business Intelligence and Retail, Data warehouse, Hadoop framework, Hive and Data Visualization. Business Intelligence is a set of methods and processes used to gather and transform the raw data into useful information to get insight of businesses. The businesses across the globe have understood the importance of various kinds of data, their unseen relationships and the data associated between products and consumers. The information systems used earlier in companies were not efficient. Analysis of data from various facet like finance, sales, marketing was missing. This resulted in loss of productivity of the businesses. This led to need of proper and systematic collection and visualization of data that would help the employees of a company to take better and faster decisions on time. Data warehouse is a large repository of historical data which is collected from different sources. Hadoop is a framework that is used to process and analyze big data. It is designed to run on commodity hardware. It has several components –1. Hadoop Distributed File System for storing huge amount of data and 2.MapReduce used for processing this huge amount of data. Hive is an application that is developed for data warehouse that provides SQL interface as well as it acts as a relational model. In the data presentation phase of BI, query and analysis results are provided to the end users in a human understandable way such as tables and charts which supports decision making and strategic thinking.

1. 2 SYSTEM IMPLEMENTATION

This paper presents a three tier architecture design. It is divided into loading data from the database, creating cubes out of it and displaying it on the dashboard.

1. Loading data from the database

The first task is to load the AdventureWorks Database from SQL Server to the Hadoop platform. This database contains data of finance, sales, products and customers of the bicycle store. This shifting of data is done by using Apache Sqoop. It is an open source tool used to transfer bulk of data from structured data stores like relational databases such as Teradata, Netezza, Oracle, MySQL, Postgres, and HSQLDB. After transferring the data, queries are fired for insertion of tables in Hive. Hive is data warehouse which is a component of Hadoop that processes the data. It is used for summarizing the big data making it easy for querying and analyzing. It is provides a

language similar to SQL called HiveQL or HQL for processing the data on it.

2. Creation of cubes

OLAP cubes are created from the Data Warehouse which is available in Hive environment. OLAP cubes is a process of storing the data in a multi-dimensional form that is classified into dimensions. It is used to gain insight of the data and analyze it. Each cell of this cube contains some number or data that represents some measure and this data is stored in schema like star or snowflake. Apache Kylin, is used to create these OLAP cubes. It is an online analytical processing engine used to handle petabytes of data. Tables are loaded into Apache Kylin from HIVE. Models are then defined to according to specific requirement by selecting the fact table and adding the lookup tables as well as selecting the dimensions specific to the corresponding fact table. A fact table is a primary table in the model and it contains measurements/facts and a foreign key to dimension table. A dimension table contains dimensions of a fact and they are joined to fact table via a foreign key. Cubes are then defined based on the selected Model.

3. Creating Dashboard

Dashboard is created in Qlikview - a powerful Data analysis tool which allows visual analyzing of data relationships.

They are created by selecting the tables which we intend to analyze. Required data is then loaded and reports in the form of pie charts, graphs, bar charts etc. are generated as per requirements. Adhoc reports are generated based on the business requirements. Reports can be consolidated to generate a Dashboard view. This helps the business analysts and users to take appropriate business decisions and increase the productivity of the business.

Data Used:

AdventureWorksDW is used which contains sales data of a fictitious company Adventure Works Cycles, which is a large multinational manufacturing company. The company manufactures and sells metal and composite bicycles to North American, European and Asian commercial markets.

Customer and sales-related information is a significant part of the AdventureWorks database.

2. SYSTEM FEATURES

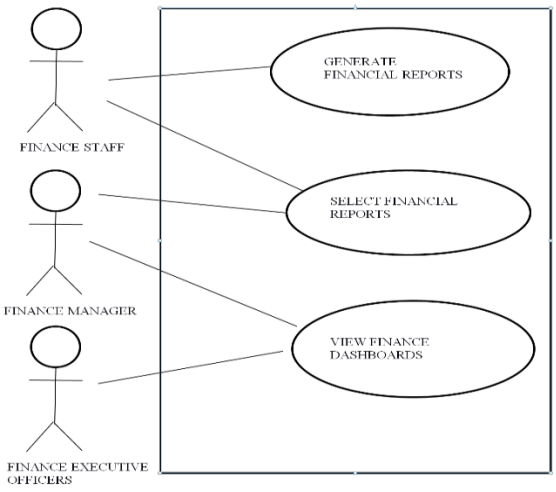
1. Financial Reporting: Supports the scenario of reporting income statements and balance sheets that include all subsidiaries. Also supports the ability to report the financial data in a specified local currency.
2. Actual versus Budget: Supports the scenario of analyzing actual expenses against budgeted expenses.
3. Product Profitability Analysis: Supports the scenario of analyzing the product sales margin by tracking costs, discounts, and selling prices.
4. Sales Force Performance: Supports the scenario of tracking the variance between sales quotas and actual sales.
5. Trend and Growth Analysis: Supports the scenario of analyzing how the current period compares to prior periods in terms of sales.
6. Promotion Effectiveness: Supports the scenario of analyzing how promotions affect current sales performance

3. ADVANTAGES

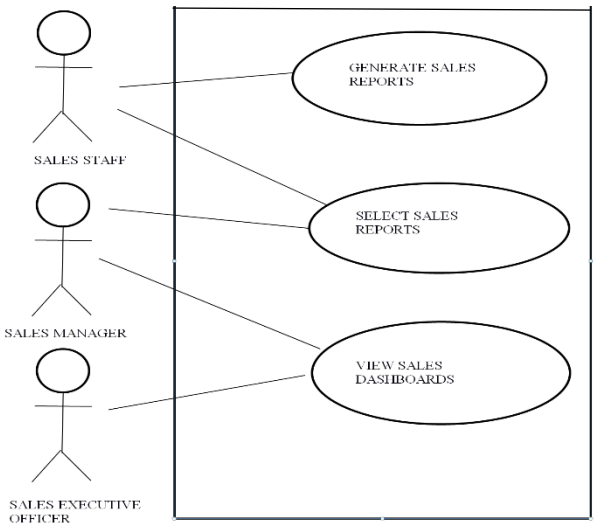
1. Data can be stored in a hierarchical way via cubes
2. Many SQL functions are supported by the OLAP tool
3. The latency time on Hadoop is very less.
4. User can reach summarized data quickly.
5. Provides simplicity of data in measures and dimensions.
6. Ability to analyze large amount of data directly in Hadoop cluster.

4. USE CASE DIAGRAMS

1. Use case diagram for Finance Department



2. Use case diagram for Sales Department



5. FUTURE WORK

1. Various OLAP frameworks as alternative to Kylin can be tried out.
2. Unstructured data can be ingested in the Hadoop environment.
3. The project can be deployed on multimode cluster instead of a single node VM.

6. CONCLUSIONS

Organizations that have developed BI successfully are thriving the downturn. It has helped them to manage inventories, cut costs, better target promotions, increases equipment utilizations and identify their most loyal customers and their preferences. BI eliminates unknowns and is playing a pivotal role in restoring confidence by throwing light upon the broader economic landscape. Traditionally BI was being used by large corporations, but now has been democratized to make it suitable for mid-sized organizations.

In a position of continuously increasing humongous data, this system focuses on helping organizations gather, store and access data with ease. The aim to use the Hadoop platform for insightful analytical thinking and visualization for the business users. The purpose of this project is to help the business analysts in taking appropriate decisions for their organization.

REFERENCES

[1] Chaudhary1, S. , Murala, D.P. , Srivastav, V.K. , "A critical review of data warehouse", Global Journal of Business Management and Information Technology , Volume 1, Number 2 (2011), pp. 95-103

[2] A. Antony Prakash, Dr. A. Aloysius, "Architecture Design for Hadoop No-SQL and Hive", International Journal of Scientific Research in Computer Science, Engineering and Information Technology , Volume 3 , Issue 1 , ISSN : 2456-3307 , 2018]

[3] Justin Hermann and Chris Manobianco, "Business Intelligence in Retail Industry", Academia.edu, 06 Nov. 2014.

[4] Jack G Zheng, Data Visualization for Business Intelligence, Book: Global Business Intelligence, Chapter 6, December 2017, DOI: 10.4324/9781315471136-6.

[5] https://www.cloudera.com/products/open-source/apache-hadoop.html

[6] https://hive.apache.org/

[7] https://intellipaat.com/tutorial/data-warehouse-tutorial/what-is-olap-and multidimensional-model/

[8] http://sqoop.apache.org/