Volume 4, Issue 8, August 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Performance Optimization in Big Data Predictive Analytics

Aditi Jain
Research Scholar, JECRC University
India

Manju Kaushik
Associate Professor, CSE, JECRC University
India

Abstract—Big Data moves around 5 Vs- volume, velocity, variety, value and veracity. Storing huge volume of data available in various formats which is increasing with high velocity to gain values out it is itself a big deal. Large business organizations in various domains are looking forward to get maximum out of this big data solutions to compete in business world. Making right decisions on right time is the logic of business. High speed Query execution from large datasets is based on the storage structure. The approach to solve the problem is to monitor the query execution speed for Predictive analytics on Big Datasets and providing solutions to speed up the query execution using various predictive models and data mining techniques which results in enhancing the predictive scores and business values.

This will results in high and more précised predictive scores on time which helps in maximizing the productivity of people, processes and assets of an organization. It can be helpful in detecting and preventing threats and frauds before they affect the organization.

Keywords—Big Data, Predictive Analytics, Datasets, Query Execution, Data Mining

I. INTRODUCTION

It is really noteworthy that "Big Data" comes in picture with astounding changes in business, government and other economy aspects. Large- Scale governmental datasets and private sector data can greatly improve the way we measure, track and describe economic activity. But it's not just companies and organizations that stand to gain from the value of Big Data, Consumers can also enjoy significant benefits. Business organizations are fast moving towards predictive analytics to predict future events and behaviors and achieve success and compete among the world's market. To have deep insight in organizational data and then to find patterns among these datasets for better decision making with high execution speed and accurate result so to predict future happenings is the only motivation behind this research work.

Big Data techniques and tools are already familiar to those parts of the business that routinely deal with sub-transactional data as part of their work, such as sensor data coming from manufacturing or process industries and especially those involved in monitoring and understanding web traffic, but these groups are typically not a part of enterprise IT groups. While enterprise IT uses primarily relational databases and SQL, the other groups using MapReduce and Hadoop are more comfortable with programming languages like Java, Python or Perl. There is already quite a bit of culture clash going on between these fractions.

Big data analytics can be defined as the combination of traditional analytics and data mining techniques along with large volumes of data to create a fundamental platform to analyze, model and predict the behavior of customers, markets, products, services and the competition, thereby enabling an outcome-based strategy precisely tailored to meet the needs of the enterprise for the market and customer segment. Big data analytics provides a real opportunity for enterprises to transform themselves into an innovative organization that can plan predict and grow markets and services driving towards higher revenue.

II. BIG DATA and PREDICTIVE ANALYTICS

Large business organizations are now using Big Data to extract values and handling business processes and for developing predictive models also. Business intelligence and analytics helps in improving the efficiency of companies. Predictive modeling is the reason behind the drastic change in products and services in recent years.

Google search engine uses predictive models and algorithms in displaying search results and news feeds and even predicts the rest of one's text. Amazon also relies on predictive models of what kind of book user might purchase. Online advertisements that displays on users screen showing various offer in which user might have interest is totally based on predictive models.

The applications of predictive algorithms are not limited to online world. Health care industries are also making use of it in providing quality services to humanity. It is now common for insures to adjust costs and quality measures based on 'risk scores', which are derived from predictive models of individual health costs and outcomes. Predictive models also have their use in credit card companies to maintain their underwritings, pricing and promotion actions.

Predictive analytics uses algorithms to find out patterns in data that might predict similar behavior in future. To explain predictive analytics, consider a very common example to find out a predictive model that predicts which customers likes to switch mobile network. Telecom industries can use the customer data which includes the details of call made by him,

text messages, internet usages, bill amount, and many other variables to develop a model that will predict which customers likes to switch mobile network. If the model results successful to predict it, then the company will try to find out solution to make customer to stop to churn from their network.

Predictive analytics is a continuous process. To maximize the success with predictive analytics, following steps must be followed by an organization:

Identify business goals: First step is to clearly identify business goals. Clearly defined business goal can only leads to a successful predictive model. For example- business goal might be to suggestions of items when a customer is adding products to his shopping cart. It will helps in increasing sells of the store and customer has to apply less efforts in finding similar items or what he wants to purchase more. This will gain customer satisfaction and increase market value.

Data understanding from various sources: After deciding business goal, next step is to collect data from variety of sources available. Large organizations store their valuable data in multiple silos. Data from external sources might be collected for analysis purpose. These external sources can be social media websites, government data, public sector data and many other sources and data collected from variety of sources helps in augmenting internal data. Data visualization tools can help data analysts to explore data from variety of sources to determine which data is relevant for predictive purpose.

Data preparation: Raw data can be collected from variety of sources but preparing that data for predictive analysis is the key challenge. Raw data is unsuitable for analysis. Data preprocessing must be required for run predictive algorithms on the data. Data analyst must perform preprocessing to make data suitable for input to the predictive models.

Development of predictive model: Predictive analytics modeling tools are used to run analysis algorithms for the data. Data analysts use one or more of these tools to perform analysis. Hundreds of machine learning algorithms and statistical algorithms are used by data analysts to find predictive models. Analysis is performed only on a subset of data which is called training data and the remaining data sets is called test data that analysts may use to test model. Training data which is input to the algorithms is only 70 % of the entire data sets and remaining 30% is used to evaluate the model [1]

Evaluation of the model: Predictive analytics is all about probability not absolutes. Before performing analysis on the test data, organizations used to set a probabilistic output that they will use to compare with the results of the predictive models. To evaluate the efficiency of a predictive model, data analysts run it against the test data sets. It the predictive output is found more effective than the random selected output, and then this model is effective predictive model. Data analysts can run various other algorithms to find most predictive model. If no results find then it is assumed that data is not suitable for prediction or not enough to perform predictive analytics

Deployment: Once an effective predictive model is identified, then it is deployed in the production application by the analysts. This deployed model consists of logic to run predictive rules, formulas, and method to get the data required by the model and finally to obtain the results**.**

Examine the effectiveness of model and result analysis: It is necessary to continuously evaluate the effectiveness of the model. It might happen that organization is performing predictive analytics with previously selected data sets and market scenario has been changed. Organizations must continue the predictive analytics process to stand in competitive market.

III. STRENGTHS and WEAKNESS of BIG DATA PREDICTIVE ANALYTICS

The strengths of Big Data Platform

- The increasing data volume has very high velocity and can only be managed through Big Data technology.
- Traditional Data base Management system RDBMS is not suitable for processing massive datasets on real time due to low bandwidth architecture. Big Data architecture overcomes this problem
- Big Data can process structured, semi-structured, unstructured and quasi structured data which RDBMS can't process.
- With big data, the value is discovered through a refining modeling process, make hypothesis, create statistical, visual, semantic models, validate and then makes a new hypothesis.
- It can easily identify which data is valuable and then transform that data and analyze it.

The Strength of Big Data Predictive analytics

- Predictive analytics will help organizations predict with confidence what will happen next so that they can make smarter decisions and improve business outcomes.
- It can be helpful in detecting and preventing threats and frauds before they affect the organization.
- Predictive modeling helps in increasing ROI of the products[2]
- Market value is increased in the field of retail markets, share markets, health care industries, and in government sectors also.
- Government has been started predictive analytics to conduct election campaigns.

Drawbacks of Big Data Technology

- Hadoop is designed for processing large jobs only. On executing small jobs, performance goes down. Solutions are also designed by implementing different processing models for jobs in map and reduce phase.
- Map reduce provides high level of abstraction but it is not suitable for data mining. It is best for data parallelism.
- Programmers need to go deep and wide to use this technology.

- Less experts are available in market for big data technology therefore requires training programs
- There is a need for a scalable distributed computing framework that provides both abstraction and performance.

IV. PERFORMANCE OPTIMIZATION in BIG DATA SETS

In big Data architecture, there are large numbers of software vendors in market who are trying to enhance the existing architecture whenever there is a limitation. The basic tool in each solution is Hadoop and its component as MapReduce, Pig, HBase and Hive. Hadoop is a software framework which is designed for processing large datasets. The main motive which was kept in mind of developers while designing this platform was the execution of large amount of data. Performance factor was not a major issue but now companies are looking forward for high performance output. This performance is based on the number of jobs executed in a time cycle.

Researchers across the world are finding the different ways to provide the solution to the problem of high performance. The major components where the enhancements can be made are:

- Map Reduce
- Pig
- Hive

In Map Reduce, there are three main phases: Mapper, Shuffle and Reduce. At each of these phases, nodes are assigned the task for execution. The results from Mapper are transferred to the shuffle. There is a check point mechanism introduced by some researchers to store this intermediate result so that when node or task failure occurs, it can be recovered from this point.

The push down mechanism [3] used in pig programs are replaced by pull down mechanism so that on executing small jobs, performance remains the same, it doesn't goes down.

Researchers designed performance architecture for Pig and Hive programs to monitor the time taken by the resources to complete the execution of the job and also the number of resources assigned to each node for job execution. This helps in improving the performance by efficient resource allocation.[4]

From various studies and research works, we consider that for performance optimization, enhancement in the design architecture of MapReduce, Pig and Hive components can be the solution.

When we consider big data predictive analytics, then the performance factor is also affected by predictive models. Algorithms used in predictive models should be according to the type of results we want and processing models also. Algorithms as clustering, classification, collaboration are used in creating predictive models. These algorithms are implemented in mahout which is also a component a Hadoop.

Thus to make the predictive models effective and highly optimized, pig programs and Hive components must be efficiently designed.

V. QUERY PERFORMANCE OPTIMIZATION TOOLS: HIVE AND IMPALA

Hive is an open source data warehousing solution which is on the top of Hadoop. It supports queries structured in SQL like language. This language is named as HiveQL. These are compiled into map reduce jobs and then executed on Hadoop. It also supports custom MapReduce scripts. These scripts are plugged in into queries.

Data in Hive is organized into 3 ways [5]

i. Tables
ii. Partitions
iii. Buckets

Tables: Tables in Hive are much similar to tables in relational database. Each table has its own HDFS directory. The tables are stored in files within the defined directory in serialized way. Users can associate tables in serialized way of the data. The format of serialization of each table is stored in system catalog and hive automatically use this format while compilation and execution. Tables from HDFS, NFS or local directories are also supported in Hive.

Partitions: Each table in Hive can have multiple partitions which show the data distribution within directories or sub directories.

Buckets: Data of each table in each partition is then further divided into buckets. This division is performed on the basis of hash of a column in the table. After division, each bucket is stored as a file in the partition directory.

HiveQL is a SQL like programming language that supports select, project, join, aggregate, union all and sub-queries in from clause. HiveQL also supports DML and DDL statements. DDL statements are used to create tables with specific serialization format and partition and bucket column. HiveQL doesn't support updating and deletion of rows in existing tables. It allows multi-table insert, where multiple queries on the same input data can be executed using a single HiveQL statement. Hive allows User Defined Column Transformation and Aggregation functions.

Impala and Hive both are the open source projects that are competing for interactive SQL in Big Data deployment. It has been analyzed by Cloudera that impala runs 6 to 69 times faster than Hive 0.12. [6]The drawbacks of impala over hive are that Impala's sub-query support, aggregation and window function are not better than Hive's support. Another drawback is on installation of both the products. Installing Impala in an existing cluster means to run the whole processes again in the running cluster while hive can easily get installed in the same running environment.

Impala supports only some of the familiar file formats used in Hadoop. It loads and query data files generated by various Hadoop components as Pig, MapReduce and vice versa also. The file format supported by impala has a significant impact on performance. Some of the file formats support compression that helps in reducing the size of the data on the disk and the amount of I/O resources also. With the help of data compression, smaller number of bytes is transferred to the memory and thus reduces the time taken to transfer the data.

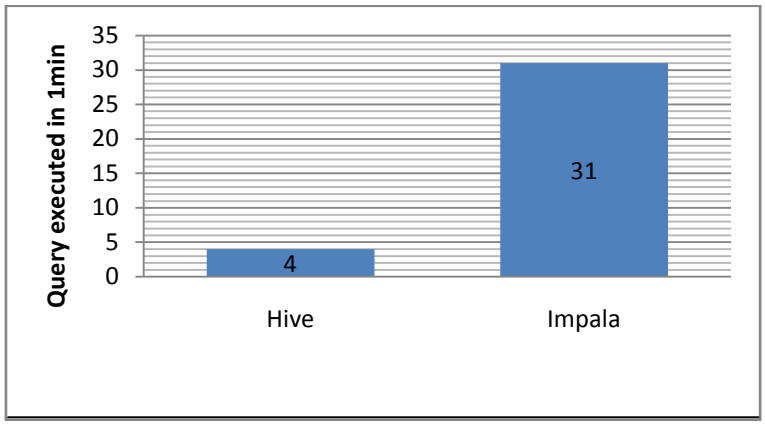

Fig1: Query Execution Graph of Hive and Impala

VI. CONCLUSION

In performing analysis using Hadoop and its various components, we have tried to find out the limitations of PIG, MapReduce, Hive and Pig and have planned future work according to this only. Pig programs take more time in compilation as these programs are first converted into Directed Acyclic graphs and then fetch the data from HDFS. It consumes a lot of time because the data in stored only on a single name node server. Similarly in Hive, file formats decreases the performance of the query execution which can also be optimized using a mechanism to store the file only in those formats which takes less time in data retrieval and takes less storage space with compression also.

REFERENCES

[1] G. Mike, "The forrester Wave ™: Big Data Predictive Analytics Solution, Q1, 2013", *Forrester Research Inc*., Cambridge, USA, 2013

[2] V. Dan, M. Henry D., "The Business Value of Predictive Analytics", IDC *analyze the future*, 2011

[3] Yan J., Yang X., "Performance optimization for short MapReduce job execution in Hadoop", *IEEE, 2nd International Conference on Cloud and Green Computing (CGC),* Xiangtan, 688-694, 2012

[4] Zhuoyao Z, Cherkosova L, "Optimizing Completion Time and Resource Provisioning of Pig Programs, *IEEE 12th International Symposium on Cluster, Cloud and Grid Computing(CCGrid)*, Ottawa, ON, 811-816, 2012

[5] T. Ashish, et.al, "Hive: A warehousing solution over a Map-Reduce Framework", Facebook Data Infrastructure Team, Brown University, 2009

[6] G. Alex, "Interactive SQL in Apache Hadoop with Impala and Hive", available at http://www.infoq.com/news/2014/02/SQL-Apache-Hadoop-Impala-Hive, 2014

