ANALYSIS OF RETAIL DATA AND DISTRIBUTED FRAMEWORKS USING HADOOP TECHNOLOGY

[1]Smita Tiwari, [2]Veena Mishra

Jawaharlal Nehru College of Technology (RGPV) University
Madhya Pradesh (Bhopal)

***Abstract* - This paper is all about retailer we all knows there is many of websites (amazon, flipkart) etc. Where we are using Hadoop technology to perform creation of data, fetching and storing (pig and hive). Here, retailer is having access to an object of production. In Bigdata, this article we attempt to focus on the value created for retail industry. We focused on the value that is to be create, stored and resolve by big data for retailers easiness. While major of paper are published of big data and their analysis around their technical execution there is few of paper that work on retails. So, that is not exclusive we all live in this era where we won't do any paper work. In retail data analysis we work on Hadoop to listing out the customers detail regarding their objects and their module that are used. And also there is little bit study of Hadoop distributed file system is to store bulk of data reliable, clustering and streaming. An also studied of various components as spark and flume. In distributed files there is thousands of users access an application like as facebook and twitter so all about the management of the unstructured data.**

***Keywords*- Retail data analysis, hive, Distributed frameworks, Big data, Hadoop study, Big data components.**

1. INTRODUCTION

Through this era information is kept rising and that heavily growth generated their needs to change the information is managed and executes. Apache Hadoop is an distributed Infrastructure. It is used a singleton machine and in advantages get full potential. There is thousands of computers with multiples of processors. its designed efficiently and distributing their jobs. Same thing happen in retailers side. Big data refers to set of data that is difficult to store, manage and analyse using software and databases. Big data we can say a terminology that deals with all these problem and provide a solution. Now a days thousands of company working and creating a huge amount of data in daily basis many of ecommerce site too. Walmart collecting more than 2.5 pb data of customer transactions alone in regular basis. All of data that being produced we analysed to obtain insights. Retailers can use big data analytics to improve their marketing. So through big data analytics we get more about how customers behave its potential gains or pricing. Hadoop has many of techniques using this we fetch the no. of records using an object. Basically here perform extraction transformation and loading (ETL), in this tool that allows to extract raw data from variety of sources. It transforms according to structure and loads it into a data warehouses. By transforming data into a structure format, which can help to obtain valuable information? Data warehouse is a data library used to storing data which are structure for analysis. Distributed system it is a set of multiple systems connected together is used to solve a computational problem. At last Hadoop is introduced is an apache software that is open software framework and it perform loading, storing and querying. Here we only were targeting on customer's individuals behaviours we analysing customers through their product recommendation. And fetch according to their brands chain wise end all of data stored in an other value. The main aim of this analysis is the fetching the no of records from retail industry.

2. DEFINTION OF BIG DATA

Big data means really a big data it is a collection of large datasets that cannot be processed using traditional computing techniques. Big data is not a merely data rather it has become a complete subject, which involves various tools, techniques and framework. So, Big data is a Buzzword and a problem which deals with 3-v's.that is generated –

1. **Volume -**

It is defined as the amount of data available to an organization or a firm, as long as it can access it, is a data volume. In simple manner we can say it is a data of particular application and firm data. Weightage of data usage.

2. **Velocity -**

Velocity is defined as no. of users. It can we say streaming and aggregation of data and the continuous flow of data or information.

3. **Variety -**

Data variety is the richness of data representation. Many of things happen it takes in the form of images uploading, post, updates, chatting(messaging)etc. variety of data means types of data it is structured, unstructured or semi-structured. There is another "v" also introduced i.e. stands for veracity it means big data request to change from one machine to another.

So the big data includes huge volume of, high velocity and extensible variety of data. The data in it will be three types.

- Structure data- Relational data.
- Semi-structured data – xml data

- Unstructured data- word ,pdf, text, media logs.

3. ANALYSIS OF BIG DATA

MapReduce-

This includes systems like massively parallel processing (MPP) database systems and map reduce that provide analytical capabilities for retrospective and complex analysis that may touch for all of the data. MapReduce provide a new method of analysing data, that is complementary to the capabilities provide by SQL and a system based on MapReduce that can be scaled up from single servers to thousands of high and low end machines. In MapReduce is mainly a data processing component of hardtop

It is a programming model for processing large number of data sets. It contain the task across the nodes. It consist of two phases-

- Map
- Reduce

Map converts a typical dataset into another set of data where individual elements are divided into intermediates pair key and value.

Reduce task that output files from a map considering as an input than integrate the data tuples into a smaller set of tuples.

Terminology-

Mapper-this application helps to maps the input key/value pairs.

Namenode-this node manages the HDFS.

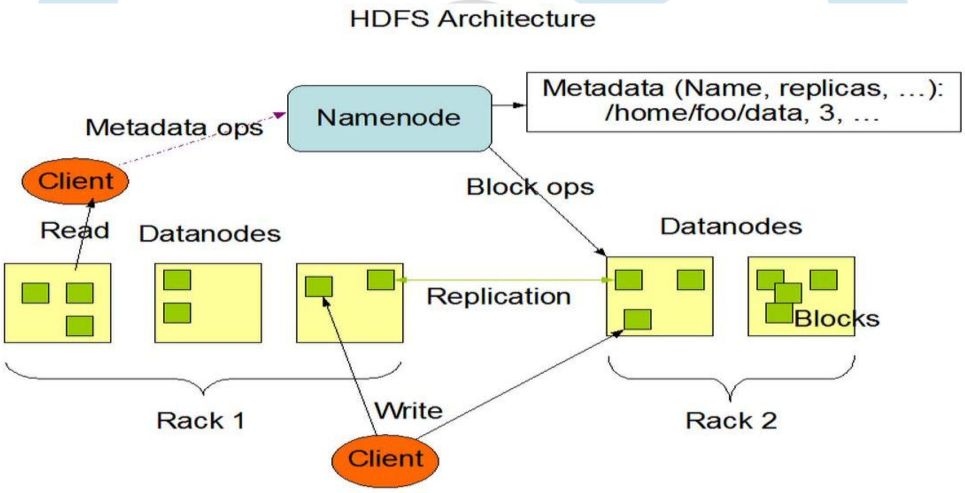
Datanode- datanode is used where data is presented before processing.

Masternode- It is used where jobtracker runs and receives job request from clients.

Slavenode- Map and reduce program run particularly in this node.

Job- It is an execution process of a mapper and reducer.

Task- Task of an execution of a mapped or a called as reducer on a slice of data.



4. FRAMEWORKS BACKGROUND AND HADOOP RELATED PROJECT

HDFS-Hadoop distributed file system is part of hardtop framework, used to store and processed datasets. It provides fault tolerant file system to run on commodity hardware. In Hadoop ecosystems there is contain different sub-projects (tools) such as Sqoop , hive, pig that are used to helphardtop modules[8].

Sqoop- It is used to export and import data to and fro between HDFS and RDBMS. The traditional application management system that is the interaction of applications with relational database using RDBMS is one of the sources that generate big data. Such big data generated by RDBMS is stored in relational databases servers in the relational database structure. SQOOP-“SQL to Hadoop and Hadoop to SQL” Sqoop is a tool designed to transfer data between Hadoop and relational databases servers.

PIG- It is procedural language platform used to develop a script for MapReduce operations. Apache pig is a high level platform for creating programmes that run on apache Hadoop. The language for this platform is called pig Latin. Pig can execute its Hadoop jobs in MapReduce , apache tez, or apache spark. Pig Latin abstracts the programming from the java MapReduce idiom into a notation which makes MapReduce programming high level, similar to that of SQL for relational database management systems. Pig Latin can be extended using user-defined functions (UDFs). Which the user can write in java, python, JavaScript’s.

Hive– It is platform used to develop SQL type scripts to do MapReduce operations. Apache hive is a data warehouse software project built on top of apache Hadoop for providing data summarization, query and analysis. Hive gives an sol-like interface to query and analysis .Hive gives an SQL-like interface to query data stored in various databases and file systems that integrate with Hadoop[1]. Traditional SQL queries must be implemented in the MapReduce java API to execute SQL applications and queries (HiveQL) into the underlying java without the need to implement queries in the low level Java API. In retail data analysis we are using these tools to manage the records and perform various of operations. Retail sales represent purchases of finished goods and services by consumers and businesses. They occur with products that have made it to the end of the supply chain. The chain starts with the goods producer or provider and the ends with retailer. The beginning of the supply chain includes commodities and other raw material. Manufacturers create the product. The middle of the supply chain is wholesales they distribute the goods and services to retailers. The retailers sell them to the consumers. So that all about my project run here we perform many of operations like as

- Creation of production table
- Exporting of data from hive to MySQL.
- Perform partitioning and bucketing-Partitioning and bucketing-
- Partitioning and bucketing in hive will let you do faster querying.
- For dynamic partitioning, load the data in to staging table which is already done.
- Now create a production table, and insert data
- Grouping by id and chain.
- Again export data from hive to pig
- Top 10 entries will fetch(brands, customers)
- Stored all result.

For example-Transaction is a production table queries to load data on pig-transactions = LOADTransactionsData/partm00000'USINGPigStorage(',')as(id:chararray,chain:chararray,dept:chararray,category:chararray,company:chararray,brand:chararray,date:chararray,productsize:float,productmeasure:chararray,purchasequantity:int, purchaseamount:float);

Retail Data Analysis using Pig and Hive –

- In retail data analysis there a transactions name cave file is created first and then it performs extraction, transformation and loading (ETL). Retail stores daily generate millions of transactions logs.
- Analyzing these logs would generate beautiful insights and improve business.
- Storing these logs on traditional databases would be costly and scalability will be a big challenge.

Volume of transactions

- Stores like Wal Mart are spread across different locations.
- Daily millions of customers visit these stores and generate billions of logs.
- These billions of logs contribute to huge volume of data.
- Velocity of transactions
- In peak hours, 1000's of transactions will happen in any given second.
- 1000's of transactions/sec across all stores contribute to high velocity.
- Variety of transactions
- The most widely known varieties of data generated by transactions:
- Jason Format
- Xml format
- CSU Format

Flume-Apache Flume is a reliable and distributed data ingestion tool that can ingest streaming and aggregate large amounts of data from different sources onto target data store there are variety of data sources generating data in the form of server logs, user-content on social media platform, user engagement on web service applications, network systems data.

Spark -Apache Spark is a general-purpose and fast in memory cluster computing platform configuring simple in Java, Scale, Python and SQL. Spark back-up its intensive computations by extending the Map Reduce programming model. When it comes to speed, it has the ability to run the massive stream, iterative type of computations in memory. The Spark system execution is faster than application workloads when running with Map Reduce on disk.

5. **Conclusions**

Spark is one of the newest tool in MapReduce field. Its purpose to make data analysis fast to write and run. spark always in memory querying of data in distributed machines too. So in future work I work on spark and flume for faster result and betterment in distributed file system whereas hive and pig tool reduce the no. of codes and easily to managed records. It reduce paper work we all knows its sometimes difficult to maintain. Using flume provide interface This study also analyses the uses of frameworks like Flume and Hive in order to work on the semi structured data formats. This study reproduces the data architecture by using the existing ones depending on the use case requirement on a low cost hardware. Finally, this thesis also discusses the integration of distributed framework components at a single level of design to improve execution time, storing as well as processing of data for better maintenance

REFRENCES

[1] Edward Capriolo, Dean Wampler, and Jason Rutherglen, Programming hive, O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472., October 2012.

[2] Natkins Jon cloudera BLOG, http://blog.cloudera.com/blog/2012/09/analyzing-twitter-data-with-Hadoop/.

[3] Jeffrey Dean and Sanjay Ghemawat, MapReduce: Simplified data processing on large clusters, Commun. ACM 51 (2008), no. 1, 107{113}..

[4] Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung, The google file system, Pro-ceedings of the Nineteenth ACM

Symposium on Operating Systems Principles (New York, NY, USA), SOSP '03, ACM, 2003, pp. 29{43.
[5] Ganglia github, https://github.com/ganglia/monitor-core/wiki/ganglia-quick-start.
[6] Apache Hive Developer Guide, https://cwiki.apache.org/con/uence/display/hive/ developerguide.
[7] Apache Hadoop, Apache Hadoop, 2011.
[8] Apache HDFS, https://Hadoop.apache.org/docs/r1.2.1/hdfs design.html.
[9] Apache Hive, https://hive.apache.org/.
[10] Steve Ho man, Apache ume: Distributed log collection for Hadoop, Packt Publishing Ltd, 2013.
[11] Karau Holden, Andy Konwinski, Patrick Wendell, and Zaharia Matei, Learning spark, O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.,Febru-ary 2015.
[12] Alex Holmes, Hadoop in practice, Manning Publications Co., 2012.
[13] Hortonworks, http://hortonworks.com/Hadoop/ ume/.
[14] Mohammad Islam, Angelo K. Huang, Mohamed Battisha, Michelle Chiang, Santhosh Srinivasan, Craig Peters, Andreas Neumann, and Alejandro Abdelnur, Oozie: Towards a scalable work ow management system for Hadoop, Proceedings of the 1st ACM SIGMOD Workshop on Scalable Work ow Execution Engines and Technologies (New York, NY, USA), SWEET '12, ACM, 2012, pp. 4:1{4:10.
[15] NatkinsJon,http://blog.cloudera.com/blog/2012/10/analyzing-twitter-data-with-Hadoop-part-2-gathering-data-with-ume/.Alex Holmes, Hardtop in practice, Manning Publications Co., 2012.
[16] Horton works, http://hortonworks.com/Hadoop/ ume/.
[17] Mohammad Islam, Angelo K. Huang, Mohamed Battisha, Michelle Chiang, Santhosh Srinivasan, Craig Peters, Andreas Neumann, and Alejandro Abdelnur, Oozier: Towards a scalable work own management system for hardtop, Proceedings of the 1st ACM SIGMOD Workshop on Scalable Work own Execution Engines and Technologies (New York, NY, USA), SWEET '12, ACM, 2012, pp. 4:1{4:10.
[18] Natkins Jon, http://blog.cloudera.com/blog/2012/10/analyzing-twitter-data-with-Hadoop-part-2-gathering-data-with- ume/.
[19] Mohammad Islam, Angelo K. Huang, Mohamed Battisha, Michelle Chiang, Santhosh Srinivasan, Craig Peters, Andreas Neumann, and Alejandro Abdelnur, Oozier: Towards a scalable work own management system for hardtop, Proceedings of the 1st ACM SIGMOD Workshop on Scalable Work own Execution Engines and Technologies (New York, NY, USA), SWEET '12, ACM, 2012, pp. 4:1{4:10.
[20] Natkins Jon ,http://blog.cloudera.com/blog/2012/10/analyzing-twitter-data-with-Hadoop-part-2-gathering-data-with-ume/.RohitKhare, Doug Cutting, KragenSitaker, and Adam Rifkin, Nutch: A exible and scalable open-source web search engine, Oregon State University 32 (2004).
[19] Vania Marangozova-Martin and Vania Marangozova, Introduction to distributed sys-tems.
[20] Thomas N•agele and FW Vaandrager, MapReduce framework performance comparison, (2013).
[21] Apache Oozier, http://oozie.apache.org/.
[22] Philip Russom et al., Big data analytics, TDWI Best Practices Report, Fourth Quarter (2011)
[23] Konstantin Shvachko, HairongKuang, Sanjay Radia, and Robert Chansler, The hardtop distributed le system, Proceedings of the 2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST) (Washington, DC, USA), MSST '10, IEEE Computer Society, 2010, pp. 1{10.
[24] Apache Spark, http://spark.apache.org/.
[25] Apache Spark SQL, http://spark.apache.org/docs/1.2.1/SQL-programming-guide.html.
[26] Apache Sqoop, http://sqoop.apache.org/.
[27] AshishThusoo, JoydeepSenSarma, Namit Jain, Zheng Shao, Prasad Chakka, Suresh Anthony, Hao Liu, Pete Wycko , and Raghotham Murthy, Hive: A warehousing solution over a map-reduce framework, Proc. VLDB Endow. 2 (2009), no. 2, 1626{1629.
[28] Ubuntu, http://www.ubuntu.com/.
[29] Vinod Kumar Vavilapalli, Arun C. Murthy, Chris Douglas, SharadAgarwal, Mahadev Konar, Robert Evans, Thomas Graves, Jason Lowe, Hitesh Shah, Siddharth Seth, Bikas Saha, Carlo Curino, Owen O'Malley, Sanjay Radia, Benjamin Reed, and Eric Balde-schwieler, Apache hardtop yarn: Yet another resource negotiator, Proceedings of the 4th Annual Symposium on Cloud Computing (New York, NY, USA), SOCC '13, ACM, 2013, pp. 5:1{5:16.
[30] vmware, http://www.vmware.com/.
[31] Tom White, Hardtop: The definitive guide, 3rd ed., O'Reilly Media, Inc., 1005 Graven-stein Highway North, Sebastopol, CA 95472., May 2012.
[32] Reynold S Xin, Josh Rosen, Mateo Zaharias, Michael J Franklin, Scott Shenker, and Ion Stoica, Shark: SQL and rich analytics at scale, Proceedings of the 2013 ACM SIGMOD International Conference