International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

E-Commerce Data Analysis Using Hadoop

***S.Sai Vamsi*[a], *K.Deeksha*[a], *S.Amruta Varshini* [a], *J.Srija*[a], *S.Jashwanth Kanna*[a], *P.Srihari*[b],*U.Archana*[b]**

[a]*Student,*[b]*Asst. Professor*
Department of Information Technology, GMR Institute of Technology,
Rajam – 532127, Andhra Pradesh, India.

A B S T R A C T

The idea of Big Data adapted a few years ago, but now many enterprises have concluded that the use of big data in the stream of business offers the best analysis of the brands. Considerable results are calculated through big data analysis. Hadoop is the ideal site for collecting and examining massive quantities of data after getting information from social networking sites like Instagram and Twitter, the data were organized using distributed file system as hadoop distributed file system (HDFS). As big data analytics tools become more attainable, the technology's impacts grows and expands into new sectors. This is a definite trend in the retail sector. HDFS, Map-Reduce, Hive, Pig, Spark, Oozie, and Sqoop are all suitable for extracting and analysing the required data. Some use cases were detected that would stand to gain the retailer, customers, and employees. It is essential for both large and small retailers to be able to not only react to, but also predict exactly industry trends and complexities. Data analytics is performed on the retail dataset in order to increase the organization's financial gain.

Keywords:Big data, Hadoop, HDFS, Hive,Tableau.

1. Introduction

The primary objective of large and small retail stores is not only to react to the product, but also to predict the requirements, income, and fraud analysis of the product. This analysis can be carried out by estimating the product's profit or by tracking the company's growth. The three main aspects to consider when analysing the product are the retailers,the client,the worker. It quantifies the impact of an advertising campaign in store sales.It delivers insights into a specific product range to optimise sales.At the time of shopping it check list generated which helps in personalized shopping.Recommendations from retailers help in availing offers.Getting updates on new arrivals. Rewarding scheme based on building ability of cashier.Leave approval managing during peak days.The Apache Hadoop framework, which is also free software, is used to process and store data and data sets. It offers a full source implementation of a distributed file system as well as a template for map reduction parallel processing. It is a storage device that is distributed (HDFS). Because a computer system cannot handle huge amounts of data at once, Hadoop allows for the grouping of multiple computers to evaluate large datasets. To obtain instant results, the data is processed in parallel on multiple machines. It can also be used to solve processing issues. Hadoop is also a software platform for trying to support statistics distributed applications that work with thousands of computationally distributed computers and petabytes of data. Hadoop is a distributed file system. is large enough to hold a substantial amount of data and data Increase production incrementally and avoid data loss in the event that significant parts of the storage infrastructure fail.A proposed methodology is map reduce. Map reduce is composed of two functions: map () and reduce (). The map () function is being used to map one set of key-value pairs to another set of key-value pairs. The datasets operate on a block or data size basis. Reduce () works on decreasing the amount of massive blocks of data to a mono and small block of data.Map reduce is linearly scalable, which has the benefit of allowing for easy dimensioning of data processing across multiple connecting nodes. The basic task of map reduce is to scale the application to run on 1000s, hundreds, or tens of machines in a cluster; this is referred to as configuration change. This simple scalability is the primary function of map reduce, which makes it valuable to some limited extend.The map reduce program runs in three stages they areMap stage, Shuffle stage, Reduce stage. Map stage the mapper's purpose is to prepare the input. The input data is in the form of a specified directory that is stored in a distributed file system such as Hadoop (HDFS). The input file is passed frame by frame to the mapper function, which processes the data and produces some many small pieces of data.Shuffle stage is used to demonstrate the number of words in the map function. Reduce function is used to process the data from the mapper; after processing, a fresh batch of output is stored and processed in HDFS.During a Map Reduce job, Hadoop distributes the Map and Reduce tasks to specific cluster data centres.The implementation of information all of the details of information, such as issuing activities, confirming task execution, and transferring files around the cluster line by line carried to the map function.

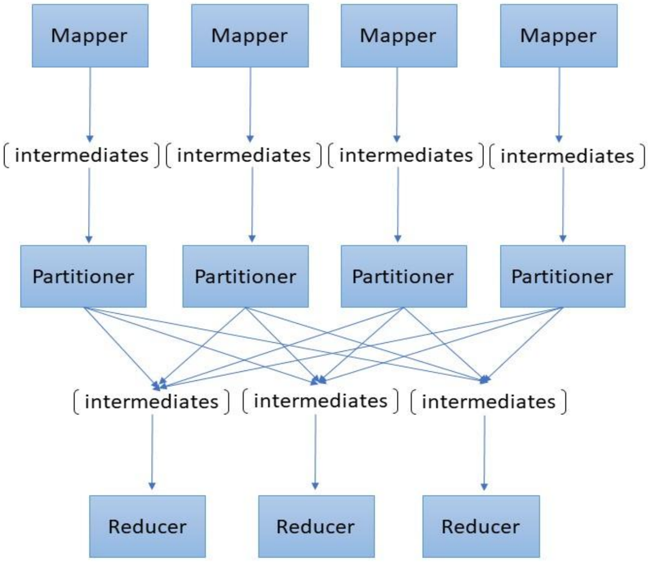

Figure 1 - MapReduce Architecture

2. Literature Review

2.1Apache Hadoop Ecosystem:

Apache Hadoop is good choice for big data analysis as it works for distributed big data. Apache Hadoop is an open-source software framework for distributed storage and large-scale distributed processing of data-sets on clusters. In short, Hadoop framework is capable enough to develop applications capable of running on clusters of computers and they could perform complete statistical analysis for a huge data and there are different tools each having different roles for extracting, storing and analyzing huge amount of data. Advantages of Apache Hadoop are Scalable, Cost-effective. Disadvantages are as issue with small files, vulnerable by nature.

2.2 Real-Time Analysis in Fashion Retail Industry:

This paper describes about the customer analytics in fashion retail industry in the era of big data.It mainly focuses on the challenges has been facing by fashion industries.Major target is to focus on customer analytics,big data customer behaviour in fashion retail industry. It predicts the customer analytics such as how to generate values in fashion industries and also strategies like customer acquisition,customer development,customer retention and acquisition,retention and optimisation. Coming to the methodology it mainlyfocuses on business understanding,types of data understanding,data pre-processing,modelling,evaluation and development. The strategies for customer analytics are inherently granular,behavioural,longitudinal,forward looking,broadly applicable and multi-platform. Customer analytics shows the features of the customer analysis,value generation in fashion retail industry,strategy,customer analytics methodologies.Customer data analysisin these datasets mainly focus on apparel industry from this data segments are created. To create these segments recencies are used so these recencies indicates value of each customer as the main indicator.Segments are based on the above recency'svalues for six months so these segments are classified into less active,active, highly active and new customer. To design and analysis of data three processes are used they are histogram of recency,frequency and purchase amount. Recency symbols how frequently a product has been purchased. when coming to frequency it defines about how often a product has been purchased. Purchased amount denotes how many times the payment has been done by a particular website or app. The main results of the fashion retail industry are to identify five different customer classes purchase behaviour over time including revenues so this helps to identify which segment of customer generates more value to the organisation and how long they can retain the period so finally it focuses on customer analytics in the era of big data and so that it could be more benefitted to the industries ever in the history.

2.3 Analysis of Twitter Data Using Apache Hive:

In today's competitive environment, businesses Wondering if customers will buy their products, find their products and applications Fun and easy to use, or the banking they need to understand how customers are transacting, big data analytics is one solution. Hence, Big Data Analysis is the process of examining large amounts of data The amount of data to discover hidden patterns or Unknown correlation, this article uses Apache Hive which is known as open-source data warehouse software for reading, writing and managing largedatasets and Pig is a scripting language similar to PigLatin.Hadoop and Hiveare standard-like with SQLlike queries.For hadoop, it allows developers to write Hive QueryLanguage (HQL). Hive is recommended for thoseFamiliar with SQL;Hive was originally created by Facebook later acquired by Apache SoftwareFoundation and continue as an open-source nameApache Hive. Built for OLAP, Hive is fast and scalableand an extensible query language, this is where the MapReduce architecture is applied to big data problems in small units of workand processed in parallel. The basic meaning of the mapReduce divides large tasks into smaller partsthentreat them appropriately, it will speed up the computationand improve system performance. card stage work. Map is a function that splits the input text of a task. This function takes key/value pairs as input and generatesA set of intermediate key/value pairs. and how the reduce function worksRely on merging all intermediate values to produce the final result. In Apache Hive architecture, it consists of user interface, metastore, Hive QL process engine, execution engine and HDFS. 1.User interface: Hive supports three user interfaces, such asWeb UI, command line and Hive HD Insight. Meta Store: stores schema or metadata for tables, table columns, their data types, and HDFS mapping databasesthe server is for Hive. HiveQL Process Engine: HiveQL is similar to SQLAlternative to traditional methods of MapReduceprogram execution engine. The execution engine is used to processQuery and generate MapReduce-like results.HBASE or HDFS: HBASE or Hadoop Distributed FileThe system is a data storage technology for storing data inFile system because this case study is about analysing Twitter data with Apache Hive, Hive queries are run on data stored in Hive tables, andCalculated results for many tweets. ThatThe result of running the Hive query on the HDInsight cluster isthe analysis is based on two parameters they are first is the Total MapReduce CPU time spent executing Hive queries and second is the total time required to run this job. Hive provides an easy-to-use platformDevelopers familiar with Map's SQL languageLess programming. this paperAlso discussed retrieving and executing Twitter tweetsHive queries on HDInsight clusters.

2.4 Map-Reduce Functionality in Retail Industry:

Now-a-days e-commerce sales are growing rapidly in such a way that the online transactions and sales are performed using point of sale(POS).Purchasing products using POS is mostly recommended to identify the purchase patterns of the customers.In POS map-reduce is used to reduce the size of the data and the data is divided into blocks.MapReduce works on two functions such as map() and reduce() where mapping() function is used to map data and count the frequency of the words and reduce()function is useful to reduce the frequency of the data by deleting the duplicate data and reducing the repeated data and forming the data as a single unit which reduces the size of the data.MapReduce is more scalable,maintains enough replication and data is distributed throughout the file systems to retrieve data from the distributed file system.It is also useful to relieve storage locations.POS works on the tool called IRM whichdetects the rise and decrease of the products demand.

2.5 Big Data Visualization Using Tableau:

Tableau is used for analysing large data sets, visualize and share knowledge. Tableau is used to provide facility to connect to different data sources with many systematic types they are data systems organised in file formats like CSV, JSON, XML, MS-EXCEL. It consists of two types of data systems such as relational data system and non-relational data systems coming non-relational data systems and some of the examples are PostgreSQL, MYSQL, SQL server, Mongo DB. Tableau also contains cloud systems like Oracle Cloud, Microsoft Azure, Google Big Query, AWS. Tableau also has a special feature of data blending which means it is a process of combining data from multiple sources. Tableau also consists of another important and unique feature and that is ability for collaboration in real time that makes it a valuable investment for commercial and non-commercial organizations. and some of the advantages are ease of use, security, advanced analysis and cost.

3. Proposed Methodology

Hive is a resource in the Hadoop Ecosystem that's also known as a data center Connectivity tool for structured data acquisition in Hadoop. Hive is used to sum up massive quantities of data which are collected in or put on top of Hadoop. Hive makes it simple to query and analyze big data to gain useful values.Hive was developed by facebook and finally acquitted as open source under the Apache

Software Foundation. Since Hive is an open-source database inventory tool, it's used by many companies for accessing and analyzing massive amounts of data. Aws Elastic MapReduce, for example, tends to make use of Hive.Hive is not at all like a relational database, which is designed for OLAP (online analytical processing) and is written in a programming language.

a) Architecture:

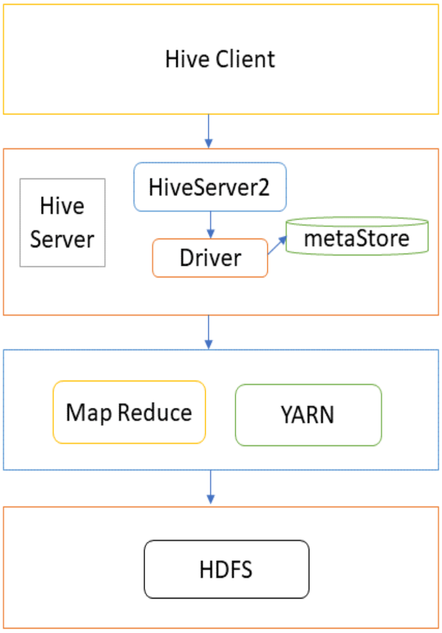

Figure 2 - Flowchartof Architecture

b) Proposed System:

Hive is a free to download and install for assessing and trying to verify large datasets stored in the Distributed Hadoop File System (HDFS). Hive is a query processing language that is parallel to SQL, then there's no need for new knowledge to query massive amounts of data using Hive. Type is processed as Structured data, Unstructured data, and Semi Structured data as it is collected from variety of sources and is available as message, clip, stereo, and pics. Hive is related to SQL in that it is simple to learn and operate queries on data to generate valuable insights because the data resides in HDFS. Hive is a tool in the Hadoop ecological unit a Map Reduce function. which parallelizes the computation of the raw data Hive utilises a query term known Hive Query Language, which is equivalent to SQL. Hive's key purposes are information recap, query processing, and analyzation. Hive QL serves as an interface, able to convert SQL commands into Hadoop Map-reduce jobs which were then implemented on Hadoop. Hive is composed of three main modules: the Meta Store, the Driver, and the Compiler.Meta Store this is a device for collecting metadata. Data regarding data is referred to as metadata. The meta store contains information such as location and schema of tables. When device is leaked, metadata serves as a back-up plan.Driver it receives Hive QL commands and starts acting as a controller.Compiler it is a code generator is used to generate Hive Query language expressions into Map reduce as required.To work with visualization techniques, a basic understanding of the Powerful query language is permitted. Connecting hive and Apache tez can provide real-time processing capabilities. For the complex network project, the project Management team uses a wide range of communication tools. email, video chat, and slack are the best among the known.The primary issue with all these messaging services is that swapping from one tool to another requires awhile. Hive performs as a centrally controlled communication tool, letting users to exchange messages via native online chat integration. The major challenges for the team are truthful and up - to - date on the moving parts of a project. The main challenges for the team are teaching about transparency and staying up to date on the project's moving parts. Consumers of hive benefit from functionalities such as agile data, extensibility, and marking to keep all of us up - to - date on the project. Though the various work layouts, hive makes it possible to structure tasks and projects. Customers have the opportunity of using a Project plan or a calendar board and a make - to - order board, and turn between them on the ease. Moreover, because new features are observable across all project views, the entire team is brought up to date. While working on complicated tasks, the project is broken down into tasks that are harder to monitor by team in order for the member of the project team.The steps to implement hive are such as data set gathering and collection, loading data into hive,storing and accessing the data into hive from HDFS, analysis of data usinghive, data visualization, output conclusion.

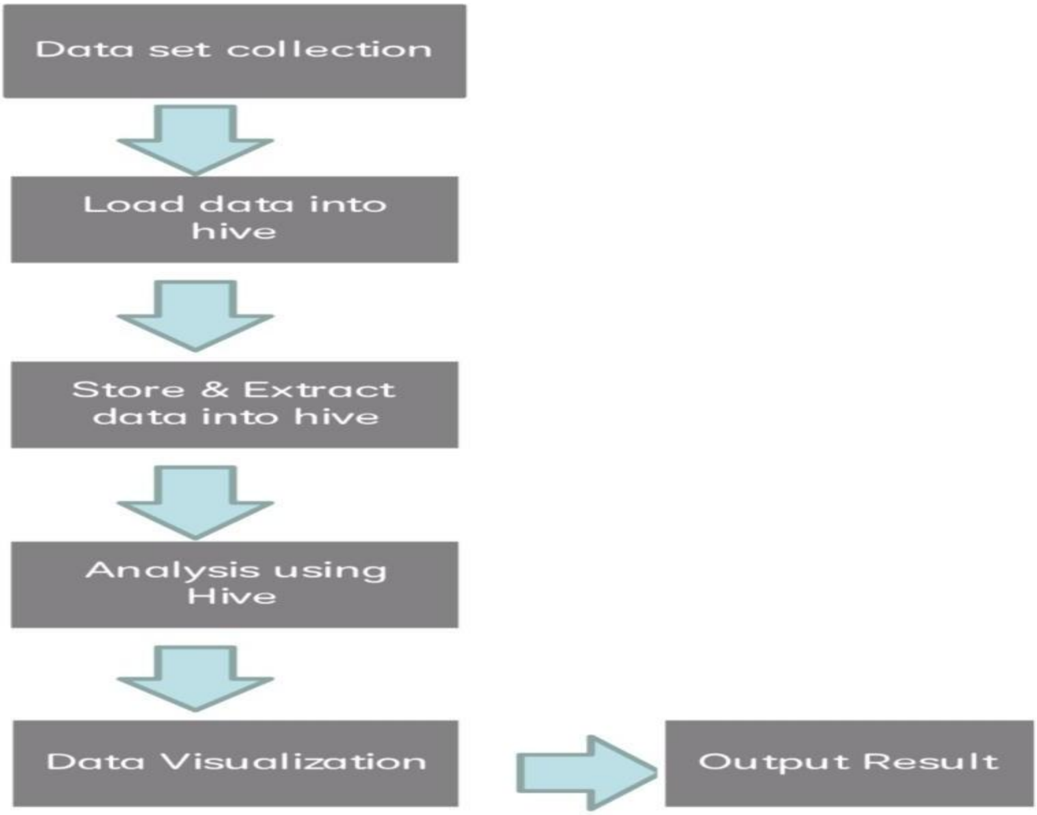

Figure 3 - Flowchart of Proposed System

Results

Based on the analysis of the placement of orders the results are concluded on the basis of these categories such as client categorization,month - to - month pattern prediction, analysis of 24/7 sales, product-based analysis, recommendation for payment, prospective user's location, supplier rating, observations into logistics-based minimization. Client categorization defines the amount spent by the customer.Month-to-month pattern predictionvisualize the trend of monthly sales.Analysis of 24/7 sales determines which hour have more number of sales.Product based analysisdefines the type of category which is more sold, which category product has more ratings, order count for each rating and it also determines the top 10 highest and least ratings.Recommendation for Paymentdescribes the type of payment and count of orders. Prospective user's location defines the location of the customers and describes how many customers are located in a particular area.Supplier ratingdescribe about which seller got more ratings and who sold more.Observations into logistics- based minimization determines that which city buys heavy weight products and low weight products.

Conclusion

Big data is not only a growing concept, but also a necessity. Various companiesincluding googlehave made significant investments in Big Data in recent years Amazon, Microsoftand Facebook will all be increasing their effectiveness, tacticand expertise. Hive is best suited for batch processingsuch as data warehouses. It is built on top of Hadoop, HDFSand map-reduce. Large amounts of data can be analysed in big dataand yet tools are required to forecast sales and profitability and requirement. Hive, HDFS and map-reduce are used to evaluate large datasets with great precision in predicting customer regular purchasing money transfers based on customer suggestions. Retailers provide recommendations based on the frequent point of sale transactions.

REFERENCES

1. Aditya Bhardwaj, Vanraj, Ankit Kumar, Yogendra Narayan, Pawan Kumar **“Big Data Emerging Technologies: A Case study with analyzing Twitter Data Using Apache Hive”** IEEE, December-2016.

2. J. Nandimath, E. Banerjee, A. Patil, P. Kakade, S. Vaidya and D. Chaturvedi, **"Big Data Analysis Using Apache Hadoop, “IEEE** 14th International Conference on Information Reuse & Integration (IRI),2018.

3. Chandadevi Giri, Sebastien Thomassey and Xianyi Zeng,**” Customer Analytics in Fashion Retail Industry**. International Journal of Scientific & Engineering Research, Volume 8, Issue 4, April-2019.

4. Neha Verma, Dheeraj Malhotra & Jatinder Singh, **“Big Data Analytics for Retail Industry Using MapReduce-Apriori Framework”**, Journal of Management Analytics,2020.

5. Ahmed M. Amer; Mohamed M. EL-Hadi.**” Tableau Big Data Visualization Tool In The Higher Education Institutions For Sustainable Development Goals “**IJCSMC, Vol. 8, Issue. 7, July 2019.

6. K.V.Shvachko, **“ The Hadoop Distributed File System Requirements”** Proceeding of the 2010 IEEE 26th Symposium on Mass Storage System and Technologies.

7. Samirana Aacharya, Bamrah Jagjit Kaur, Bandari Sharath Chandra, B. Vijaya Lakshmi **“Analyzing the frequently viewed videos from a YouTube log dataset using Apache Hive”** International Journal of Scientific & Engineering Research, Volume 8, Issue 4, April-2017.

8. N Sarkar, J Michael, V Kumar, R Chaurasia **“Hit Count Analysis using Big Data Hadoop”** – 2016.

9. Irida Gjermeni, D Hoxha,**“Big Data Analysis Use of E-Commerce Data”**. Control Theory and Informatics, 2019.

10. Dhruba, jssarma, jgray, amitanand.s , **“Apache Hadoop Goes Realtime at E- Commerce”**, Proceedings of the ACM SIGMOD International Conference on Management of data , June 2018.