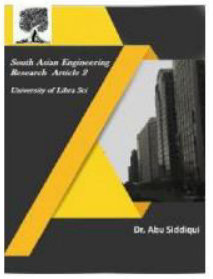
International Journal For Recent Developments in Science & Technology
2581-4575
Crossref
A Peer Reviewed Research Journal



BIG DATA ANALYSIS IN E-COMMERCE SYSTEM USING HADOOP MAP REDUCE THROUGH AMAZON

KANKATA VEMULA SAI KRISHNA [1], GUTHA APURUPA[2] , DEVAGANUGULA VIJAYA DURGA[3], BATTULA GIRIJA[4] , THUMU SUBBA REDDY[5]

[1,2,3,4] Student, Tirumala Engineering College, Jonnalagadda, Narasaraopet.

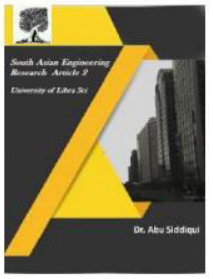
[5]Assistant Professor, Tirumala Engineering College, Jonnalagadda, Narasaraopet.

Abstract: Analysis of big data is a challenging task as it involves large distributed file systems. The infrastructure require for analyzing big data is different from Amazon analysis technology and data mining on various types of data. Map reduce is widely popular for analysis of big data. Map reduce is working with mapping, sorting, shuffling and reducing using Master/Slave architecture. Similarly Amazon Map Reduce programming model over large data set is introduced by Amazon, on the web especially used for ecommerce. In this paper Amazon EC2 cloud computing model used for central part of designed web and for collection and storing of large data Amazon uses S3. Amazon clusters is a group of servers which is working together to perform any type of tasks on distributed database on different servers in parallel. Amazon services are used in analysis of big data and to increase business efficiency

1.INTRODUCTION

As the technology get improving day by day the usage of the data and the increasing of the data becoming more and these create a lots of problems in the way of processing and the retrieving process, which leads to the big data challenges and these ever-growing data is becoming challenging to the traditional database system to access the content and process the data from that, her to overcome these challenges we need an innovation thoughts to make a gap between the data storage and the processing. Here comes the solution for this data challenges, big data tools and technologies offers to overcome and analyze the data effectively for the better user understanding and gain a competitive advantage in the present marketplace. These tools provide the better architecture to manage the traditional data-warehousing model to more complex architecture that addresses the more requirements, they are real-time processing, batch processing, structured and unstructured data. In the present market, many of the tools got evolved as a solution for big data challenges, but the most popular tool is Hadoop, here Hadoop provides data processing and also storage of the data in distributed file systems. It provides a better solution in processing of a large amount of the data. Hadoop provides many eco-systems for data processing and has many

International Journal For Recent Developments in Science & Technology
2581-4575
Crossref
A Peer Reviewed Research Journal



services like the hive, pig, sqoop, Hbase, Cassandra, etc. Each component has its own specialty in the processing. Increasing of the users' usage for the data will effects on the servers which will not provide the proper service to the end-user, when the OLTP Online transaction process is happening the availability of the data needs to be improved and it should be in secure mode, for processing of the data with high availability here amazon web services came into picture with number of services. Amazon web services(AWS), it provides the many services which are available same as the Physical services are completely implemented in the amazon services, these works as a · Platform as a service · Software as a service · Infrastructure as a service These are the above services amazon provides on-demand and whenever we required amazon provides, if we don't want to use the service then we can opt-out from the service, which will help in the cost-effective levels. 1 Amazon also provides a big data solution as in service of EMR (Elastic map-reduce) which is works as a Hadoop. EMR provides a seamless scale end-end big data applications. AWS provides the infrastructure and tools to tackle big data projects. Here no infrastructure is required and no maintenance is required amazon takes care of the whole setup and gives us as a service. To analyze a large amount of the data with the required setup is not possible for the computer capacity may vary when the amount of the data increased in the processing and at that time it will effects on the allocation of the data in the HDFS when we come to the data storage in the AWS service. When the data is expanded more than the computer capacity, then on-demand it automatically increased the services. In addition to this, the data gets more the flexibility of the availability in the regions is provided by the regions in the amazon web services When the data increasing is they're the horizontal addition of the system is made easy in this service, with the help of these services the data can be accessed and processed from anywhere when we have a system with an internet connection is enough for this. This service also disables when the analysis is not performing and can run the service by creating a Cron, which is a scheduling tool helps to run the processing deployment whenever required and processing will happen there. The capabilities of the AWS platform make it ideal for big data solution and analytics. Her for more information we have a document provided by amazon web services. Below are some of the big data analytics services provided by amazon.

· Amazon kinesis
· AWS Lambda
· Amazon Glue
· Amazon DynamoDB
· Amazon Redshift ·

Amazon Elastic Map Reduce Along with this amazon provides the EC2 instance for the storage of the data, for the self-managed big data applications.

2.EXISTING SYSTEM

The existing system uses Hadoop and Hive on Hadoop's environment. The system has

International Journal For Recent Developments in Science & Technology
2581-4575
Crossref
A Peer Reviewed Research Journal

three stages: 1) Dataset generation 2) Usage of Hadoop 3) Usage of Hive for Analyzing

2.1 Dataset generation In this stage of dataset generation data can be extracted from the following url: https://data.world/datasets/open-data

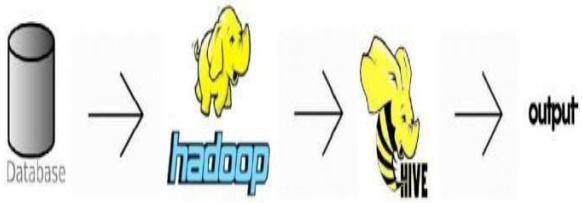


Fig. 1: Dataset Generation Existing System Figure

2.2 Hadoop In this stage, the Hadoop environment is being set up by, installing Hadoop software on system. Hadoop file system supports the distributed processing and allows the massive data in compressed format, which can be facilitates the big data processing tasks.

2.3 Hive In this stage, Hive is being installed on Hadoop environment. Data is loaded into hive. Queries are executed according to our obligation. Examined data is being displayed on the terminal. 3.1.4 Limitations The persisting system uses Hadoop and Hive on Hadoop environment. This model does the great work with large datasets and produces high accuracy in analyses data. The one drawback of this model is that do not visualize the analyzed data obtained as production. We need to use any of the visualization tools for visualization.

3. PROPOSED SYSTEM

The proposed system makes an up gradation to the existing model by presenting a new visualization method. The proposed system addresses the issues of existing model. The main aim of this system is to visualize the examined data.

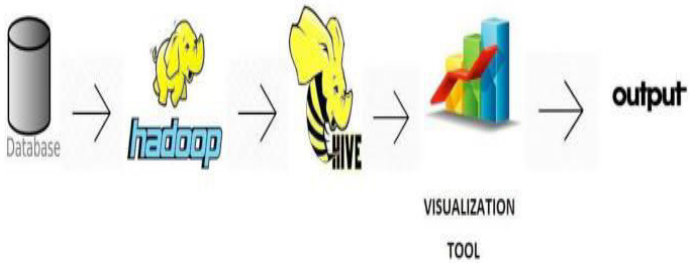


Figure 2: Dataset Generation Proposed System Figure

3.2.1 Working of the proposed system The proposed system employs a visualization tool. After the data is analyzed by performing hive queries Hive installed on Hadoop environment, it is imported into the visualization tool and been visualized in any of the forms mentioned below. · Column Chart · Line Chart · Pie Chart · Doughnut Chart · Bar Chart · Area Chart · XY (Scatter) Chart · Bubble Chart · Stock Chart · Surface Chart · Radar Chart · Combo Chart Here comes the beauty of the Amazon web service for the big data analysis, EMR provides great services which can provide the Availability, elasticity, cost-effective, High Availability, Security. Before the Amazon service, as it struggled a lot for the environment setup for the Hadoop Architecture and 14 faced many issues and wasted lots of time in spending the time for the setup when we know about the amazon web services, it helped a lot in the environment setup. EMR it provides the all the services which will be provided by the Hadoop and overcome many issues which have faced in the local server setup, now it ready to retrieve any amount of the data and

International Journal For Recent Developments in Science & Technology
2581-4575
Crossref
A Peer Reviewed Research Journal

done with the dynamical allocation of the storage when it's out of the bound. Steps for EMR setup: 1) Create Amazon Account 2) Open Amazon EMR service a 3) Click on Create Cluster and follow the steps for the creation 4) create a key pair to login using the PUTTY 5) Select number of instance 6) And it's Done.

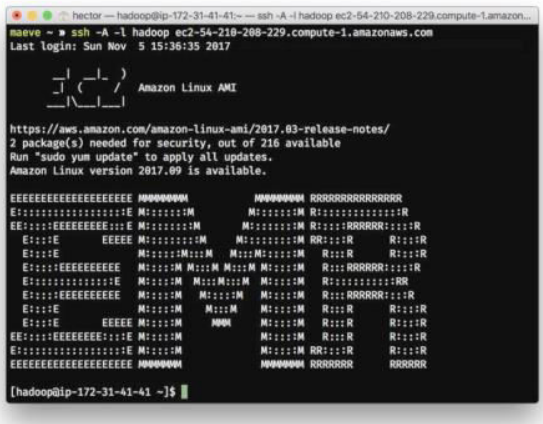
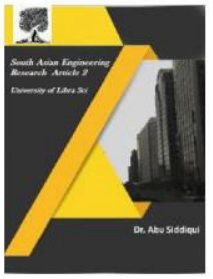


Figure 3: EMR Setup

The above screen is how it shows when we login to EMR using putty, if everything went right. From here we can enter into any Hadoop component and access the HDFS storage. For our requirement we need the hive for the data loading and create a schema, Type hive in EMR cli to enter into hive cli. Creating schema: 15 Create table using sample data in S3. Note: you can replace this S3 path with your own. CREATE EXTERNAL TABLE IF NOT EXISTS logs (DateObject Date,Time STRING,Location STRING,Bytes INT,RequestIP STRING,Method STRING,Host STRING, Uri STRING,Status INT,Referrer STRING,OS String,Browser String,BrowserVersion String) ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.RegexSerDe' WITH SERDEPROPERTIES ("input.regex" = "^(?!#) ([^]+)\\s+([^]+)\\s+([^]+)\\s+([^]+)\\s+([^]+)\\s+([^]+)\\s+([^]+)\\s+([^]+)\\s+([^]+)\\s+([^]+)\\s+([^]+)\\s+([^]+)\\s+([^]+)\\s+[^\(]+[\(]([^\;]+).*\%20([^V]+)[V](.*) $") LOCATION '${INPUT}/twitter/data';sample schema for the log analysis for the preprocessing.

4.IMPLEMENTATON

There are three dissimilar components for performing various tasks in this work. The three modules and their description is as follows: 1) Hadoop Installation 2) Hive Installation 3) Visualization Tool 5.1. Hadoop installation After the data related to Data World are retrieved using the connection https://www.lendingclub.com/info/ This module is used to prepare an atmosphere for data investigation. Hadoop is an open-source distributed treating framework written in Java, that manages data processing and storage for big data applications running in clustered systems. It is at the center of a rising ecosystem of big data technologies that are principally used to provide advanced analytics initiatives, including all predictive analytics, data mining and machine learning applications. Hadoop can handle various forms of structured and unstructured data, giving users more flexibility for collecting, processing and analyzing the data than relational databases and data warehouses provide. It comprises of several components that allow the storage and processing of

International Journal For Recent Developments in Science & Technology
2581-4575
Crossref
A Peer Reviewed Research Journal


large data capacities in a clustered environment.

List of Hadoop Ecosystem HDFS – Hadoop Distributed File System. Hive – Data Query System. Pig – Data Query System. MapReduce – A data processing Layer. HBASE – Columnar Store. HCatalog – Data Storage System. YARN – Yet Another source Navigator. Avro.

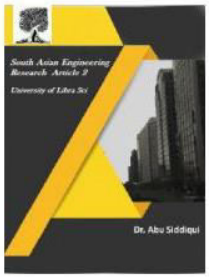
Hive Installation This section deals with the installation of Hive on Hadoop platform. Hive is an open-source data warehouse software project built on top of Hadoop for providing data queries and examining large data 20 sets that are mainly stored in Hadoop files. Hive gives an SQL-like border to query data stored in various folders and file systems that integrate with Hadoop. Traditional SQL queries must be applied in the Map-Reduce Java API to execute SQL applications and questions over distributed data. Hive provides the necessary SQL generalization to integrate SQL-like queries (HiveQL) into the original Java without the important to implement queries in the low-level Java API. Since most data warehousing applications work with SQL-based querying languages, Hive aids portability of SQL-based applications to Hadoop. After installing Hive on Hadoop ecosystem queries according to users requirement are being executed. The data is examined and displayed after this component. 5.3. Visualization tool After obtaining the examined data from the above units it is imported to an imagining tool to visualize. Here, the visualization tool we used is Microsoft Excel 2007. You can be displaying your data analysis reports in various ways in Excel. However, if your data analysis results can be visualized as graphs that highlight the notable points in the data, your audience can rapidly grasp what you want to project in your data. It also leaves a good effect on your presentation style.

List of charts with HIVE Graphical Interface Column Chart. Line Chart Pie Chart Doughnut Chart Bar Chart Area Chart XY (Scatter) Chart Bubble Chart Stock Chart Surface Chart Radar Chart Combo Chart Here, we use pie chart for visualizing examined data among the all above revealed options. The visualized data is then showed to the user as a final output.

5.RESULT

Data Visualisation Data Visualisation plays an important role in today's industry where everything needs to be impactful and easier to understand. In our project also instead of representing the data, in just tabular format we have made an effort to represent the results more effectively, which is very easy to understand, do not need any technical knowledge, any person could understand by viewing the figures. The project makes use of Python Spark which includes an inbuilt Zeppelin Visualisation tool, using this tool we can represent the outcomes of the project in desired visualisation format. The sample data is obtained from IBM which consists of 1 lakh records based on online sales, which we are using for our project and running queries on these data and results can be in seen in the python spark. We have received the data in the form of csv and process it in

International Journal For Recent Developments in Science & Technology
2581-4575
Crossref
A Peer Reviewed Research Journal



Spark In this data we are going to analyse the data like Retailer Country, Order Method Type, Retailer Type, Product Line, Product Type, Product, Year, Quarter, Revenue, Quantity, Gross margin.

6.CONCLUSION AND FUTURE WORK This project work concentrates on effective processing E-commerce big data using proposed distributed visualization methodology. Its methodology effectively works for query processing on e-commerce data for data analysis of sales and profit reporting with multi-dimensional perspective. Experimental results are demonstrated with various visualizations in the interfacing of HIVE with matplot modules on large dimensional big datasets. Future work of this project is to extend for classification analysis with various scalable machine learning algorithms.

REFERENCES

[1]. Pradeep Adluru ; Srikari Sindhoori Datla ; Xiaowen Zhang, "Hadoop eco system for big data security and privacy", Long Island Systems, Applications and Technology, IEEE Explore, 2015

[2]. Wencheng Sun; Zhiping Cai, "Data Processing and Text Mining Technologies on Electronic Medical Records: A Review", J Healthc Engg., 2018

[3]. Taiwo Kolajo; Olawande Daramola, "Big Data Stream Analysis: a systematic literature review", Journal of Big Data, Vol.6, Issue. 47, 2019

[4]. Salman Salloum; Ruslan Dautov, " Big Data Analytics on Spark", Int. Journal Data Science Anal. 2016

[5]. Ram Sharan Chaulagain ; Santosh Pandey ; Sadhu Ram Basnet ; Subarna Shakya, "Cloud Based Web Scraping for Big Data Applications", IEEE International Conference on Smart Cloud (SmartCloud), 2017

[6]. Krishna Das ; Smriti Kumar Sinha," Essential pre-processing tasks involved in data preparation for social network user behaviour analysis", International Conference on Intelligent Sustainable Systems (ICISS), 2017

[7]. Sitaram Asur, Bernardo, "Predicting the Future with Social Data, IEEE Int. Conf. on Web Intelligence and Intelligent Agent, 2010

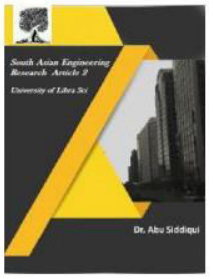
[8]. Ashish Juneja ; Nripendra Narayan Das, "Big Data Quality Framework: Pre-Processing Data in Weather Monitoring Application", International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), 2019

[9]. Kyounghyun Park ; Minh Chau Nguyen ; Heesun Won, "Web-based collaborative big data analytics on big data as a service platform",17th International Conference on Advanced Communication Technology (ICACT), 2015

[10]. Kumari Punam ; Rajendra Pamula ; Praphula Kumar Jain, "A Two-Level Statistical Model for Big Mart Sales Prediction", International Conference on Computing, Power and Communication Technologies, 2018

[11]. Shakila Shaikh ; Sheetal Rathi ; Prachi Janrao, "Recommendation System in E-Commerce Websites: A Graph Based Approached", IEEE 7th International

International Journal For Recent Developments in Science & Technology
2581-4575
Crossref
A Peer Reviewed Research Journal



Advance Computing Conference (IACC), 2017

[12]. Hongyong Yu ; Deshuai Wang, "Research and Implementation of Massive Health Care Data Management and Analysis Based on Hadoop", Fourth International Conference on Computational and Information Sciences, 2012

2581-4575

International Journal For Recent Developments in Science & Technology

Crossref

A Peer Reviewed Research Journal