JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

A STUDY ON PRACTICE OF HADOOP ORGANISM IN ENORMOUS DATA

GOWSALYA.S, Dr.V.RAMYA, K.SUKITHA

ASSISTANT PROFESSOR

SRI BALAJI ARTS AND SCIENCE COLLEGE

ABSTRACT

Big data is a term for massive data sets having large, more varied and complex structure with the difficulties of storing, analyzing and visualizing for further processes or results. The process of research into massive amounts of data to reveal hidden patterns and secret correlations named as big data analytics. These useful information for companies or organizations with the help of gaining richer and deeper insights and getting an advantage over the competition. For this reason, big data implementations need to be analyzed and executed as accurately as possible. This paper presents an overview of big data's content, scope, samples, methods, advantages and challenges and discusses privacy concern on it.

Keywords: Hadoop, Analytics, Cloud, Distributed System, Map Reduce, Big Tables

INTRODUCTION

BIG DATA is a vague topic and there is no exact definition which is followed by everyone. Data that has extra-large Volume, comes from Variety of sources, Variety of formats and comes at us with a great Velocity is normally refer to as Big Data. Big data can be structured, unstructured or semi-structured, which is not processed by the conventional data management methods. Data can be generated on web in various forms like texts, images or videos or social media posts. In order to process these large amount of data in an inexpensive and efficient way, parallelism is used [1]. There are four characteristics for big data. They are Volume, Velocity, Variety and Veracity.Volume means scale of data or large amount of data generated in every second. Machine generated data are examples for these characteristics. Nowadays data volume is increasing from gigabytes to petabytes [2]. 40 Zettabytes of data will be created by 2020 which is 300 times from 2005 [3]. Second characteristic of Big Data is velocity and it means analysis of streaming data. Velocity is the speed at which data is generated and processed. For example social media posts [2]. Variety is another important characteristic of big data. It refers to the type of data. Data may be in different forms such as Text, numerical, images, audio, video, social media data [2]. On twitter 400 million tweets are sent per day and there are 200 million active users on it [3].Veracity means uncertainty or accuracy of data. Data is uncertain due to the inconsistency and incompleteness [2].

HADOOP :

Hadoop is an open-source software structure that ropes dataintensive distributed applications. It allows applications to work with thousands of computationally self-governing computers and with petabytes of data. Hadoop increases the storage space and the processing power by uniting many computers into one. Hadoop devises two parts: HDFS (Hadoop Distributed File System) file system and MapReduce programming paradigm. It has been formed by Dong Cutting and Mike Cafarella in 2005. It was developed to upkeep distribution for search engine project. It is

certified under APACHE LICENSE 2.0. This is written in Java Runtime Environment (JRE) 1.6 or advanced version. The operating system is cross-platform. It was developed by Apache Software Foundation. Hadoop came as a derivative from Google's Map Reduce and Google File System (GFS). The principal of Apache Hadoop consists of a storage part (Hadoop Distributed File System (HDFS)) and a processing part (MapReduce). Hadoop separates files into large blocks and allocates them among the nodes in the cluster. To work on the data, Hadoop MapReduce handovers packaged code for nodes to process in parallel, based on the data each node needs to process. This approach takes advantage of data locality—nodes manipulating the data that they have on hand—to allow the data to be processed faster and more efficiently than it would be in a more predictable supercomputer architecture that depends on on a parallel file system where computation and data are connected via highspeed networking. The HDFS is a distributed file system that provides fault tolerance and is designed to run on commodity hardware. HDFS delivers high throughput access to application data and is appropriate for applications that have large data sets. Hadoop provides a distributed file system (HDFS) that can store data across thousands of servers, and a way of running work (ie; Map/Reduce jobs) across those machines, running the work near the data. HDFS devises master/slave architecture. Large data is automatically split into chunks which are managed by different nodes in the Hadoop cluster.

HADOOP FRAMEWORK

Hadoop is open any one software used to process the Big Data. It is very famous used by administrations/researchers to analyze the Big Data. Hadoop is influenced by Google's structural design, Google File System and MapReduce. Hadoop procedures the large data sets in a spread calculating environment.

HADOOP CONTAINS OF TWO MAIN MECHANISMS:

STORING:

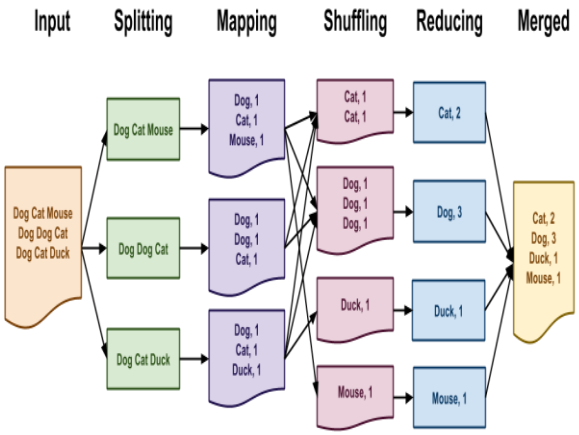
The (HDFS)Hadoop Distributed File System: These are dispersed file system which brings responsibility taking and measured to run on creation hardware. HDFS brings high amount entree to application data and is suitable for requests that have vast data sets. HDFS can stock data over thousands of servers. HDFS has master/slave construction [5]. Files added to HDFS are separated into fixed-size masses. Mass size is configurable, but avoidances to 64 megabytes.

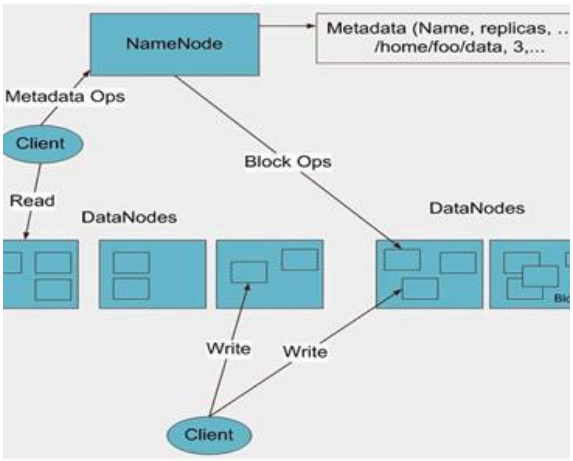
Processing: Map Reduce:

It is a software project classical presented by Google in 2004 for effortlessly writing applications which procedures enormous volume of data in equivalent on huge bunches of hardware in responsibility. This functions on huge data set, separations the problem and data sets and run it in equivalent way. Two utilities in MapReduce are as following:

a) **Map** - The Map function always runs first typically used to filter, transform, or parse the data. The output from Map becomes the input to Reduce.

b) **Reduce** - The Reduce function is optional normally used to summarize data from the Map function.





APPLICATIONS IN DATA MINING:

Big Data is very useful for Business Organizations as well as to the researchers to observe the data patterns in big data sets. Extracting useful information from large amount of big data is called as Data Mining. There is huge amount of data on Internet in form of text, numbers, social media posts, images and videos. 40 Zettabytes of data will be created by 2020 which is 300 times from 2005 [3]. To analyze this data to get useful information for security, health, education etc., we need to introduce new data mining system which is effective. There are many Data mining techniques which can be used with big data, some of them are:

A. Classification Analysis:

It is a systematic process for obtaining important information about data and metadata. Classification can also be used to cluster the data.

B. Cluster Analysis:

It is the process to identify data sets that are similar to each other. This is done to get the similarities and differences within the data. For example clusters of customers having similar preferences can be targeted on social medial [6].

C. Evolution Analysis:

It is also called as genetic data mining mainly used to mine data from DNA sequences. But can be used in Banking, to predict the Stock exchange by previous years' time series Data [7].

D. Outlier Analysis:

Some observations, identifications of items are done which do not make a pattern in a Data Set. In medical and banking problems this is used.

LITERATURE REVIEWS

Anupam Jain, Rakhi N K and Ganesh Bugler studied Indian Recipes and discovered that the presence of certain spices makes a meal much less likely to contain ingredients with flavors in common. Jain and others chose an online website TarlaDalaa.com and downloaded more than 2500 recipes for their research. 194 different ingredients were found in these recipes. Then they studied Network of links between these recipes. They found that Indian cuisine is characterized by strong negative food pairing that even higher than any before. According to them, "Our study reveals that spices occupy a unique position in the ingredient composition of Indian cuisine and play a major role in defining its characteristic profile". "Our study could potentially lead to methods for creating novel Indian signature recipes, healthy recipe alterations and recipe recommender systems," conclude Jain and mates [8,9].

Vidyasagar S. D did a survey on Big Data and Hadoop system and found that organizations need to process and handle petabytes of Data sets in efficient and inexpensive manner. According to him if there is any node failure then we can lose some information. Hadoop is an Efficient, reliable, Open Source Apache License. Hadoop is used to deal with large data sets. Author explained its need, uses and application. Now days, Hadoop is playing an important role in Big Data. Vidyasagar S.D concluded that "Hadoop is designed to run on cheap commodity hardware, it automatically handles data replication and node failure, it does the hard work – you can focus on processing data, Cost Saving and efficient and reliable data processing.

CONCLUSION AND FUTURE SCOPE

By the above comparative survey we have come to know that HADOOP is the best technique for handling Big Data compared to that of RDBMS. As world moves on, the data used increases and therefore a better way of handling such huge amount of data is becoming a tedious task. Analysis and storage of the so called Big Data is handy only by the help of the new Hadoop eco-system than the traditional RDBMS being used till now. Hadoop is a large scale, open source software framework dedicated to scalable, distributed, dataintensive computing. The framework breaks up large data into smaller parallelizable chunks and handles scheduling, maps each piece to an intermediate value, Fault tolerant, reliable, and supports thousands of nodes and petabytes of data, tried and tested in production, many implementation options.

REFERENCES:

[1] Harshawardhan S. Bhosale, Prof. Devendra P. Gadekar "A Review Paper on Big Data and Hadoop" in International Journal of Scientific and Research Publications, Volume 4, Issue 10, October 2014.

[2] SMITHA T, V. Suresh Kumar "Application of Big Data in Data Mining" in International Journal of Emerging Technology and Advanced Engineering Volume 3, Issue 7, July 2013).

[3] IBM Big Data analytics HUB, www.ibmbigdatahub.com/infographic/four-vs-big-data.

[4] Mrigank Mridul, Akashdeep Khajuria, Snehasish Dutta, Kumar N "Analysis of Bidgata using Apache Hadoop and Map Reduce" in International Journal of Advance Research in Computer Science and Software Engineering, Volume 4, Issue 5, May 2014.

[5] Apache Hadoop Project, http://hadoop.apache.org/, 2013.

[6] Smitha.T, Dr.V.Sundaram, "Classification Rules by Decision Tree for disease prediction" International journal for computer Application, (IJCA) vol 43, 8, No-8, April 2012 edition. ISSN0975-8887; pp- 35-37

[7] Mucherino A. Petraq papajorgji P.M.Paradalos 1998. A survey of data mining techniques alied to agriculture CRPIT.3(3): 555560.

[8] Anupam Jain, Rakhi N K and Ganesh Bagler, arxiv.org/abs/1502.03815 Spices Form The Basis Of Food Pairing In Indian Cuisine.

[9] MIT Technology Review, http://www.technologyreview.com/view/535451/data-mining-indian-recipes-reveals-new-food-pairing-phenomenon/.

[10] Vidyasagar S. D, A Study on "Role of Hadoop in Information Technology era", GRA - GLOBAL RESEARCH ANALYSIS, Volume : 2 | Issue : 2 | Feb 2013 • ISSN No 2277 –8160.

[11] BIG DATA: Challenges and opportunities, Infosys Lab Briefings, Vol 11 No 1, 2013.

[12] Divyakant Agrawal, Challenges and Opportunities with Big Data, A community white paper developed by leading researchers across the United States.

[13] Puneet Singh Duggal, Sanchita Paul, Big Data Analysis: Challenges and Solutions in International Conference on Cloud, Big Data and Trust 2013, Nov 13-15, RGPV.

[14] Big Data, Wikipedia, http://en.wikipedia.org/wiki/Big_data Webster, Phil. "Supercomputing the Climate: NASA's Big Data Mission". CSC World. Computer Sciences Corporation. Retrieved 2013-01-18.