Middle-East Journal of Scientific Research 24 (9): 2734-2738, 2016
ISSN 1990-9233

DOI: 10.5829/idosi.mejsr.2016.24.09.23804

Tools and Platforms for Big Data Analytics in Mobile Telecommunication: A Review

[1]H. Karthikeyan and [2]T. Menakadevi

[1]Department of IT, Adhiyamaan College of Engg., Hosur - 635109, Tamil Nadu, India
[2]Department of ECE, Adhiyamaan College of Engg., Hosur - 635109, Tamil Nadu, India

Abstract: Big data is a collection of huge set of data with different types. Data is a raw material and an asset to improve a business. Big data analytics provides an opportunity to improve the business by examining the huge amount of data.In recent days Mobile Network Users and Mobile Network Service Providers are increasingly high. Now a days in India Telecom Regulatory Authority of India (TRAI) has given an option to Mobile Network Users (MNU) to change the Mobile Network Service Provider (MNSP). For this reason MNSP should identify their customers' needs to retain them in own network. Big data analytics provides an opportunity to identify and extracts customer expectations by analyzing customer data. Big data analytics will be useful to predict the network usage and to enrich the business opportunities.There are some issues and challenges of processing, analyzing and storing the big data. This paper focuses on architecture of Hadoop framework, components and comparisons of variousplatforms and also comparisons of different clustering algorithms forBig Data Analytics.

Key words: Big Data Analytics · Clustering Algorithms · Hadoop · K-means Algorithm

INTRODUCTION

Big data analytics refers to the process of collecting, organizing and analyzing large set of data to discover and develop data-driven business model. Big data is used to discover hidden patterns, market trends customer preferences and other useful information. Data will be generated from different sources such as sensor networks, company market lead databases, web-logs and social networking sites. The generated data is in different formats either structured or unstructured. Representation of data in tabular format (RDBMS) is an example for structured data and text, audio and video is an example for unstructured data. Big data analytics uses both structured and unstructured data for analytics.

National Oceanic and Atmospheric Administration (NOAA) uses big data analytics to assist with climate, ecosystem and environment, weather forecasting and pattern analysis and commercial translational applications. Big data is used to predict the fraud detection in banking industry, Wal-Mart handles more than 1 million transactions per hour and it contains more than 2.5 petadata and Pharmaceutical companies are using big data analytics for drug discovery, analysis of clinical trial data side effects and reactions.

In recent years, usage of mobile phones, number of MNU and MNSP are increasing rapidly. India is currently the second-largest telecommunication market and has the third highest number of internet users in the world. The total mobile services market revenue in India is expected to touch Rs.2, 51, 300 crores in 2017, registering a Compound Annual Growth Rate (CAGR) of 5.2 per cent between 2014 (ie, from 1, 28, 729 crores to 2, 51, 300 crores)and 2017, according to research firm International Data Corporation (IDC). Mobile Network Users are expecting effective service from the MNSP. This is a challenging task for the MNSP to satisfy the user`s expectations and retain the MNU with their own MNSP. In tradition approach MNU can change the MNSP using Mobile Number Portability introduced by TRAI in India but the Mobile Network User has to retain minimum three months on that Network [1].

The Platforms for big data analytics was compared in "A survey on platforms for big data analytics" The author had described various platforms based upon scalability, Data I/O Performance and Fault tolerance [2]. Big Data Mining model was proposed in "Data Mining with Big Data", The author had illustrated data- driven model and data analysis from Big Data.

Corresponding Author: H. Karthikeyan, Department of IT, Adhiyamaan College of Engg., Hosur - 635109, Tamil Nadu, India.

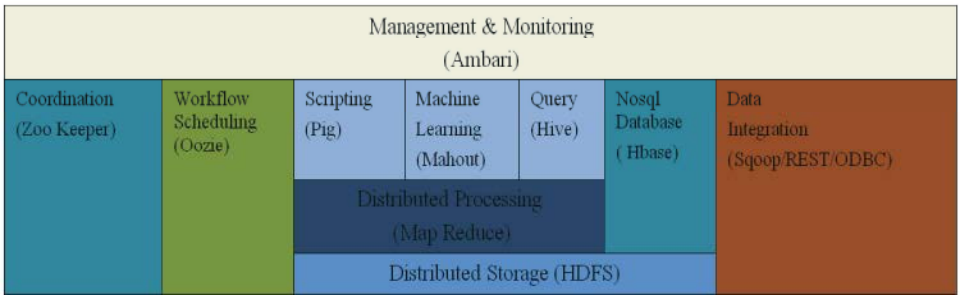

Fig. 1: Apache Hadoop Eco System

1010data Inc. is a market leader in Big Data Discovery and data staring it uses columnar databases for analyzing and it uses Hadoop framework it's a proprietary design with its own query language that supports a subset of SQL functions. IBM has the broadest data-management portfolio and it uses DB2 and Netezza for Analytical Databases. Also it uses InfoSphere for stream-processing.

In General Big data can be characterized into 5Vs

- Volume: represents the size of data.
- Velocity: represents Increasing speed of data.
- Variety: represents the structure of the data.
- Veracity: represents the quality of data.
- Value : represents hidden values in the large set of data

The remainder of the paper is structured as follows. Section II deals with Literature review ofHadoop Framework for Big Data Analytics, Section III describes Data Mining Algorithms used in Big Data Analytics. Section IV discusses the comparative study of different frame work. Section V summarizes the conclusion.

Hadoop Framework Description: The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models and it is a part of Apache project sponsored by Apache Software Foundation. Doug Cutting is the creator of Hadoop named the framework after his child`s stuffed toy elephant. Hadoop is used by many companies like Google, IBM and Yahoo.

For our proposed Work Apache Hadoop Framework will be used to implement. This paper reports the core elements of Hadoop and its usage. Fig.1represents arrangements of components in Hadoop Eco System [2].

Hadoop [3] [4] is a framework which has two main Components.

- Hadoop Distributed File System (HDFS)
- Hadoop MapReduce

HDFS allows collecting data from various data centers or various clients and added into clustersas shown in Fig. 2. HDFS works by breaking large files into smaller pieces called blocks. The blocks are stored on data nodes and it is the responsibility of the Name Node to manage all access to the files. HDFS requests the file name to the HDFS Name Node and Name Node responses with Block id and location of the file. Then client uses that block id to access the data from HDFS data node. HDFS uses transaction logs and checksum validation to ensure integrity across the cluster.

Transaction logs are used to track every operation for rebuilding of the file system. Data blocks are replicated across several data nodes to avoid data loss. Hadoop MapReduce allows processing the data in parallel and results are stored in a single cluster. Hadoop MapReduce composed of map function and reduce function. Hadoop map function performs filtering and sorting operation from various clusters then reduce function performs summarization operation and results are stored in HDFS as shown in Fig. 3. Hadoop Map-Reduce will split the data and sort the data by dividing the file into splits, it allows several map tasks to operate on a single file in parallel. The Record Reader (RR) loads data from its source and converts it into (K, V) pairs suitable for reading by Mappers. The Mapper performs the user-defined work of the first phase of the MapReduce program.

The Reducer performs the user-defined work of the second phase of the Map Reduce program. The output format defines the way (K, V) pairs produced by Reducers are written to output files.Whereas K-Key and V-Value. Apache Hadoop Eco system components are tabulated along with its usage in Big data analytics in Table 1.

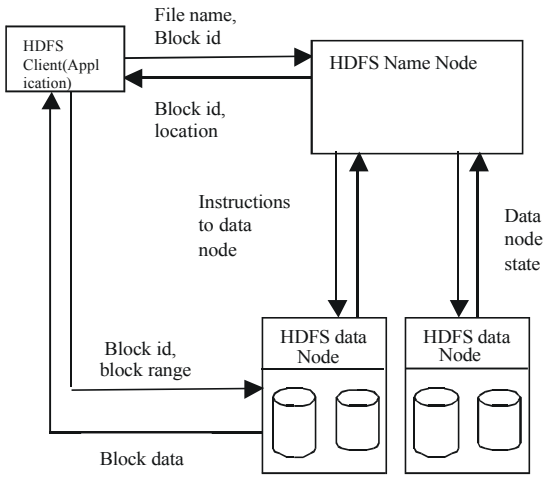

Fig. 2: Basic architecture of HDFS in Hadoop

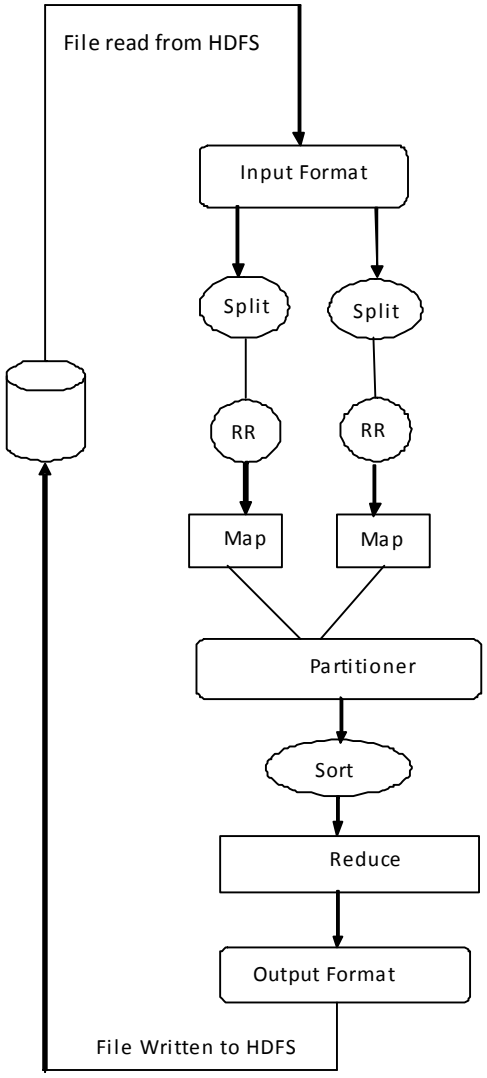

Fig. 3: Hadoop MapReduce Processing

Table 1: Hadoop Ecosystem and its usage

Hadoop EcoSystem	Usage
Hadoop Distributed File System (HDFS)	• Distributed file system to access the data
MapReduce	• For Data processing
Apache Pig	• Designed for performing a long series of data operations such as Extract-transform-load and Iterative data Processing.
Apache Hive	• Data warehouse system for Hadoop that facilitates easy data summarization. • Provides a mechanism to query the data ware house SQL(Structured Query Language).
Apache Mahout	• Scalable machine learning. • Supports of different data mining algorithms such as clustering, classification
Apache HBase	• Column oriented data base and Hbase uses Hadoop HDFS
Apache ZooKeeper	• Configuration Management supports for naming, distributed synchronization
Ooize	• Work flow co-ordination systems.
Sqoop	• To transfer data from Hadoop to structured data stores

Data Mining Algorithms: Data mining algorithms are needed on the data to do statistical analysis. There are many data mining techniques and algorithms are there. Such as Cluster Analysis, Classification Analysis, Associationanalysis and Prediction analysis [5]. Clustering Algorithms divide their data into small segments which will used to identify similar customers are grouped into one cluster.

Amazon e-commerce uses recommendation algorithms to recommend a product to a user based upon existing wish-list and users search this is one way of following clustering technique to reduce large(X) to small data sets (y1, y2..) [4]. Machine learning algorithms for query processing was described in "MapReduce: Distributed Computing for Machine Learning". The author had reported the performance on searching and sorting tasks to investigate the effects of various system configurations in Map Reduce.

K-Means Algorithm: It is an unsupervised learning algorithm data sets are classified as clusters. The main idea in k-means algorithms is for each cluster it should be a centroid and results may be vary based upon chosen centroid.

$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c^i} \left(|x_i - v_j| \right)^2 \quad (1)$$

c - Number of cluster center
c_i - Number of data points in the i^{th} cluster
$|x_i-v_j|$- Euclidian distance between x_i and v_j

K-Medoids Algorithm: it chooses data points as centersand works with an arbitrary matrix of distances between data points.

Agglomerative Hierarchical Clustering: This algorithm works by grouping the data one by one on the basis of the nearest distance measure of all the pairwise distance between the data point.

Density-Based Clustering Algorithm: It starts from an arbitrary point and retrieves all points density-reachable from that arbitrary point.

Here various types of clustering algorithms along with its advantages and disadvantages are tabulated in Table 2 [6] [7]. The proposed work is extended using K-means algorithm along with Hadoop framework.

With the advantages as described in the Table 2. The proposed work is extended using K-means algorithm along with Hadoop framework.

Comparison of Framework for Big Data Analytics: The Table3 describes various Frame work and its advantages for Big data analytics [8]

Disco: Disco is a project of Nokia Research developed in 2008 by Ville Tuulos. Disco Core is aopen-source framework for distributed computing. This is used to analyze and process large data sets.

Cloud MapReduce: It is developed at Accenture Technology Labs for Amazon Web Services and it is implemented on the top of the Amazon cloud operating system using Amazon cloud services.

Spark: It is open-source framework developed by AMPLab and donated to Apache Software foundation.

Bash Reduce: Mark Clarke presented a solution for the business on Software Freedom day 2014.

Table 2: Advantages of clustering Algorithms

Clustering Algorithms	Advantages	Disadvantages
K-means	• Each cluster is represented by the center of the cluster. • Data are automatically assigned to cluster. • Simple to implement.	• Unable to handle noisy data and outliers.
K-medoids	• Each cluster is represented by one of the objects in the cluster. • Result and run time depends on the initial partition	• Relatively more costly
Agglomerative Hierarchical Clustering	•Hierarchical clustering by starting with each point as a singleton cluster and then repeatedly merging the two closest clusters until a single, all-encompassing cluster remains. •Use of different distance metrics for measuring distance between clusters will produces different results.	• Relocation of objects is not possible.
Density-based clustering algorithm	• Produces a partitional clustering. • High Sensitivity to the setting up of input parameters.	• Not work in high dimensions of data.

Table 3: Advantages of various platforms for Big data analytics

Framework	Programming Model	Computation Model	Data Processing Model
Disco	Erlang and Python	Distributed	Iterative
Cloud MapReduce	Google`s Mapreduce	Amazon cloud	Amazon Cloud
Bash Reduce	Shell Scripts	Distributed	Iterative
Spark	Java, Python or Scala	Distributed	Iterative
Hadoop	Java	Distributed	Batch

CONCLUSION

The research paper has been reviewed various platforms, frameworks and different clustering algorithms for Big Data Analytics. The advantages and disadvantages of different clustering algorithms are compared. This paper also presented the detailed architecture of Hadoop framework and functions of various components in the Hadoop Eco System. This paper analyzed the various frameworks for the Big Data Analytics. Finally this work will be extended by using Hadoop Framework and k-means algorithm for MNSP.

REFERENCES

1. Ericsson White paper Uen 288 23-3211 Rev B | October 2015.
2. Dilpreet Singh and Chandan K. Reddy, 2014. “A survey on platforms for big data analytics” Journal of Big Data, 1:1, 8.
3. Tom White, 2015. “Hadoop The Definitive Guide”, O`Reilly Publications, 4th Edition April 2015.
4. Greg Linden, Brent Smith and Jeremy York, 2003. “Amazon.com Recommendations Item-to-Item Collaborative Filtering”, IEEE Internet Computing.
5. Adhil Fahad, Najlaa Alshatri Zahir Tari, Abdullah Alamri, Ibrahim Khalil, Albert Y. Zomaya, Sebti Foufou and Abdelaziz Bouras, 2014. “A Survey of Clustering Algorithms for Big Data: Taxonomy and Empirical Analysis” in IEEE Transactions on Emerging Topics in Computing.
6. Huan Li, Kejie Lu and Shicong Meng, 2015. “BigProvision: A Provisioning Framework for Big Data Analytics” IEEE Network.
7. Pang-Ning Tan, Michael Steinbach and Vipin Kumar, 2005. “Introduction to Data Mining ” Addison-Wesley.
8. Huan Liu, Dan Orban, “Cloud MapReduce: a MapReduce Implementation on top of a Cloud Operating System”, Accenture Technology Labs.