Big Data Analytics using Hadoop Collaborative Approach on Android

Altaf Shah*, Amol Bhagat** and Sadique Ali***
*Department of Computer Science and Engineering,
Prof Ram Meghe College of Engineering and Management, Badnera, Amravati, India
altafshah7@gmail.com
**Innovation and Entrepreneurship Development Center,
Prof Ram Meghe College of Engineering and Management, Badnera, Amravati, India
amol.bhagat84@gmail.com
***Department of Computer Science and Engineering,
Prof Ram Meghe College of Engineering and Management, Badnera, Amravati, India
softalis@gmail.com

Abstract: Big data also known as data sets which are so large and complex that they are not easy to understand or handle using traditional versions of data processing or database management systems. The Data is always important for searching, transferring, visualizing, storing this data the Data Storage technologies are responsible. Now as the business grows the data pertaining to the business organization grows drastically - hence storage technologies are emerging as critical IT system on which the business depends. At the same time, businesses across the board are also increasing their investment in various mobility solutions. Areas those are getting popular like 'the integration of videos, images, personal information and streaming, displaying, them in cross platform mobile devices – and streaming it in cross platform mobile devices. This paper presents how to mine data on a mobile from a big-data on the remote server.

Keywords: Android, Big Data, Data Analytics, Hadoop, Mapreduce, Mobile Data Analytics.

Introduction

Recent development of various areas of Information and Communication Technology (ICT) has contributed to an explosive growth in the volume of data. According to a report published by IBM in 2012 [1], 90 present of the data in the world was generated in the previous two years. As a consequence, the concept of the big data has emerged as a widely recognized trend, which is currently attracting much attention from government, industry, and academia. Figure 1 show the various by which big data can be gathered. Big data is data that exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or doesn’t fit the strictures of your database architectures. To gain value from this data, you must choose an alternative way to process it.

The hot IT buzzword of 2012, big data has become viable as cost-effective approaches have emerged to tame the volume, velocity and variability of massive data. Within this data lie valuable patterns and information, previously hidden because of the amount of work required to extract them. To leading corporations, such as Wal-Mart or Google, this power has been in reach for some time, but at fantastic cost. Today’s commodity hardware, cloud architectures and open source software bring big data processing into the reach of the less well-resourced. Big data processing is eminently feasible for even the small garage start-ups, who can cheaply rent server time in the cloud.

The value of big data to an organization falls into two categories: analytical use, and enabling new products. Big data analytics can reveal insights hidden previously by data too costly to process, such as peer influence among customers, revealed by analysing shoppers’ transactions, social and geographical data. Being able to process every item of data in reasonable time removes the troublesome need for sampling and promotes an investigative approach to data, in contrast to the somewhat static nature of running predetermined reports.

The past decade’s successful web start-ups are prime examples of big data used as an enabler of new products and services. For example, by combining a large number of signals from a user’s actions and those of their friends, Facebook has been able to craft a highly personalized user experience and create a new kind of advertising business. It’s no coincidence that the lion’s share of ideas and tools underpinning big data has emerged from Google, Yahoo, Amazon and Facebook. The emergence of big data into the enterprise brings with it a necessary counterpart: agility. Successfully exploiting the value in big data requires experimentation and exploration. Whether creating new products or looking for ways to gain competitive advantage, the job calls for curiosity and an entrepreneurial outlook.

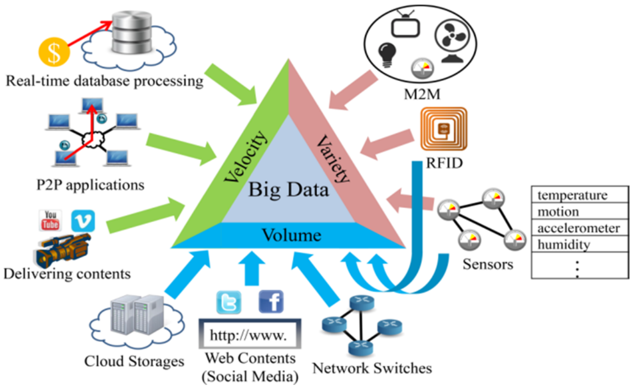

Figure 1. Major trends of big data gathering

The primary goal of big data analytics is to help companies make more informed business decisions by enabling data scientists, predictive modellers and other analytics professionals to analyse large volumes of transaction data, as well as other forms of data that may be untapped by conventional business intelligence (BI) programs. That could include Web server logs and Internet clickstream data, social media content and social network activity reports, text from customer emails and survey responses, mobile-phone call detail records and machine data captured by sensors connected to the Internet of Things. Some people exclusively associate big data with semi-structured and unstructured data of that sort, but consulting firms like Gartner Inc. and Forrester Research Inc. also consider transactions and other structured data to be valid components of big data analytics applications.

Big data can be analysed with the software tools commonly used as part of advanced analytics disciplines such as predictive analytics, data mining, text analytics and statistical analysis. Mainstream BI software and data visualization tools can also play a role in the analysis process. But the semi-structured and unstructured data may not fit well in traditional data warehouses based on relational databases. Furthermore, data warehouses may not be able to handle the processing demands posed by sets of big data that need to be updated frequently or even continually -- for example, real-time data on the performance of mobile applications or of oil and gas pipelines. As a result, many organizations looking to collect, process and analyse big data have turned to a newer class of technologies that includes Hadoop and related tools such as YARN, MapReduce, Spark, Hive and Pig as well as NoSQL databases. Those technologies form the core of an open source software framework that supports the processing of large and diverse data sets across clustered systems [17].

For years SAS customers have evolved their analytics methods from a reactive view into a proactive approach using predictive and prescriptive analytics. Both reactive and proactive approaches are used by organizations, but let's look closely at what is best for your organization and task at hand. Enterprises are increasingly looking to find actionable insights into their data. Many big data projects originate from the need to answer specific business questions. With the right big data analytics platforms in place, an enterprise can boost sales, increase efficiency, and improve operations, customer service and risk management.

Webopedia parent company, QuinStreet, surveyed 540 enterprise decision-makers involved in big data purchases to learn which business areas companies plan to use Big Data analytics to improve operations. About half of all respondents said they were applying big data analytics to improve customer retention, help with product development and gain a competitive advantage. Notably, the business area getting the most attention relates to increasing efficiencies and optimizing operations. Specifically, 62 % of respondents said that they use big data analytics to improve speed and reduce complexity.

Related Work

The different aspects of hadoop distributed file system are described in [1]. It presents the working of hadoop components. It describes comparison of hadoop technique with other system technique and concludes that Hadoop is possibly one of the best solutions to maintain the Big Data. Paper provide study of other techniques such as Grid Computing tools, Volunteering Computing and RDBMS techniques, paper presents that Hadoop is capable enough to handle such amount of data and analyse such data. Big data analytics define the analysis of large amount of data to get the useful information and uncover the hidden patterns [2]. Big data analytics refers to the Mapreduce Framework which is developed by the Google. Apache Hadoop is the open source platform which is used for the purpose of implementation of Google's Mapreduce Model. In this the performance of SF-CFS is compared with the HDFS using the SWIM by the Facebook job traces. SWIM contains the workloads of thousands of jobs with complex data arrival and computation patterns.

To explore Big Data, [3] have analysed several challenges at the data, model, and system levels. To support Big Data mining, high-performance computing platforms are required, which impose systematic designs to unleash the full power of the Big Data. At the data level, the autonomous information sources and the variety of the data collection environments, often result in data with complicated conditions, such as missing/uncertain values. In other situations, privacy concerns, noise, and errors can be introduced into the data, to produce altered data copies. Developing a safe and sound information sharing protocol is a major challenge. At the model level, the key challenge is to generate global models by combining locally discovered patterns to form a unifying view. The report [4] is an attempt to assess the potential value of data equity to the companies that gather and store big data, and who combine this with high-performance analytics to harness the big data's full potential. they undertook this at the sector level by reviewing and developing an understanding from the available literature of how and by what means big data could be expected to impact on each of twelve sectors: Retail Banking, Insurance, Investment Banking, Retail, Central Government, Healthcare, Transport & Logistics, Telecommunications, Energy & Utilities, Manufacturing, Professional Services and Other Activities.

In [5], authors provide an overview of state-of-the-art research issues and achievements in the field of analytics over big data, and they extend the discussion to analytics over big multidimensional data as well, by highlighting open problems and actual research trends. Their analytical contribution is finally completed by several novel research directions arising in this field, which plays a leading role in next-generation Data Warehousing and OLAP research. [6] have characterized four requirements for the data placement structure: (1) fast data loading, (2) fast query processing, (3) highly efficient storage space utilization, and (4) strong adaptively to highly dynamic workload patterns. they have examined three commonly accepted data placement structures in conventional databases, namely row- stores, column-stores, and hybrid-stores in the context of large data analysis using MapReduce. In this paper, they present a big data placement structure called RCFile (Record Columnar File) and its implementation in the Hadoop system.

Starfish, a self-tuning system for big data analytics is introduced in [7]. Starfish builds on Hadoop while adapting to user needs and system workloads to provide good performance automatically, without any need for users to understand and manipulate the many tuning knobs in Hadoop. While Starfish's system architecture is guided by work on self-tuning database systems, they discussed how new analysis practices over big data pose new challenges; leading us to different design choices in Starfish. The approach in the paper enables Starfish to handle the significant interactions arising among choices made at different levels. In [8] authors found Facebook daily operation results, certain types of queries are executed at an unacceptable low speed by Hive (a production SQL-to-MapReduce translator). In this paper, they demonstrate that existing SQL-to-MapReduce translators that operate in a one-operation-to-one-job mode and do not consider query correlations cannot generate high-performance MapReduce programs for certain queries, due to the mismatch between complex SQL structures and simple MapReduce framework. They propose and develop a system called YSmart, a correlation aware SQL-to-MapReduce translator. YSmart applies a set of rules to use the minimal number of MapReduce jobs to execute multiple correlated operations in a complex query.

The article [9] discussed big data techniques and technologies, the transformative potential of big data in five domains. They focussed on data have swept into every industry and business function and are now an important factor of production, big data creates value in several ways. The book [10] is the culmination of five years' worth of in-memory research presenting: overview of their vision of how in-memory technology will change enterprise applications, technical foundations of in-memory data management, in-depth description of how we intend to realize our vision, resulting implications on the development and capabilities of enterprise applications. In [11] authors shown from an application perspective, many websites dedicated to social media are among the most popular Wikipedia (collective knowledge generation), MySpace and Facebook (social networking), YouTube (social networking and multi- media content sharing), Digg and Delicious (social browsing, news ranking, and bookmarking), Second Life (virtual reality), and Twitter (social networking and microblogging), becoming the source of Big-Data and how users, customers, volunteers getting interacted with these sites to make businesses. In [12], author report their survey on a selection of state-of-the- art VA systems as a basis for analysing current market and trend, discussing space for improvement and identifying future research directions. They evaluate the functionality and performance of each system by surveying the vendor with a structured questionnaire as well as testing with real world data and detailed findings and outline the main characteristic of each system. Their survey provides a comparative review of ten products on the market. We also investigate a larger number of systems, including Cognos, SQL Server BI, Business Objects, Teradata, PowerPivot, Panopticon, KNIME, Oculus, Palentir and in-Spire to gain a better overview of the VA software market.

The paper [13] is about how the SP theory of intelligence and its realization in the SP machine may, with advantage, be applied to the management and analysis of big data. The SP system introduced in this paper and fully described, may help to overcome the problem of variety in big data; it has potential as a universal framework for the representation and processing of diverse kinds of knowledge, helping to reduce the diversity of formalisms and formats for knowledge, and the different ways in which they are processed. It has strengths in the unsupervised learning or discovery of structure in data, in pattern recognition, in the parsing and production of natural language, in several kinds of reasoning, and more. In [14], authors have investigated the privacy challenges in the big data era by first identifying big data privacy requirements and then discussing whether existing privacy-preserving techniques are sufficient for big data processing. They have also introduced an efficient and privacy-preserving cosine similarity computing protocol in response to the efficiency and privacy requirements of data mining in the big data era. Although they have analysed the privacy and efficiency challenges in general big data analytics to shed light on the privacy research in big data, significant research efforts should be further put into addressing unique privacy issues in some specific big data analytics.

In [15] described Smart Data that is realized by extracting value from Big Data, to benefit not just large companies but each individual. He considered If his child is an asthma patient, for all the data relevant to his child with the four V-challenges, what he care about is simply, "How is her current health, and what are the risk of having an asthma attack in her current situation (now and today), especially if that risk has changed?" As he shown, Smart Data that gives such personalized and actionable information will need to utilize metadata, use domain specific knowledge, employ semantics and intelligent processing, and go beyond traditional reliance on ML and NLP.

In [16], authors proposed architecture for real-time Big Data analysis for remote sensing application. The proposed architecture efficiently processed and analysed real-time and offline remote sensing Big Data for decision-making. The proposed architecture is composed of three major units, such as 1) RSDU; 2) DPU; and 3) DADU. These units implement algorithms for each level of the architecture depending on the required analysis. The architecture of real-time Big is generic (application independent) that is used for any type of remote sensing Big Data analysis. Furthermore, the capabilities of filtering, dividing, and parallel processing of only useful information are performed by discarding all other extra data. The proposed architecture welcomes researchers and organizations for any type of remote sensory Big Data analysis by developing algorithms for each level of the architecture depending on their analysis requirement.

Big Data Hadoop Android a Collaborative Approach

In this stage an in-depth analysis is performed to obtain a detailed understanding of the business needs as defined in the business case and scope documents. The challenges, features, requirements of big data platform and how important Big-Data analytics are? analysed in this paper. For the ease of use and for ease of business, analytics should be available of smart phone (Android). For solving this issue we have proposed logic. Contemporary mobile users show an increasing trend in consuming information as data analytics, with bigger screens and smarter visualization on mobile devices. Instead of showing the information on computers now it can be available on mobile phone. Applications running on a mobile device (Android, iOS) want to access file systems stored in the Hadoop server. For that, applications use the HTTP requests.

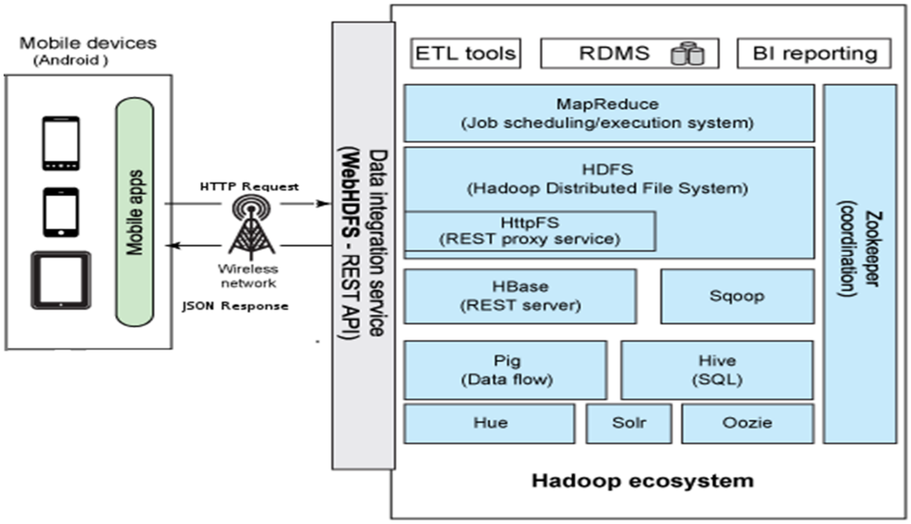

Fig 2. Big data analytics on Android platform

The procedure then fires a REST (JSON/HTTP) request to the web service running inside the Hadoop container system as MapReduce service. Upon receiving the REST request processing starts, server then returns data back in JSON format. Once the application retrieves data in JSON format then Application running on mobile device (Android, iOS) parse it and represent it in desired format. The approach utilized for analysing Big data on android platform is depicted in figure 2. In this paper an approach for showing Big-Data analytics on mobile devices running android OS is developed. The main objectives of this paper is addressing challenges, requirements, and importance of Big-data, application of hadoop for data analytics, utilization of JSON and web-services for Big-data analytics on Android mobile phone. The proposed work is implemented by using following steps:

Step 1: Copy input file- copy input file from local file system to Hadoop Distributed File System (HDFS) using the command *$ hdfs dfs -put /home/altaf/Desktop/MyHadoop/ finalin*

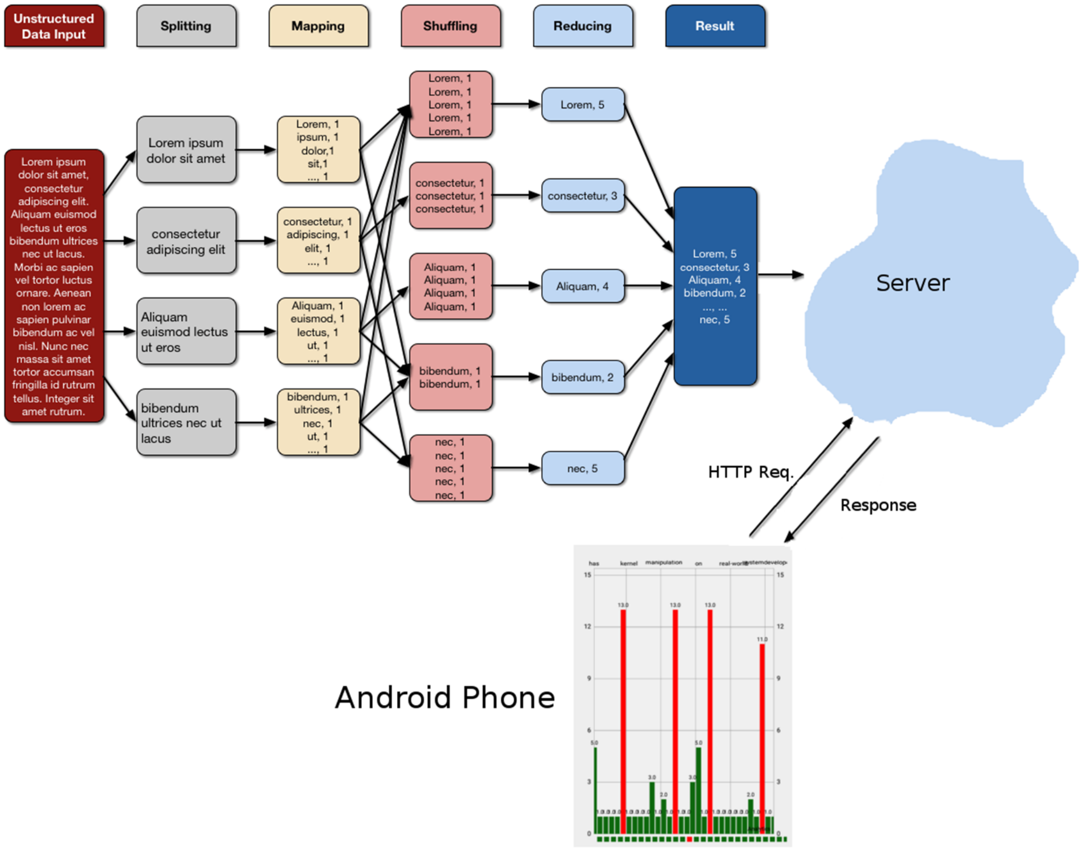
Unstructured Data Input
Splitting
Mapping
Shuffling
Reducing
Result
Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aliquam euismod lectus ut eros bibendum ultrices nec ut lacus. Morbi ac sapien vel tortor luctus ornare. Aenean non lorem ac sapien pulvinar bibendum ac vel nisl. Nunc nec massa sit amet tortor accumsan fringilla id rutrum tellus. Integer sit amet rutrum.
Lorem ipsum dolor sit amet
consectetur adipiscing elit
Aliquam euismod lectus ut eros
bibendum ultrices nec ut lacus
Lorem, 1 ipsum, 1 dolor,1 sit,1 ...., 1
consectetur, 1 adipiscing, 1 elit, 1 ...., 1
Aliquam, 1 euismod, 1 lectus, 1 ut, 1 ...., 1
bibendum, 1 ultrices, 1 nec, 1 ut, 1 ...., 1
Lorem, 1 Lorem, 1 Lorem, 1 Lorem, 1 Lorem, 1
consectetur, 1 consectetur, 1 consectetur, 1
Aliquam, 1 Aliquam, 1 Aliquam, 1 Aliquam, 1
bibendum, 1 bibendum, 1
nec, 1 nec, 1 nec, 1 nec, 1 nec, 1
Lorem, 5
consectetur, 3
Aliquam, 4
bibendum, 2
nec, 5
Lorem, 5 consectetur, 3 Aliquam, 4 bibendum, 2 ...., ... nec, 5
Server
HTTP Req.
Response
Android Phone
has
kernel
manipulation
on
real-world
15
12
9
6
3
0
13.0
13.0
13.0
11.0
5.0
5.0
3.0
3.0
2.0
2.0

Fig 3. Proposed Big-Data Hadoop Android collaborative approach

Step 2: Execute the job- execute the job and obtain desired output which is going to be store in directory of HDFS. This is done by command *$ hadoop jar /usr/local/hadoop/share/hadoop/mapreduce/hadoop-mapreduce-examples-2.6.0.jar wordcount finalin/final final*

Step 3: Display output file- The output file can be displayed using command *$ hdfs dfs -cat finalop/**

Step 4: Copy output to local file system- After getting output in output file copy it from HDFS to Local file System using (Local Server) command $ hdfs dfs -copyToLocal finalop/* /var/www/html/xyz/

Step 5: Android User- Android makes the HTTP connection to local server and reads the copied file. The analytics are shown in the form of bar chart. If output file contain word whose count is greater than threshold value (11) then represent that word with red bar.

The overall implementation is carried out using Android framework, XML, PHP, and Hadoop. In the experimental controlled environment we have created a Mobile (Android) and Hadoop Application. The developed system contains two maor components: Hadoop Application and Android Application (for End User). Hadoop application which operates on data or dataset and perform number of operations like mapping, reducing, counting words, creating output file and so on. Android application shows the number of time word occurrence of word in file of text or in dataset. The complete process followed for the implementation of the proposed collaborative approach is shown in the figure 3.

Experimental Results

The word count operation takes place in two stages a mapper phase and a reducer phase. In mapper phase first the test is tokenized into words then we form a key value pair with these words where the key being the word itself and value ‘1’. In the reduce phase the keys are grouped together and the values for similar keys are added. So here, there is only one pair of similar keys ‘tring’ the values for these keys would be added.

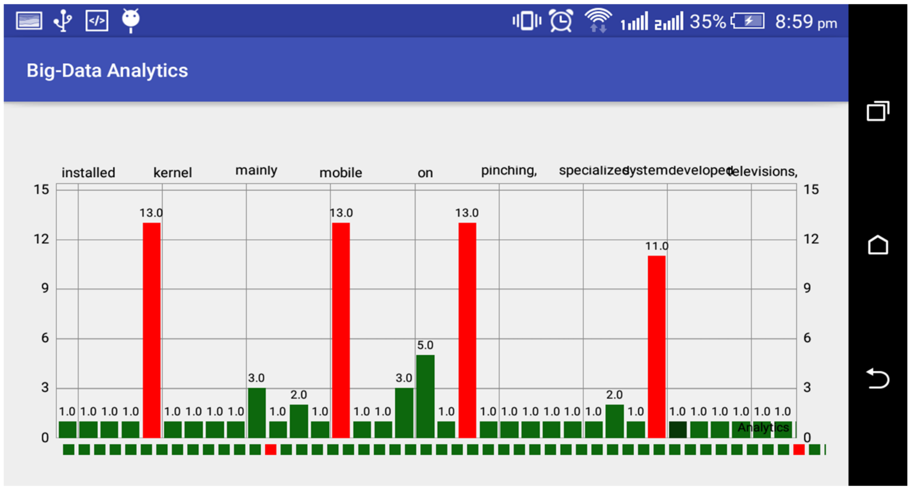

Fig 4. Big-data analytics obtained on Android 5.1.1

The point to be noted here is that first the mapper class executes completely on the entire data set splitting the words and forming the initial key value pairs. Only after this entire process is completed the reducer starts. Say if we have a total of 10 lines in our input files combined together, first the 10 lines are tokenized and key value pairs are formed in parallel, only after this the aggregation/ reducer would start its operation. Now coming to the practical side of implementation we need our input file and map reduce program jar to do the process job. In a common map reduce process two methods do the key job namely the map and reduce, the main method would trigger the map and reduce methods. After getting the number of occurrences of a word then output is uploaded to the server. The end user make request and get the analytics in the form of bar chart. Figure 4 shows the analytics obtained on the Android 5.1.1 in landscape view.

The analytics are easily available on users smartphone, because Android phones are very common. Enables access anywhere with a Mobile connection i.e. globalised the work. Faster, better decision making by observing chart. It enables access anywhere with a web connection i.e. globalised the work. Device and location independence: Enables users to access systems using a Mobile Phone regarding of their location. No need to buy updates or newer versions of software and hardware. Updating and managing software or applications i.e. cost can be reduce by spending on technology. No need to worry about your lot of data and files to store, this provides more data to save the files in server. Here depending upon the data and usage you can choose the plans. Everything is online, store your entire data in cloud and can access at any time in browser.

Conclusion and Future Scope

The proposed work strives to load the Big-Data analytics on Mobile phone (Android) which is done by using Hadoop and analytics are easily available to user with high throughput. In experimental platform we have implemented Mobile application (Android) and hadoop application and analytics are shown on a Android Application. Mobile has become one of the most active topics in today’s world. In future work, we can use the Mobile as commodity hardware by enhancing the configuration of mobile like, RAM, secondary storage, etc. for Hadoop configuration and also we can extend the applications for machine learning.

References

[1] Rathod, H.; Chauhan, T.: A Survey on Big Data Analysis Techniques. 9. (2013).

[2] Mukherjee A., Datta J., Jorapur R., Singhvi R., Haloi S., Akram W.: Shared disk big data analytics with Apache Hadoop. Pp. 18-22, (2012).

[3] Xindong W., Xingquan Z., Gong-Qing W., Wei D.: Data Mining with Big Data. (2014).

[4] Cebr: Data equity Unlocking the value of big data. SAS Reports, pp. 1–44, (2012).

[5] Cuzzocrea A., Song I., Davis K.C.: Analytics over Large-Scale Multidimensional Data: The Big Data Revolution! in Proceedings of the ACM International Workshop on Data Warehousing and OLAP, pp. 101–104 (2011).

[6] He Y., Lee R., Huai Y., Shao Z., Jain N., Zhang X., Xu Z.: RCFile: A Fast and Space-efficient Data Placement Structure in MapReduce-based Warehouse Systems. IEEE International Conference on Data Engineering (ICDE), pp. 1199–1208, (2011)

[7] Herodotou H., Lim H., Luo G., Borisov N., Dong L., Cetin F.B., Babu, S.: Starfish: A Self-tuning System for Big Data Analytics. In Proceedings of the Conference on Innovative Data Systems Research, pp. 261–272 (2011).

[8] Lee R., Luo T., Huai Y., Wang F., He Y., Zhang X.: Ysmart: Yet Another SQL-to-MapReduce Translator. In IEEE International Conference on Distributed Computing Systems (ICDCS), pp. 25–36 (2011).

[9] Manyika J., Chui M., Brown B., Bughin J., Dobbs R., Roxburgh C., Byers A. H.: Big Data: The Next Frontier for Innovation, Competition, and Productivity. In: McKinsey Global Institute Reports, pp. 1–156 (2011)

[10] Plattner, H., Zeier, A.: In-Memory Data Management: An Inflection Point for Enterprise Applications. Springer, Heidelberg (2011)

[11] Zeng D., Hsinchun C., Lusch R., Li S.H.: Social Media Analytics and Intelligence. IEEE Intelligent Systems 25(6), 13–16 (2010).

[12] Zhang L., Stoffel A., Behrisch M., Mittelstadt S., Schreck T., Pompl R., Weber S., Last H., Keim D.: Visual Analytics for the Big Data Era—A Comparative Review of State-of-the-Art Commercial Systems. in IEEE Conference on Visual Analytics Science and Technology (VAST), pp. 173–182 (2012).

[13] Gerard J.W.: Big Data and the SP Theory of Intelligence. IEEE Translations on Content Mining, (2014)

[14] Rongxing Lu, Hui Zhu, Ximeng Liu, Joseph K. Liu, and Jun Shao: Toward Efficient and Privacy-Preserving Computing in Big Data Era, IEEE Network, (2014)

[15] Amit Sheth, Knoesis, : Smart Data - How you and I will exploit Big Data for personalized digital health and many other activities. 2014 IEEE International Conference on Big Data.

[16] Muhammad Mazhar, Paul Anand, Ahmad Awais, Chen Bo-Wei, Huang Bormin, Ji Wen: Real-Time Big Data Analytical Architecture for Remote Sensing Application. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. (2014).

[17] Badase P. S., Deshbhratar G. P., Bhagat A. P.: Classification and analysis of clustering algorithms for large datasets. in Proceedings International Conference on Innovations in Information, Embedded and Communication Systems. (2015).