BIG DATA CHALLENGES: TOOLS PERSPECTIVE

[1]V.N.V.SRINIVASA RAO, [2]M.S.S.SAI

[1,2]KKR & KSR INSTITUTE OF TECHNOLOGY AND SCIENCES
E-mail: [1]srinivasvuddanti@gmail.com, [2]msssai@emailid.com

Abstract— Depending on modern information systems and digital technologies is flourishing day-by-day, it results in expansion of enormous data by terabytes and petabytes of storage. In order to easy up the things to human, technology is growing, but it leads to generation of huge data, which committed to analyze, process, computation, and finding results. All phases are confide in data only. So to handle these things with technology it becomes challenge and showing a path for Research and Development. These expansion of data is entitled as Big Data analytics. The objective of this paper is to show the challenges of big data and tools suited with it

Index Terms— Big Data Analytics, Structured and Unstructured Data.

I. INTRODUCTION

In digital world, data is generating from various sources and the speed conversion from digital technologies has led to development of big data. It invoke to the collection of large and complex collection of datasets which is difficult to process using classical database tools and processing applications. These are available in the formats like structured, unstructured and semi-structured in petabytes and beyond. Suppose these all the data formats are considered as formerly it can be considered as 5 V's.

A. **First V- refers to volume** specifies huge amount of data that is generating every day from various sources, simply size of the data at present larger than terabytes and petabytes. Advantage of data growing it creates the hidden information through analysis.

B. **Second V- refers to velocity** specifies how fast the data rise and gathered for analysis, simply it shows the data coming from different sources. It is not limited to only input sources but also a data movements. For example data is generating all the time from sensors to database.

C. **Third V- refers to variety** specifies the types of data we deal with text, image, audio, video, images, data logs i.e.
   - Structured- This type of data tagged and storing is in the form rows and columns format. And results generating based on queries on functional and operational needs.
   - Semi-structured- it does not belong to fixed format, data is separated with tags to separate data elements. It is not adapt any stable schema.
   - Unstructured- this type of data is difficult to analyze it involves formats which not on relational schema. Examples are images, audio and video etc.

D. **Fourth V- refers to Value-** it is very important feature of big data, is high finding of hidden facts from huge datasets. It is used for finding knowledge from huge repositories of data. These results highly help in business movements and decision making systems.

E. **Fifth V- refers to veracity**- specifies the quality of gathered data that matches to affecting accurate analysis. The main intention of big data analysis is to proceed data of huge volume, velocity, variety, and veracity using different classical data processing tools and techniques.

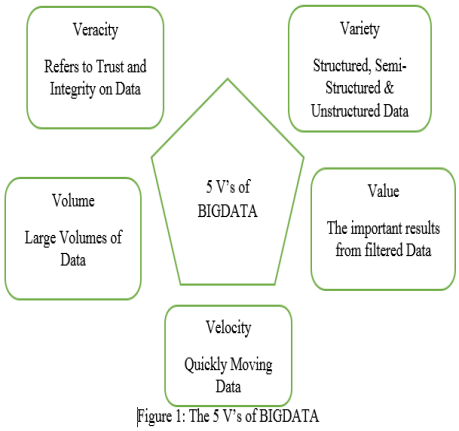

Figure 1: The 5 V's of BIGDATA

The above figure specifies the big data definition, exact definition is not specified but there is a belief that it is problem specific. Purpose of Data warehouses is to manage the large datasets. In this case finding the knowledge in large datasets is foremost challenge. Most of the approaches in data mining are not able to handle large datasets successfully.

All the available data in the form of big data are not used for decision making and analysis process. We

focus on available techniques and challenges of big data.

II. CHALLENGES IN BIG DATA

Big data has been increasing now a day in several domains like public administration, health care, retail, bio-chemistry, scientific researches. Web-based applications facing big data frequently, such as social networking, internet text and documents, and internet searching. Social networking includes social network analysis, online communities, stock market analysis, and search engine indexing like Google, yahoo, bing, amazon. These are some advantages it provides a new challenges in knowledge processing tasks. However opportunities always follow some challenges. To handle that challenges we need to know computational complexities, information security, and method of computation to analyze the big data. Challenges of big data analytics are shown below.

A. Data Storage and Analysis

Now a Days data size is epidemic by various means such as increase of internet usage, mobile devices, aerial sensory technologies, security infrastructures etc. These data are stored on mispent much amount of cost whereas they ignored or deleted finally because there is no enough space to store them. So our first challenge is storage mediums and higher input and output speed. In past decades, analyst use hard drives to store data but, it slower random I/O performance than sequential I/O. To overcome this drawback, the concept of solid state drive (SSD) was introduced. The available storage technologies cannot possess the required performance for processing of big data.

Another challenge with big data analysis is diversity of data, with the growing of datasets. Clustering of huge datasets that help in evaluating the big data is of prime concern. But now tool for big data is hadoop and map Reduce is taking large data sets and transform semi-structured and unstructured to structured data and compute. Major challenge in this case to pay high attention for making storage systems and heighten the data analysis tool that provides guarantees on output.

In the first challenge we need to put more concentration on output evaluation strategy when the input from multiple input sources.

B. Computational complexities and knowledge discoveries

The reason for computational complexities is knowledge discovery and representation is major issue in big data. So data analysis from large repositories from warehouse is also a challenge. It may be difficult to establish mathematical system is applicable to big data. But a data analytics which is based on specific domain can be done easily by understanding particular complexities. This is another challenge i.e developing computational complexity that maps the requirements of this condition.

C. Scalability and Data visualization

From the past decades when data is increasing dramatically than processors speeds, it automatically speeds of the processor also increasing. By increasing the number of cores in a processor. It automatically compete with the data sizes. And also the processing can be done parallel rather than serial processing. Real-time like GPS navigation, social networking, internet search, banking and finance etc.., requires the parallel computing.

And next is Data Visualization is also challenging because visualizing data is more adequately using some techniques in Graph theory. Visualization of data with graph manner provides the link between data with proper interpretation. For example, online market flipkart, amazon, ebay having billions of goods sold for millions of customers each month, this generates lot of data. So understand the company's sales employees need some visualization tools. This is also another challenge factor for big data.

D. Information Security

In Big data analysis security is decisive, because gigantic amount of data is related, analyzed, and mined for meaning decisions. To provide safety to this type of sensitive information companies implement high security mechanisms. And it is also a big challenge in big data. Security of big data can be strengthen by using some techniques like authorization, authenticity, s, and decryption. So attention to be on multi-level security mechanism model and preventing system

III. BIGDATA Processing Tools

High number of tools for processing Big Data is available, here we discuss important tools that deals with Big Data with three important tools namely Apache Spark, Map Reduce and Spark and storm. Available tools for Big Data mainly focuses on batch processing, stream processing and interactive analysis. Batch Processing tools are based on Apache hadoop infrastructure such as mahout and dryad. Stream data applications are mostly used for real-time analytic, examples are strom and splunk. Interactive analysis tools for analyzing data in real-time applications

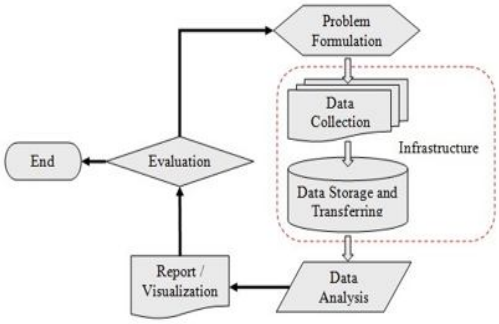

Figure 2: Work Flow of Big Data Project

Apache hadoop MapReduce is a framework for easily writing applications for huge amount of data. It contains hadoop kernel, map reduce, hadoop distributed file system (HDFS) and apache hive etc. Map Reduce is a programming model for processing large dataset is based on divide and conquer method, and map step and reduce step. Hadoop works on two kinds of nodes such as master node and slave node. This master node divides the input problem into sub problems and assigns work to slave systems and this is map step and master node combines all the problems result this is reduce node.

A*pache Mahout* goal is to build an programming environment and framework for building machine learning and scalable algorithms. It is mainly focuses on filtering, clustering, classification. The goal of Apache Mahout is to develop responsive, analytical, highly eminent system. These are some basic tools used for big data processing and analytics. But these may not beat all the challenges of big data. Whereas a new tool which is not related to big data has some of its challenges had met, i.e

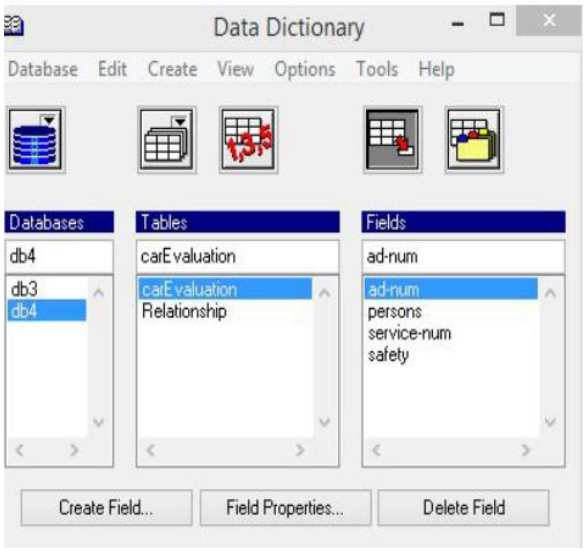

Figure - 3 Multiple Data Sources

OPEN EDGE ABL is a framework which is a combined package with frond end programming language Advanced Business Language which is intended to write business logic for software and backend for Progress Database which is for storage the data purposes and app builder is used to design the front end logic for user interactions purpose. OpenEdge programming has several features which also supports oops concepts also. Our Big data challenges are Data storage and analysis, computational complexities and knowledge discoveries, scalability and data visualization, and information security. Open Edge ABL on applying the sample project with is based on analytical processing related to keyword searching paradigms. And the input sources is not from single. Data input sources are taken from multiple and based on that multiple sources. The data can be placed into multiple databases this part can be take care by Progress database. End user can give input as keyword based on that keyword it analyze and find the relevant results from multiple sources. This part is done by ABL and Appbuilder. And after that computation can be performed and display the results based on computation. Results can be shown in graphical manner

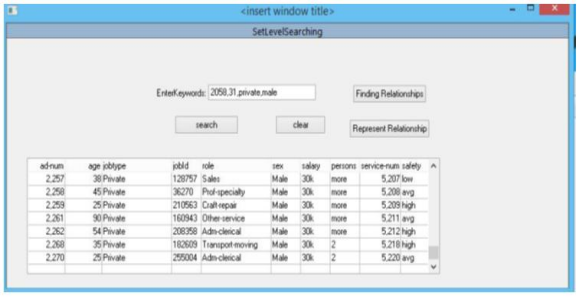

Figure – 4 :Computational Complexity and Analysis

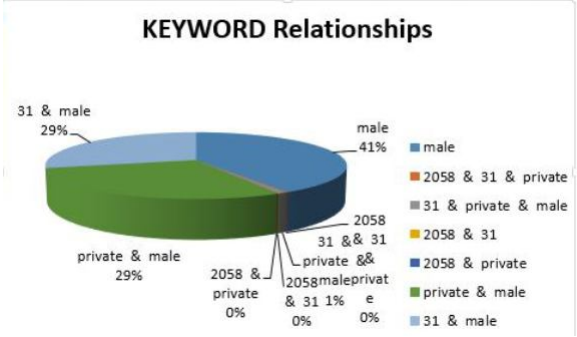

Figure - 5 Data Visualization

CONCLUSIONS

So in this application view there are several phases first one is data sources which is from multiple points, and analysis from multiple sources this is one challenge of big data it is also phased in this application and next is finding the source and maps the relevant results from multiple sources which matches input keyword, this part shows the knowledge discovery and computational complexities. And the next phase is to display the results which is in the form of graph representation. And this belong to third challenge data visualization

and scalability. So partially challenges of the bigdata can beat by openEdge ABL tool but it is not belongs to Bigdata tool. It takes input as normal data and solves the challenges as big data.

REFERENCES

[1] "A Survey on Big Data Analytics: Challenges, Open Research Issues and Tools", D. P. Acharjya, Kauser Ahmed P, IJACSA,Vol. 7, No. 2, 2016

[2] "H2Hadoop: Improving Hadoop Performance using the Metadata of Related Jobs ", *Hamoud Alshammari, Jeongkyu Lee and Hassan Bajwa, DOI 10.1109/TCC.2016.2535261, IEEE Transactions on Cloud Computing*

[3] Tzu-Chi Huang, Kuo-Chih Chu, Xue-Yan Zeng, Jhe-Ru Chen, Ce-Kuen Shieh5, "CURT MapReduce: Caching and Utilizing Results of Tasks for MapReduce on Cloud Computing", 978-1-5090-2179-6/16 $31.00 © 2016 IEEE DOI 10.1109/BigMM.2016.10

[4] Bo Wang, Jinlei Jiang, Yongwei Wu, Guangwen Yang, "Accelerating MapReduce on Commodity Clusters: An SSD-Empowered Approach" DOI 10.1109/TBDATA.2016.2599933, IEEE Transactions on Big Data

[5] Swathi Prabhu, Anisha P Rodrigues, Guru Prasad M S & Nagesh H R, "Performance Enhancement of Hadoop MapReduce Framework for Analyzing BigData", 978-1-4799-608S-9/1S/$31.00©2015 IEEE

[6] Dili Wu*, Aniruddha Gokhale*, " A Self-Tuning System based on Application Profiling and Performance Analysis for Optimizing Hadoop MapReduce Cluster Configuration ",978-1-4799-0730-4/13/$31.00 ©2013 IEEE

[7] Vasiliki Kalavri, Vladimir Vlassov, "MapReduce: Limitations, Optimizations and Open Issues", 978-0-7695-5022-0/13 $26.00 © 2013 IEEE DOI 10.1109/TrustCom.2013.126

[8] Wei Jiang, Gagan Agrawal, "MATE-CG: A MapReduce-Like Framework for Accelerating Data-Intensive Computations on Heterogeneous Clusters", 1530-2075/12 $26.00 © 2012 IEEE DOI 10.1109/IPDPS.2012.65

[9] K. Wang, X. Lin, and W. Tang, "Predator - An Experience Guided Configuration Optimizer for Hadoop MapReduce," in Cloud Computing Technology and Science (CloudCom), 2012 IEEE 4th International Coriference on, 2012, pp. 419-426.

[10] "MapReduce: A Flexible Data Processing Tool," Commun. ACM, vol. 53, no. 1, pp. 72-77, Jan. 2010. [Online]. Available: http://doi.acm.orgIl0.1145/ 1629175.1629198

[11] Azza Abouzeid, Kamil BajdaPawlikowski, "HadoopDB: An Architectural Hybrid of MapReduce and DBMS Technologies for Analytical Workloads", VLDB '09, August 2428,2009, Lyon, France Copyright 2009 VLDB Endowment, ACM 0000000000000/00

[12] Anooshiravan Saboori, Guofei Jiang and Haifeng Chen, "Autotuning Configurations in Distributed Systems for Performance Improvements using Evolutionary Strategies ",1063-6927/08 $25.00 © 2008 IEEE DOI 10.1109/ICDCS.2008.11

★★★