ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 1, January 2017

A Preliminary Study on Big Data Landscape

Manu M N[1], Ramesh B[2]

Assistant Professor, Department of Computer Science & Engineering, SJBIT, Bengaluru, Karnataka, India[1]

Professor, Department of Computer Science & Engineering, MCE, Hassan, Karnataka, India[2]

ABSTRACT:The Big data consists of huge sets of data with various sizes beyond the ability of frequently used software tools to capture, curate, manage, and process data within the required elapsed time. A retailer handles a millions of customer transactions per hour and transfers those into DBs and it contains more than 2.5 petabytes of data. Radio frequency identification systems only can generate 1,000 times the data of conventional advanced bar code systems. Social media like Facebook and Twitter handles more than 250 million photo uploads and the interactions of 800 million active users with more than 900 million pages and twits – each day. More than 6 billion people are, tweeting, sending text messages, calling and browsing on mobile phones worldwide. Over 1 billion people used Facebook in a single day, In Aug 2015. On an average of 31.25 million messages were sent and 2.77 million videos were viewed by Facebook users every minute. A massive growth in video and photo data was seen, alone on YouTube upto 300 hours of videos were uploaded every minute.Understanding and Handling Big data is a big challenge. In this paper, we discuss the challenging issues, organizational issues emerging tools and platforms for Big Data.

KEYWORDS:Big data; Hadoop; HDFS; MapReduce; Hive; Storm; Hive; Mahout.

I. INTRODUCTION

The modern world generates a staggering volume of data and the capacity to store, broadcast and compute this information continues to grow exponentially, with one estimate suggesting that the installed capacity to store information would reach 2.5 zeta-bytes in 2013. International Data Corporation research suggests that the world's digital information is doubling every two years and will increase fifty times by 2020 [27]. The commercial sector is increasingly exploiting these vast quantities of data in a variety of ways, from sophisticated market analysis that allows precisely targeted advertising, to the real-time analysis of financial trends for investment decisions and ultimately to complete new business models.

According to Cisco by 2018, the IoT (Internet of Things) will generate 400 zettabytes (ZB) of data per year [28].Advances in digital sensors, communications, computation, and storage have created hugecollections of data, capturing information of value to business, science, government, andsociety. For example, search engine companies such as Google, Yahoo!, and Microsoft havecreated an entirely new business by capturing the information freely available on the WorldWide Web and providing it to people in useful ways. These companies collect trillions ofbytes of data every day and continually add new services such as satellite images, drivingdirections, and image retrieval. [25]. Just as search engines have transformed how we access information, other forms of big-datacomputingcan and will transform the activities of companies, scientific researchers, medical practitioners, and our nation's defense and intelligence operations.

In 2015, a staggering 1 trillion photos will be taken and billions of them will be shared online. Nearly 80% of photos will be taken on smart phones, by 2017. This year, over 1.4 billion smart phones will be shipped; all of it is packed with sensorscapable of collecting all kinds of data, not to mention the data the users create themselves. The Forecast to grow at a compound annual growth rate 58% surpassing$1 billion by 2020 is the Hadoop market.Retailers who leverage the full power of big data could increase margins as much as 60% their operating [29]. A set of techniques and technologies are required by big data with a new type of integration to reveal insights from datasets that are diverse, complex and are of massive scale. To expose valuable business by applying data mining and analytics information embedded in structures and unstructured, streaming data and data warehouses, The insights can be used to help the

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 1, January 2017

revamp supply chains and to improve program planning and some of the marketing activities, measure the performance across the channel and transforms into an on-demand business. The strategy of big data gives the business the capability to analyses better.

The key to success of any big data analytics initiative is to primarily identify the needs and opportunities of business and then select the proper platform. Some of the identified used cases from the business will be best suited by existing technologies such as a data warehouse while others require a combination of existing technologies and new big data systems. Big data has high volume, high velocity and high variety information assets that require new type of processing to enable enhanced decision making, insight discovery and process optimization, this definition was updated. While understanding the value of big data continues to be a challenge, other practical challenges including return on investment and funding and skills continue to remain at the most important position for many of different industries that are adopting big data. With that said, more than 75% of companies are investing or are planning to invest in big data in the next two years, Gartner Survey for 2015 showed this [30]. This also represented significant increases from a similar survey which was done in the year 2012 which indicated that 58% of companies invested or were planning to invest in big data within the next 2 years.Users perform more than 40,000 search queries every second, On Google alone. But when every second or millisecond can lead to mountains of lost data, each business needs a dedicated platform to analyze and capture data at these increasingly rapid speeds [31].

5 V's of Big Data, they are:

(i) Volume: The quantity of data that is generated is very important in this context. It is the size of the data which determines the value and potential of the data under consideration and whether it can actually be considered as big data or not.
(ii) Variety: It means that the category to which Big Data belongs to is also a very essential fact that needs to be known by the data analysts. This helps the people, who are closely analyzing the data. Managing, integration and governing different varieties of data is still struggle to handle by many organizations.
(iii) Velocity: The term 'velocity' in this framework refers to the accumulating at unpredicted high speed rates or how fast the data is generated. This up-to-date information is important for decision making. Reacting quickly enough to deal with data velocity is a challenge for most organizations.
(iv) Variability: This is a factor which can be a problem for those who are analyzing the data. This refers to the inconsistency which can be shown by the data at times, thus hampering the process of being able to handle and manage the data effectively.
(v) Value: Ability to change data into value by clear understanding of business needs

II. BIG DATA TYPES

Big data types are: Unstructured, semi-structured or structured data. Unstructured data refers to databases of structured information don't contain half of the information which is available for our use. 80% of business-relevant information originates in unstructured form. Not all unstructured data can be simply or easily be converted into a structured model. Some of the limited list of unstructured data types: Emails, PDF files, Spreadsheets, Images, Video, Audio, Social Media data etc..,

Structured data is information usually a text file, which are displayed in titled columns and rows which can be ordered and processed easily by data mining tools. This can be visualizes as a perfectly organized filing cabinet where in everything will be identified labeled and easily accessed. Many organizations are likely to be familiar with this form of data and they are already using it effectively.

Data streaming is also one of the critical issue. Streaming data can be processed as a steady and continuous stream [4]. Streaming technologies are becoming increasingly important with the growth of the Internet because most users do not have fast enough access to download large multimedia files quickly. Analysing data streaming helps in applications like business, astronomy and scientific applications [5] & [6]. In our work, we have considered unstructured type of

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 1, January 2017

data in Big Data.

For analytics we requires different data formats from different sources which may be further integrated using various tools like falcon, sqoop etc., further data can be processed usingHive or HBase etc., which can be further used for analysis for business intelligence.

III. BIG DATA CHALLENGES

The Challenging Issues of Big Data

Capturing the most important data in unstructured format and to deliver that to right people in real-time, Handling unstructured data, storing and management of raw data, Extracting and changing to meet business needs, Finding patterns in mass volume to help organization to make better business decisions, How to mine big data when it is in unstructured format, Mining meaningful of data for analysis etc.,

Organizational Issues

Some of the technical challenges facing in current situation are:

- Handling real-time data sources.
- Cost effectively manages, store and analyze increasingly structured and unstructured data.
- Data flooding into the enterprises to identify patterns and Business insights, intelligent automation is necessary for competitive advantages so smart computing and analytics is required to automate complex infrastructure management.

IV. LITERATURE REVIEW

Numbers of researchers have addressed [27] various issues of Big Data. This section gives the work and major contributions contributed by the researchers under various domains.

JasminAzemovic and Denis Music [8] have presented storing techniques for unstructured data. The three methods are: i) Unstructured data inside relational environment. ii) Unstructured data outside relational environment was setup and iii) Hybrid way of storing unstructured data. The paper presenting the advantages and disadvantages of three methods. The testing environment and drawn (result to shown average upload time to measure overhead and performance using 100KB, 1MB and 10MB files) charts for average upload time to measure overhead and performance using 100KB, 1MB and 10MB files.

The issues of I/O subsystems to provide efficient storage and accessing continuous growing vast repository unstructured data considered in [9]. The paper also discusses other benchmarks available for evaluating different aspects of performance.

Boris et al. [10] concentrates on knowledge management, stored into single library for querying, innovations, re-use of information and for computation advantages. They claim that the metadata layer act as perfect bridge between unstructured and structured data environment.

The paper [11] focuses on data streams, need for analyzing the streams of data and problems involved in mining the data streams. In data-based category, idea is to examining a subset of the whole data set and techniques are sampling, sketching, load shedding, synopsis data structures and aggregation. In task-based solution, the techniques are used to modify existing and invent new techniques for efficient utilization of time and space. The task-based techniques are approximation algorithms, sliding window and output granularity algorithm. They considered clustering, classifications, association, frequency counting and time series analysis algorithms to extract knowledge from streaming inputs.

SHAO Xiao-Dong and LI Qiang [12] addressed two key issues in design and development strategies for huge data transmit/receiving process handling in real-time. They are i) synchronizing problem in transmit/receiving and ii)

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 1, January 2017

Receiving/handling process. Data buffer cache method is used for synchronizing in data sending and Message triggering. Management mechanism also uses data buffer cache for receiving delay problem.

Ostrowski et al., [13] designed a Framework to support machine learning from an unstructured data as sources and to integrate two or more learning mechanism with a traditional indexing method to solve issues in index based algorithm. The framework has three phases, phase i) preprocessing phase to identify all related subjects, Phase ii) implementing supervised learning mechanism using Bayesian learning algorithm and phase iii) integrate the reverse indexing strategy on number of samples.

The paper [14] extracts and classifies unstructured data of web pages and stored into a multimedia database via XML documents. The paper gives detailed study on various tools for data extraction and classification. Document Object Module tree is used in classification process and implemented prototype architecture to show how data extraction from web pages.

J. Chandrika and K.R Ananda Kumar [15] discussed the importance of resources adoption and quality awareness with respect to data stream mining. A novel framework that generalized for any stream mining techniques has been proposed. They proposed resource awareness and quality awareness system.

To reduce the time in finding top-k similar neighbors Lien et al. [16] proposed a new similarity measure based on graph models, which calculated the similarity measures from connections on a graph with vertices being items and users.

Alexander P.Pons and Hassan Alijifri [17] have designed and developed object-related DBMS for storage and manipulation of unstructured data objects. ORDBMS supports all data types and has features of RDBMS provide integration access to all heterogeneous data. The authors have analyzed the facilities of DB2 and ORACLE using object types and manipulation capacity and considered that ORACLE is more effective than DB2.

Deng Shaoling and Li Yan [18] has proposed a customer relationship management model for text data to discover the knowledge from customer unstructured data. To handle massive data they applied content analysis technique and transformed unstructured content into structured data. The decision tree method is adopted to discover customer knowledge.

Information Lifecycle management (ILM) is used in unstructured data management process by the authors Goel et al. [19]**.** They provided well defined approaches to manage unstructured data through adopting agent technology. The agents are: i) Extraction Agent to extract the documents, ii) Categorization Agent to categorize documents and iii) Retrieval Agent to retrieve information from collected documents.

Every organization is trying to get right knowledge to make right decisions at right time. Gu et al., introduced a system MAKES (Multi-faceted and Automatic Knowledge Elicitation System) [20] based on case study of public organization. MAKES architecture allows retrieving, classifying and capturing knowledge from unstructured information. The advantages of this system are quicker decision making, reduced manpower and search time.

For assessing the accuracy, several metrics have been proposed in collaborative filtering methods. They are distributed into two main classes: Decision-support accuracy metrics and statistical accuracy metrics. In this paper [21], statistical accuracy metrics are used to calculate the accuracy of a prediction algorithm by equaling the numerical deviation of the predicted ratings from the respective ratings of user.

V. VISION OF WELL KNOWN TECHNOLOGY FIRMS

The advent of technology and applications have led to Big Data, which is complex due to the volume of information being created every day in terms of terabytes, enterprises speaking in terms of petabytes. There are various applications available and industries are also designing and developing their own tools to handle Big Data problem. Few of them are discussed here.

SAP

SAP's introduced its big data tool in-memory relational database named HANA, HANA is integrates with Hadoop

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 1, January 2017

for processing terabytes of data. It can perform predictive analytics, spatial data processing, advanced analytics, text analytics and streaming analytics.

PENTAHO

Pentaho is open source-based data integration tool for business analytics. It is specifically designed for big data environments. The open source-based suite offers data mining, data integration, OLAP services and ETL capabilities. It also supports NoSQL data sources.

AMAZON SERVICES

Amazon has its own hadoop-based Elastic MapReduce to process huge data sets. It uses Kinesis Firehose for massive streaming data. Amazon offers data mining, ETL, Data Analytics.

MICROSOFT

For analyzing structured and unstructured data Microsoft's offers Hortonworks data platform and HDInsights tool. In recent trends, for big data processing Microsoft's connectors connect SQL Server 2016 to Hadoop, and it is the only big data analytics platform in R.

GOOGLE

Google offers a cloud-based analytics platform and serverlessBigQuery to process huge data and continues to expand offerings.

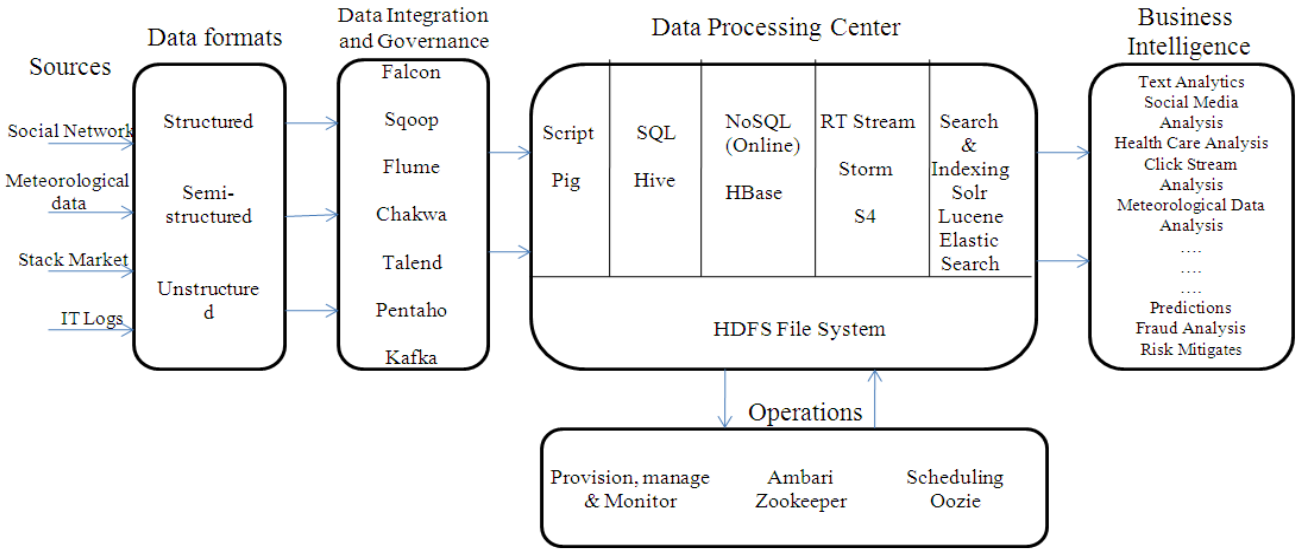

Figure 1. Big data Platform and Tools for Handling Big data

VI. EMERGING TOOLS AND PLATFORMS

Big Data Analysis Tools

1) Hadoop: Apache software is purely open source and framework for very large scale data storing and processing. The de facto standard for processing, analyzing and storing hundreds of Petabytes to Zetabytes of data and beyond. Hadoop enables distributed parallel processing of huge data cost-effectively. Hadoop's get through advantages in businesses and organizations can find value in data. Operating System: Linux, Windows and OS X.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 1, January 2017

2) MapReduce: It is the heart of Apache Hadoop and Developed by Google. It is a programming standard that associates implementation for processing and allows for massive scalability across thousands of servers in parallel and distributed manner in a Hadoop cluster. Operating System: OS autonomous/Independent.

3) Grid Gain: Grid Grain is an open source Java-based tool for big data processing in real time. Grid Gain is compatible with the Hadoop Distributed File System and an option to Hadoop's Map Reduce. It offers in-memory processing for high speed analysis of real time data. Operating Systems: Linux, Windows and OS X.

4) Apache S4: S4 is mainly designed for processing continues data streams and managing data streams, it provides applications for combining streams and processing in real time.

5) Storm: Owned by Twitter. Apache Storm is open source, fault-tolerant and real time distributed computation system. It is language independent. Storm allow constantly process boundless streams of data, it can be used with any programming language. Hadoop works for batch processing.

It can process huge amounts of data in real time with guaranteed reliability. Operating System: Linux.

File System

1) HDFS: Hadoop Distributed File System is a file system showed in Figure 1. It provides scalable and reliable data storage. It is java based distributed file system for Hadoop framework. It is scalable up to 200PB of storage and single cluster serves of 4500 servers. It processing around billion files and blocks. HDFS provide high availability. It provides name node and data node and for scheduling having job tracker and task tracker. Operating system: Linux, Windows and OS X.

2) Gluster: Developed by RedHat. It is unified file system and object storage for huge datasets. Mainly it is used to extend Hadoop capabilities. Operating System: Linux.

Data Integration or aggregation and transfer Tools

1) Apache Falcon: it is data governance and integration tool. It is a framework for data managing and pipeline processing in Hadoop cluster. It enables users to automate the movement and processing of datasets. It handling data replication and automatically manages the complex logic of data handling.

2) Apache Sqoop: it is apache project for transfer bulk data between structured data stores (RDBMS and Data warehouse) and Hadoop. It parallelizes data transfer for makes data analysis more efficient. It imports data from external data stores into HDFS or related systems like Hive, HBase and vice versa. It works with relational databases like Teradata, Netezza, Oracle, MySQL and HSQLDB. Operating System: OS autonomous/Independent.

3) Apache Flume: it is apache project for streaming logs into Hadoop. It is java based, robust, distributed, fault tolerant, reliable, aggregating and transfers large amount of log data into HDFS. It provides web logs data from multiple sources and guarantee delivery of data. Operating system: Linux, Windows and OS X.

Programming Tools

1) Apache Pig/Pig Latin: it is apache project. It is scripting platform for processing and analyzing large datasets. It translates Pig Latin script into MapReduce and it executed within Hadoop. It provide user defined functions and user can write in Java or scripting language and then call directly from the Pig Latin means it allows user to write complex MapReduce transformations such as aggregate, Join and sorting using a simple scripting language. Operating System: OS autonomous/Independent.

2) ECL: Enterprise control language with HPCC. It is a complete tool including an IDE, debugger. Documentation is available on HPCC site. Operating system: Linux.

3) R: It is developed by Bell laboratories. It is a programming language environment for statistical computing and graphics. It is easier to manipulate perform calculations, generate charts and graphs. Operating system: Linux, Windows and OS X.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 1, January 2017

Data Mining Tools

1) Apache Mahout: it is an apache project runs on top of Hadoop and it offers algorithms for classification, clustering and collaborative filtering. It provides rich machine learning libraries. Operating System: OS autonomous/Independent.

2) RapidMiner/RapidAnalytics: RapidMiner is world leading open source system for data and text mining and RapidAnalytics is server version of that product. It is available in paid version, enterprise version and open source version. Operating System: OS autonomous/Independent.

3) MOA: Open source software for stream data mining in real time. It offers implementation of classification, clustering, regression, frequent item set mining and graph mining. SAMOA is new project will combine S4 and storm with MOA for distributed stream mining.

4) Rattle: R analytical tool to learn easily is for non-programmers to use by providing a graphical interface for data mining. Using Rattle can build model, create data summaries, draw graphs and many more. Operating system: Linux, Windows and OS X.

5) Apache spark: It is used to process large datasets.It is faster and then hadoopMapReduce in memory, mainly used for large-scale data processing.

Operations

Below tools are using for administrate, monitor and operate Hadoop cluster.

1) Apache Ambari: Open operational framework for administrate (provisioning and managing) and monitoring apache Hadoop clusters. It includes collection of operator tools and a set of API's.

2) Zookeeper: It offers or provides operational services for Hadoop cluster. It is an open source server that reliably coordinates distributed processes and provides very simple interfaces and services. Key benefits of using Zookeeper are it is fast, reliable, ordered and simple. API's available for Java and c, Python. Operating system: Linux, Windows and OS X.

3) Avro: It is data serialization system. API's are available for java, C, C++ and C#. Operating System: OS autonomous/Independent.

4) Oozie: Apache project is used to coordinate the scheduling of apache Hadoop jobs. It triggers the job based on the availability. Operating system: Linux and OS X.

5) YARN(Data Operating System):it is Apache software foundation's sub-project of Hadoop. It is introduced in Hadoop 2.0 for cluster resource management. It separates the resource management and processing components which was in Hadoop 1.0. It enhances the power of Hadoop computer cluster by improved cluster utilization, scalability, compatibility with Map-Reduce.

Databases/Data warehouse Tools

1) HBase: It support non-relational database store for Hadoop that runs on top of HDFS. It is columnar and provides quick access to large data and fault tolerant storage. It allows users to conduct update, insert and delete operations. Operating System: OS autonomous/Independent.

2) Hive: de facto data warehouse for SQL queries over petabytes of data in Hadoop. It provides powerful set of tools for data summarization, queries and analysis of big data and developers to access Hadoop data. It uses HiveQL as SQL-like language. Operating System: OS autonomous/Independent.

3) Cassandra: Developed by Facebook, presently managed by the Apache Foundation. It is java based Store NoSQL database for huge datasets in "almost" SQL. Operating System: OS Independent.

4) MongoDB: MongoDB is NoSQL database and retains some friendly properties of SQL like Query and index. Operating system: Linux, Windows, OS X, and Solaris.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 1, January 2017

5) Tez: It provides more general data processing applications to the benefit of the entire ecosystem such as apache Hive and apache Pig to meet demands for fast response time and extreme throughput by eliminating unnecessary tasks. After Hadoop YARN, it comes next for Hadoop.

Hive on Spark: New technology, adapting Hive on Spark is used as parallel to MapReduce and Tez. by adopting this, increases the performance, benefits to spark users and greater Hive adoption

VII. CONCLUSION

Big data is an integral part of today's information world. The information world is doubling every two years. New information being created in 2015 is replicated such as shared documents or duplicated in 2016. Using this huge data, delivering trusted information for smarter decision is tedious task. If Facebook, Twitter and LinkedIn are generating collectively mass data per day and tripling every 6 months, so handling big data becoming really a complex and it leads to big challenges. In this paper, we have mentioned some of the challenges in big data, ranging from storage to decision making using various current tools. As per our survey on big data landscape, Apache Hadoop is the most influential and established tool for analyzing big data. Apache Hadoop is a framework for storing and processing data in a large scale, and it is completely open source. Hadoop can run on commodity hardware, making it easy to use with an existing data center, or even to conduct analysis in the cloud. Hadoop is broken into four main parts: The Hadoop Distributed File System (HDFS), which is a distributed file system designed for very high aggregate bandwidth; YARN, a platform for managing Hadoop resources and scheduling programs which will run on the Hadoop infrastructure; MapReduce, as described above, a model for doing big data processing; and a common set of libraries for other modules to use.

REFERENCES

[1] Welcome to Apache Hadoop!. 29 Jul. 2013 http://hadoop.apache.org/.

[2] "Parallel Programming and Computing Platform CUDA, NVIDIA" 2011. 29 Jul. 2013 http://www.nvidia.fr/object/cuda home new.html.

[3] C.F. Cheung, W.B. Lee, W.M. Wang, Y. Wang & W.M. Yeung. "A multi-faceted and automatic knowledge elicitation system (MAKES) for managing unstructured information", Expert Systems with Applications, 2011, vol. 38, pp. 5245 - 5258D.

[4] W.M. Wang & C.F. Cheung, "A narrative-based reasoning with applications in decision support for social service organizations", Expert Systems with Applications, 2011, vol. 38, pp. 3336 – 3345.

[5] J. K., Waters, "Managing unstructured information", Application Development Trends Articles, http://www.adtmag.com/article.aspx?id=10542, Jan. 2005 (Accessed 02.10.10).

[6] C. Moore, "Diving into Data: companies aim to control the rising tide of unstructured data and gain a strategic edge", InfoWorld, http://www.infoworld.com/article/02/10/25/021028feundata_1.html, Oct. 2002 (Accessed 20.3.10)

[7] D. Hatch, "Data Management 2.0: Making Sense of Unstructured Data", Aberdeen Group Benchmark Report, Jul. 2007.

[8] Jasmin Azemovic and Denis Music, "Efficient Model for Detection Data and Data Scheme Tempering with Purpose of Valid 85 Forensic Analysis" Computer Engineering and Applications, International Conference on Computer Engineering and Applications Manila, Philippine, June, 2009.

[9] Lee, Jay; Wu, F.; Zhao, W.; Ghaffari, M.; Liao, L, "Prognostics and health management design for rotary machinery systems—Reviews, methodology and applications". Mechanical Systems and Signal Processing, Jan 2013.

[10] "News: Live Mint". Are Indian companies making enough sense of Big Data?. Live Mint - http://www.livemint.com/. 2014-06-23. Retrieved 2014-11-22.

[11] Ibrahim; Targio Hashem, Abaker; Yaqoob, Ibrar; Badrul Anuar, Nor; Mokhtar, Salimah; Gani, Abdullah; Ullah Khan, Samee (2015). "big data" on cloud computing: Review and open research issues". Information Systems 98–115. doi:10.1016/j.is.2014.07.006.

[12] Shao Xiao-Dong and Li Qiang, "A strategy for continuously big capacity data transmit/receive process handling in real-time", 2nd International Conference on Advanced Computer Control (ICACC), 2010 (Volume:5).

[13] David Alfred Ostrowski, Bradley Graham, Oleg Yu. Gusikhin, "A Discrete Simulation Framework for Part Replenishment Optimization", SIMULTECH, 2013, 467-473.

[14] Couldry, Nick; Turow, Joseph, "Advertising, Big Data, and the Clearance of the Public Realm: Marketers New Approaches to the Content Subsidy". International Journal of Communication: 1710–1726, 2014.

[15] J. Chandrika and K.R Ananda Kumar, "Data Stream Querying: Challenges and issues", International conference on Computer Applications, ISBN: 978-981-08-7300-4, 2011.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 1, January 2017

[16] D. T. Lien, N. D. Phuong, "Collabotative filtering with a graph-based similarity measure," International Conference on Computing, Management and Telecommunications, pp.251-256, 2014

[17] Alexander P.Pons and Hassan Alijifri, "Handling Unstructured Data Type in DB2 and Oracle", Communications of the International Information Management Association, Volume 3 Issue 2.

[18] Deng Shaoling and Li Yan, "Design of an E-Procurement System Based on Business Intelligence Tools", International Conference on Management of e-Commerce and e-Government, ICMECG '08, 2008..

[19] Antony, S., & Santhanam, R. "Could the use of a knowledge-based system lead to implicit learning", Decision Support Systems, 43(1), 141–151, 2007.

[20] J. Gu, W.B. Lee, C.F. Cheung, E. Tsui, W.M. Wang, "Discovery and Capture of Organizational Knowledge from Unstructured Information", World Academy of Science, Engineering and Technology Vol:5 2011-05-29.

[21] GuangHua Cheng, SongJie Gong, "An Efficient Collaborative Filtering Algorithm with Item Hierarchy", Second International Symposium on Intelligent Information Technology Application(IITA2008), IEEE Computer Society Press, 2008, Volume3, pp.28-31.

[22] Fonseca, D. J., Uppal, G., & Greene, T. J. "A knowledge-based system for convey or equipment selection. Expert Systems with Applications", 26, 615–623, 2004.

[23] White Paper, "What to do with Big Data". Avish Technology.

[24] Tseng, Y.-H., Lin, C.-J., & Lin, Y.-I. "Text mining techniques for patent analysis", Information Processing and Management, 43, 1216–1247, 2007.

[25] Randal E. Bryant, Randy H. Katz, Edward D. Lazowska, "Big-Data Computing: Creating revolutionary breakthroughs in commerce, science, and society", Computing Community Consortium.

[26] Wang, W. K, "A knowledge-based decision support system for measuring the performance of government real estate investment", Expert Systems with Applications, 29(4), 901–912. 2005.

[27] Manu M N, AnandaKumar K R, "A Current Trends in Big Data Landscape," 2015 IEEE International Conference on Computational Intelligence and Computing Research.

[28] Dan Worth, "Internet of Things to generate 400 zettabytes of data by 2018," 06 November 2014, Cisco.[Online]. Available:http://www.v3.co.uk/v3-uk/news/2379626/internet-of-things-to-generate-400-zettabytes-of-data-by-2018.

[29] Bernard Marr, "Big Data: 20 Mind-Boggling Facts Everyone Must Read," Sep 30, 2015. [Online]. Available:http://www.forbes.com/sites/bernardmarr/2015/09/30/big-data-20-mind-boggling-facts-everyone-must-read/#65a4730b6c1d.

[30] Maryanne Gaitho, "How Applications of Big Data Drive Industries," October 20, 2015. [Online]. Available:https://www.simplilearn.com/big-data-applications-in-industries-article.

[31] Tx Zhuo, "Big Data' Is No Longer Enough: It's Now All About 'Fast Data" May 13, 2016. [Online]. Available:https://www.entrepreneur.com/article/273561.