Big Data Analytics: An Overview

Jayshree Dwivedi, Abhigyan Tiwary

Abstract: Big data is a data beyond the storage capacity and beyond the processing power is called big data. Big data term is used for data sets it's so large or complex that traditional data, it involves data sets with sizes. Big data size is a constantly moving target year by year ranging from a few dozen terabytes to many petabytes of data means like social networking sites, the amount of data produced by people is growing rapidly every year. Big data is not only a data, rather it become a complete subject, which includes various tools, techniques and framework. It defines the epidemic possibility and evolvement of data both structured and unstructured. Big data is a set of techniques and technologies that require new forms of assimilate to uncover large hidden values from large datasets that are diverse, complex, and of a massive scale. It is difficult to work with using most relational database management systems and desktop statistics and visualization packages exacting preferably massively parallel software running on tens, hundreds, or even thousands of servers. Big data environment is used to grab organize and resolve the various types of data. In this paper we describe applications, problems, and tools of big data and gives overview of big data.

Keywords: Big data, Application, Types, Tools.

I. INTRODUCTION

The big data phenomenon is that, on the other hand it's all about large amounts of data, while the other hand it is also almost running analytics on those large data sets. Big data is a set of techniques and technologies that require new forms of assimilate to uncover large hidden values from large datasets that are disparate, composite, and of massive scale. Big Data represents the Information assets characterized by such a High Volume, Velocity and Variety to require specific Technology and precise methods for its transformation into Value. Big data is not only a data, rather it become a complete subject, which includes various tools, techniques and framework. It defines the epidemic growth and availability of data, both structured and unstructured [1]. Over the past 20 years, data has increased in a large scale in various fields, According to a record from International Data Corporation (IDC) in 2011, the overall created and copied data volume in the world was 1.8ZB (≈ 1021B) which increased by closely nine times within years [1]. This figure will double at least every other years in the near future. Under the explosive increase of global data, the term of big data is mainly used to describe tremendous datasets. Compared with traditional datasets big data typically includes masses of unstructured data that need more real-time analysis. In addition, big data also brings about new opportunities for discovering new values, helps us to gain an in-depth understanding of the hidden values and also provoke new challenges, e.g., how to effectively formulate and manage such datasets. Big data is a collection of data sets so broad and synthesized that it becomes crucial to process using traditional database management tools. The challenges include storage, acquisition, search, sharing, analysis, and determination.

- *Jayshree Dwivedi, Department of Computer Science and Engineering, SIRTS Group of Institute, Bhopal, India jayshree1401@gmail.com*
- *Abhigyan Tiwary, Department of Computer Science and Engineering, SIRTS Group of Institute, Bhopal, India abhigyantiwary@gmail.com*

The trend to larger data sets is due to the additional information obtainable from dissolution of a single large set of related data, as compared to separate smaller sets with the same total amount of data, allowing interconnection to be found to spot business trends, determine quality of research, prevent diseases, link legal illustration, combat crime, and determine real-time roadway traffic conditions. Big data having 3V characteristics i.e. volume (GB, TB, PB, ZB) velocity (speed of change or speed of generation high) variety (structured, unstructured, semi structured) [3].

Volume-The quantity of generated data is important in this context. The size of the data determines the value and potential of the data under concentration, and whether it can actually be considered big data or not. The name big data itself contains a term related to size, and hence the characteristic. The data we find in the format of videos, music's and large images on our social media. It is very common to have Terabytes and Petabytes of the storage system for enterprises. As the database grows the application and architecture built to support the data needs to be re-evaluated quite often.

Velocity- In this context the speed at which the data is generated and processed to meet the demands and the challenges that lie in the path of growth and development. Data is streaming in at unprecedented speed and must be dealt with in a timely manner. RFID tags, sensors and smart apprise are driving the need to deal with torrents of data in near-real time.

Variety- The type of content is an essential fact that data analysts must know. Data can be stored in multiple formats. Data today comes in all types of formats. Structured, numeric data in traditional databases. Information created from line-of-business applications. Sometimes the data is not even in traditional format, it may me in form of video, sms and pdf. It is the need of the organization to arrange it and make it meaningful. It will be easy to do so if we have data in the same format, however it is not the case most of the time. The real world has data in many different formats and that is the challenge we need to overcome with the big data. Put another way, big data is the realization of greater business intelligence by storing, processing, and analyzing data that was previously ignored due to the limitations of traditional data management technologies. The main problem arises with big data is how to store this huge data and how to process this data. Data come

from many quarters Social media sites, Sensors, Digital photos Business transactions, Location-based data, Airlines data, Hospitality data, Shopping data, CCTV data. There are various types of application of big data such as E-Commerce and Market Intelligence, E-Government, Science & Technology, Smart Health and Wellbeing, Security and Public Safety. Due to largeness in size, decentralized control and different data sources with different types the Big Data becomes much complex and harder to handle. We cannot manage them with the local tools those we use for managing the regular data in real time. For major Big Data-related applications, such as Google, Flicker, Face book, a large number of server farms are deployed all over the world to ensure nonstop services and quick responses for local markets. In this paper we describe what are big data, big data applications and big data tools.

II. BIG DATA TYPES

Data can be grouped into two broad categories in Structured Data, Unstructured Data, and Semi-structured Data. Structured data is a data which having fixed field within a record or file this involves data contained in relational databases and spreadsheets. Structured data defines what fields of data will be stored and how the data will be store Structured data includes a data warehouse already tagged and easily sorted but unstructured data is random and difficult to analyze [3].For example, RDBMS, MYSQL. Unstructured data is a data which doesn't involves row-column database it is opposite of structured data. Unstructured data files shaving text, images, audio, videos, mail messages, presentations, webpage's and other business reports. Semi structured data is a data in which there is no separation between the data and the schema. It can be helpful to view structured data and it provides a flexible format for data exchange between different types of database. In semi structured data similar entities are grouped together and entities in same group may not have same attributes. XML is example of semi structured data. Semi-structured data does not conform to fixed fields but contains tags to separate data elements.

III. APPLICATIONS OF BIG DATA

Big data is really cynical to our life and its emerging as one of the most important technologies in modern world. There are many benefits of big data for example using the information kept in the social network like face book, the marketing agencies are learning about the respond for their campaigns, promotions, and other advertising mediums. Using the information in the social media like groove and product perception of their consumers, product perception of their consumers, products companies and retail organizations are planning their production. Using the data regarding the earlier medical history of patients, hospitals are providing better and quick service.

TABLE 1: *Applications of big data*

Application areas of big data	
E-Commerce and Market Intelligence	• Recommender system • Social media monitoring and analysis • Crowd-sourcing systems • Social and virtual games
E-Government and Politics	• Ubiquitous government services • Equal access and public services • Citizen engagement
Science & Technology	• S&T innovation • Hypothesis testing • Knowledge discovery
Smart Health and Wellbeing	• Human and plant genomics • Healthcare decision support • Patient community analysis
Security and Public Safety	• Crime analysis • Computational criminology Terrorism informatics • Open-source intelligence • Cyber security

Big data provides an infrastructure for translucence in manufacturing industry, which is the ability to unravel uncertainties such as in constant component performance and availability. Predictive manufacturing as an applicable approach toward near-zero downtime and translucence requires vast amount of data and advanced prediction tools for a systematic process of data into useful information. A conceptual framework of prognostic manufacturing begins with data acquisition where different type of sensory data is available to acquire such as acoustics, vibration, pressure, current, and voltage and controller data. Huge amount of sensory data in addition to historical data construct the big data in manufacturing. Big data analytics has helped healthcare improve by providing personalized medicine and prescriptive analytics, clinical risk intervention and predictive analytics, waste and care variability reduction, automated external and internal reporting of patient data, standardized medical terms and patient registries and fragmented point solutions.

E-Commerce and Market Intelligence: The opportunities affiliated with data and analysis in different organizations have helped generate significant interest in BI&A, which is often referred to as the techniques, technologies, systems, practices, methodologies, and applications that analyze fussy business data to help an enterprise better understand its business and market and make timely business decisions. In addition to the necessary data processing and analytical technologies, BI&A includes business-centric practices and methodologies that can be applied to different high-impact applications such as e-commerce, market intelligence, e-government, healthcare, and security. The excitement surrounding BI&A and Big Data has arguably been generated primarily from the web and e-commerce communities. Indicative market transformation has been accomplished by leading e-commerce vendors such Amazon and eBay through their ingenious and highly scalable ecommerce platforms and product recommender systems. Exceeding Internet firms such as Google, Amazon, and Face book continue to lead the development of web analytics, cloud computing, and social media platform.

E-Government and Politics: In E-Governance, government makes best possible use of internet technology to communicate and provide orientation to common peoples and businessman. Today, electricity, water, phone and all kinds of bills can be paid over the internet. All this is what government and civilian is using and doing. All are dependent on internet and when civilian depends on government internet services all that come is E-Governance. Modern infrastructure allows contemplate new large scale problem-governesses which solution was not possible before ,e.g. Banking, Media, Airlines, Telecom, Entertainment news, Sports, Astrology, Movie tickets, Public works monitoring, Electricity Board, Health etc. e-governess typically produces a vast amount of data that need to be supported by a new type of e-Infrastructure capable to store, distribute, process, preserve, and curate these data We refer to these new infrastructures as E-governess Data Infrastructure. The emerging E-GOVERNESS should allow different groups of researchers to work on the same data sets, constructing their own distributed approach and combining environments, safely store and retrieved intermediate results, and later share the discovers results. New data provide, Third party security and access control mechanisms and tools will allow researchers to link their e-governess results with the initial data and intermediate data to allow future re-use re-purpose of big data, e.g. with the enhanced research technique and tools. Use of internet by the government to provide its services at the door step of customers, business and other stakeholder.

Science & Technology: Many fields of science and technology (S&T) are reaping the benefits of high-throughput sensors and instruments, from astrophysics and oceanography, to genomics and environmental research. To facilitate information sharing and data analytics, the National Science Foundation (NSF) recently assigned that every project is required to provide a data management plan. Cyber infrastructure, in particular, has become crucial for supporting such data-sharing initiatives.

Smart Health and Wellbeing: The healthcare industry historically has generated large amounts of data, driven by record keeping, compliance & regulatory requirements, and patient care. While most data is stored in hard copy form, the current trend is toward rapid digitization of these large amounts of data. Driven by mandatory requirements and the potential to improve the quality of healthcare delivery meanwhile reducing the costs, these massive quantities of data (known as 'big data') hold the promise of supporting a wide range of medical and healthcare functions, including among others clinical decision support, disease surveillance, and population health management. By digitizing, combining and effectively using big data, healthcare organizations ranging from single-physician offices and multi-provider groups to large hospital networks and accountable care organizations stand to realize significant benefits. Potential benefits include detecting diseases at earlier stages when they can be treated more easily and effectively; managing specific individual and population health and detecting health care fraud more quickly and efficiently. Numerous questions can be addressed with big data analytics. Certain developments or outcomes may be predicted and/or estimated based on vast amounts of historical data, such as length of stay (LOS); patients who will choose elective surgery; patients who likely will not benefit from surgery; complications; patients at risk for medical complications, patients at risk for sepsis, MRSA, C. difficult, or other hospital-acquired illness; illness/disease progression; patients at risk for advancement in disease states; causal factors of illness/disease progression; and possible co morbid conditions (EMC Consulting). McKinsey estimates that big data analytics can enable more than $300 billion in savings per year in U.S. healthcare, two thirds of that through reductions of approximately 8% in national healthcare expenditures.

Security and Public Safety: Since the tragic events of September 11, 2001, security research has gained much attention, especially given the increasing dependency of business and our global society on digital enablement. Researchers in computational science, information systems, social sciences, engineering, medicine, and many other fields have been called upon to help enhance our ability to fight violence, terrorism, cyber crimes, and other cyber security concerns. Critical mission areas have been identified where information technology can contribute, as suggested in the U.S. Office of Homeland Security's report "National Strategy for Homeland Security," released in 2002, including intelligence and warning, border and transportation security, domestic counter-terrorism, protecting critical infrastructure(including cyberspace), defending against catastrophic terrorism, and emergency preparedness and response Intelligence, security, and public safety agencies are gathering large amounts of data from multiple sources, from criminal records of terrorism incidents, and from cyber security threats to multilingual open-source intelligence. Companies of different sizes are facing the daunting task of defending against Cyber security threats and protecting their intellectual assets and infrastructure. Processing and analyzing security-related data, however, is increasingly difficult. A significant challenge in security IT research is the information stovepipe and overload resulting from diverse data sources, multiple data formats, and large data volumes. Current research on technologies for cyber security, counter-terrorism, and crime fighting applications lacks a consistent framework for addressing these data challenges. Selected BI&A technologies such as criminal association rule mining and clustering, criminal network analysis, spatial-temporal analysis and visualization, multilingual text analytics, sentiment and affect analysis, and cyber attacks analysis and attribution should be considered for security informatics research.

IV. PROBLEM AND SOLUTION

The major challenges affiliate with big data are overwhelm data, storage, searching, sharing transfer, analysis, presentations. There are two types of problems originated in big data fist one is how to store this data and second one is how to process this data. For example, social media sites like Face book generated 300 logs files so the problem is how to store this data. So the traditional method is not suitable for this type of data. In 1990 to process this data Google launch two things Hdfs for storing huge data set and Map Reduce to process this huge data set of Hdfs. Hdfs and map reduce are layers of Hadoop. So here the question is what Hadoop is. Hadoop is designed to answer the question How to process big data with reasonable cost and time. Hadoop is a

distributed file system and data processing engine that is designed to handle intensely high volumes of data in any structure. The focus is on supporting redundancy, distributed architectures, and parallel processing. Hadoop is based on parallel processing. Hadoop is the core platform for structuring big data, and solves the problem of formatting it for sub sequential analytics purposes. Hadoop uses a distributed computing architecture subsist of multiple servers using commodity hardware making it relatively inexpensive to scale and support extremely large data stores. Hadoop was created by Doug Cutting and Mike Cafarella in 2005. Doug Cutting, who was working at Yahoo at the time, named it after his son's toy elephant Hadoop is open source software (framework) overseen by apache software foundation for storing processing huge data sets with use of cluster of commodity hardware.. Hadoop has two main components first one is Hdfs and second is Map reduce.

V. BIG DATA TOOLS

There are two types of problems emanate using big data how to store this data and how to process this data these problems solved by using two tools first one is Hadoop and second one is Spark.

Hadoop is an open source framework that provides solutions for handling big data along with extensive processing and analysis. It was created by Doug Cutting in 2005 when he was working for Yahoo at the time for the Notch search engine project. Hadoop has two major components named HDFS (Hadoop Distributed File System) and the Map Reduce framework. Hadoop Distributed File System is said to be inspired by Google's The Google File System (GFS) and provides a scalable, efficient, and replica based storage of data at various nodes that form a part of a cluster. A distributed file system is a client/server-based application that allows clients to access and process data stored on the server as if it were on their own computer. The main core of the Hadoop framework is functionally different from an RDBMS. Hadoop uses two types of methods for storing and processing big data.

Hdfs (Hadoop Distributed File System), it is specially designed file system for storing and processing huge data sets with the use of cluster of commodity hardware and with fixed streaming access pattern. It is a self – healing, high bandwidth clustereble storage [8]. Hadoop divide data into different machines not code. The Hadoop Distributed File System (HDFS) is the file system component of the Hadoop framework. HDFS is designed and augmented to store data over a large amount of low-cost hardware in a distributed fashion. Services of HDFS are Name node, Secondary name node, Job tracker Data node, Task tracker. Name node and data node are used for storage management. Job tracker and task manager are used for process data. The Name Node records all of the metadata, attributes and locations of files and data blocks in to the Data Nodes. The attribute records all the things like file permissions, file modification and access times, and namespace which are a hierarchy of files and directories. Data node is a type of slave node in the Hadoop, which is used for saving the data and there is task tracker in data node which is used to track on the ongoing job on the data node and the jobs which coming from name node [1]. Hadoop Distributed File System (HDFS) allows user data to be organized in the form of files and directories; it provides a command line interface called FS shell that lets a user interact with the data in HDFS accessible to Hadoop Map Reduce programs, used to save the data and there is task tracker in data node which is used to track on the ongoing job on the data node these jobs are coming from name node. The Data Nodes store the blocks and block replicas of the file system. During startup each connects and performs a handshake with the Name Node. The Data Node checks for the accurate namespace ID, and if not found then the Data Node automatically shuts down. New Data Nodes can join the cluster by simply registering with the Name Node and receiving the namespace ID.

Map Reduce is a software framework in which an application is collapsed into parts these parts are called fragments or blocks can be run on any node in the cluster. Operates on unstructured and structured data, it takes input as <key, value>and output will also in form of <key, value>. Map Reduce is a software framework for processing (large) data sets in a distributed fashion over a several machines. The core idea behind Map Reduce is mapping your data set into a collection of <key, value> pairs, and then reducing overall pairs with the same key. The Map Reduce framework consists of a single master, Job Tracker and one slave Task Tracker per cluster-node. The master is responsible for scheduling the jobs' component tasks on the slaves, observing them and re-executing the failed tasks. The slaves execute the tasks as directed by the master Map Reduce programs are contained in a Java jar file an XML file containing serialized program configuration options Running a Map Reduce job places these files into the HDFS and notifies Task Trackers where to retrieve the relevant program code [9]. Minimally, applications specify the input/output locations and supply map and reduce functions via implementations of appropriate interfaces and abstract-classes. These and other job parameters compose the job configuration. The Hadoop job client then submits the job jar executable file. Configuration to the Job Tracker which then assumes the amenability of disseminate the software configuration to the slaves, scheduling tasks and monitoring them, providing status and diagnostic information to the job-client. In below diagram we describe architecture of Hadoop.

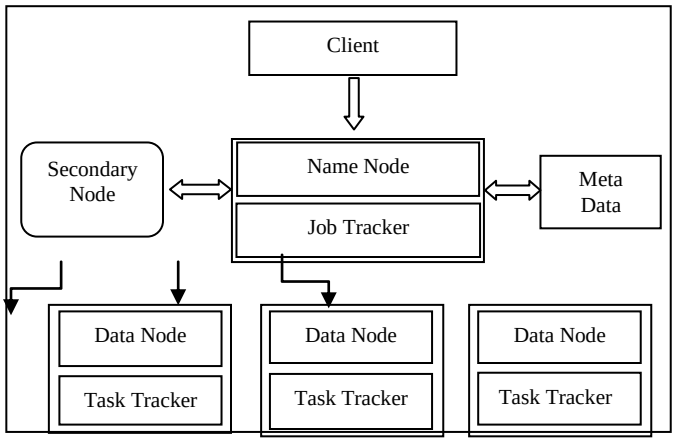

***Fig 1:-** Hadoop architecture*

There are various types of tools used by Hadoop like Pig, Hives, Scoop, Hbase, and Oazie [1].

Pig tool is a high level platform for composing Map Reduce programs used with Hadoop. The language for this platform is called Pig. Pig is a high level scripting language that is used with apache Hadoop. Pig scripts are translated into a sequence of Map Reduce jobs that are run on the apache Hadoop cluster. Pig is a platform for analyzing huge data sets that subsist of a high-level language for expressing data scrutiny programs. Pig generates and compiles a Map Reduce programs on the file. Pig is called as data flow language, Pig is working on bags/ tuple / atom. Working top of Map reduce. If we are not aware of sql queries so use pig tool.

Hive is a data warehouse infrastructure built on top of Hadoop Supports analysis of huge datasets stored in Hadoop compatible file systems like HDFS and Amazon S3 file system, Provides SQL-Like query language called Hive SQL.Provides index to accelerate queries, Hive is a data warehouse infrastructure built on top of Hadoop , Supports analysis of large datasets stored in Hadoop compatible file systems like HDFS and Amazon S3 file.

Sqoop is a tool for transferring data among Hadoop and relational databases. You can use Sqoop to import data from a My SQL or Oracle database into HDFS, run Map Reduce on the data, and then export the data back into an RDBMS. Sqoop automates these processes, using Map Reduce to import and export the data in parallel with fault-tolerance.

Hbase is an open-source, distributed, versioned and column-oriented database built on Hadoop file system on top. It supports Insert, Delete, and Update. It can manage structured and semi structured data. Hbase is a part of the hadoop ecosystem that provides random real time read and write access to data in the Hadoop file system.

Apache Oozie is server based workflow scheduling system to manage Hadoop jobs. In Oozie defined as a collection of control flow and action nodes in a directed acyclic graph. Control flow workflow nodes define the beginning and the end of workflow as well as a mechanism to control the workflow execution path.

Spark is provides a user friendly programming interface to decrease coding efforts and provide better performance in a majority of the cases with problems related to big data. Spark not just provides an alternative to Map Reduce, but also has options for SQL like querying with Shark and a machine learning library called MLib. The performance and working of spark is considerably different from that of map reduce, but is also dependent on the constraints of parallelism, the types of problems in context, and the resources available. Apache Spark started as a research project at UC Berkeley in the AMP Lab, was started with a goal to design a programming model that supports a much wider class of applications than Map Reduce, while maintaining its automatic fault tolerance. Spark offers an abstraction called Resilient distributed Datasets (RDDs) to support these applications efficiently. RDDs can be stored in memory between queries without requiring replication. Instead, they rebuild lost data on failure using lineage: each RDD remembers how it was built from other datasets (by transformations like map, join or group By) to rebuild itself. RDDs allow Spark to outperform existing models by up to 100 as in multi-pass analytics. RDDs can support a wide variety of iterative algorithms, as well as interactive data mining and a highly efficient SQL engine Shark. Spark allows us to perform stream processing with large input data and deal with only a chunk of data on the fly. This can also be used for online machine learning, and is highly appropriate for use cases with a requirement for real time analysis which happens to be an almost ubiquitous requirement in the industry. It also allows us to cache the data in memory, which is beneficial in case of iterative algorithms such as those used in machine learning. Table 2 shows difference between Hadoop and Spark.

Table 2:- *Comparison between Hadoop and Spark*

Hadoop	Spark
Batch processing	Real time processing
Verbose	Compact
High latency	Lower latency
Slower	Faster

VI. CONCLUSION

The newest convolution of big data is generating new opportunities and new challenges for businesses across every industry. The challenge of data integration includes data from social media and other unstructured data into a traditional environment. Apache Hadoop provides a cost-effective and extensive scalable platform for big data and preparing it for analysis. Using Hadoop to deceive the traditional processes can reduce time to analysis by hours or even days. Running the Hadoop cluster readily means selecting an optimal infrastructure of servers, storage, networking, and software. The amount of data has been accumulating and data set analyzing become more combative. The challenge is not only to collect and manage large volume and different type of data, but also to extract meaningful value from it. We have the capabilities to analyze data set fast, quickly and cost effectively in terms of efficiency, capacity, wages and judiciousness.

REFERENCE

[1] Ms. Vibhavari Chavan, Prof. Rajesh. N. Phursule “Survey Paper on big data” ,JSPM’s Imperial College of Engineering and Research, Pune, Vol. 5 (6), 7932-7939, 2014.

[2] Yuri Demchenko, Canh Ngo, Peter Membrey “Architecture Framework and Components for the Big Data Ecosystem Draft Version 0.2”.

[3] Seref SAGIROGLU and Duygu SINANC, “Big Data: A Review” 978-1-4673-6404-1/13/$31.00 © IEEE, 2013.

[4] C.L. Philip Chen, Chun-Yang Zhang, “Data-intensive applications, challenges, techniques and technologies: A survey on Big Data” in Information Sciences 275 314–347, 2014.

[5] Rohit Pitre, Vijay kolekar " A Survey Paper on Data Mining With Big Data" Volume 1 Issue 1 (April 2014).

[6] Yuri Demchenko, Canh Ngo, Peter Membrey, "Architecture Framework and Components for the Big Data Ecosystem" 12 September 2013.

[7] Xue-wen chen1, Xiaotohg lin2, " Big Data Deep Learning: Challenges and Perspectives" Received April 20, 2014, accepted May 13, 2014, date of publication May 16, 2014, date of current version May 28, 2014.

[8] Min Chen, Shiwen Mao, Yunhao Liu, "Big Data: A Survey" Springer Science Business Media New York 2014 Mobile Netw Appl (2014)19:171–209.

[9] Wullianallur Raghupathi, and Viju Raghupathi, Raghupathi, "Big data analytics in healthcare: promise and Potential Health" Information Science and Systems 2014.

[10] Big Data and Analytics 2 for Government Innovation © Springer International Publishing Switzerland 2015 Morabito, Big Data and Analytics, DOI 10.1007/978-3-319-10665-6_2.

[11] Hsinchun Chen, Roger H. L. Chiang, Veda C. Storey, "Business Intelligence and analytics from big data to big impact" Vol. 36 No. 4, pp. 1165-1188/December 2012.

[12] Shubham Kalbande, Sumant Deshpande and Prof. Mohit Popat**,** "Review Paper on Use of Big Data in E-Governance of India" International journal for research in emerging science and technology volume-2, special issue-1, March-2015.