IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 3 Issue 3, March 2016.

www.ijiset.com

ISSN 2348 – 7968

A Review on Big Data Concepts

Dr.K.Venkataramana[1], Dr.M.Sreedevi[2]

1 Assistant Professor,Dept of Computer Science, KMM Institute of Post graduate Studies,Tirupati

2. Assistant Professor,Dept of Computer Science,S.V.University,Tirupati

Abstract

Today we are revolving around Digitalized world which is based on data that we are generating, using in various ways. All the attempts are to how to control and process large quantities of data by using various techniques, but before that we should know what the big data we are taking about is. So in this paper we have reviewed and studied the concepts of big data and the resources of its generation, and how it is analyzed by using various analytics for effective decision. Finally we concluded this paper by giving brief information on big data tools used to process it at different levels which is open source software's which are currently used. By this study we expect the researchers can gain information and issues relating to big data and analytics.

Keywords: *Big data, Analysis, Data analytics, big data tools, data cleansing, data extraction.*

1. Introduction

Continuous data is generated by various data sources and applications at an ever-increasing rate like Mobile phones, social media, imaging technologies to determine a medical diagnosis—all these and more create new data, and that must be stored somewhere for some purpose. Devices and sensors automatically generate diagnostic information that needs to be stored and processed in real time. Due to the variety of disparate data sources and the sheer volume, it is difficult to collect and integrate data with scalability from distributed locations. Merely keeping up with this huge influx of data is difficult, but substantially more challenging to provide function and performance guarantee, in terms of fast retrieval, scalability, and privacy protection. Especially when it does not conform to traditional notions of data structure, to identify meaningful patterns and extract useful information is difficult.

Think of the web which currently covers more than 100 million domains and is still growing at the rate of 20,000 new domains every single day. The data that comes from these domains is so massive and mind boggling that it is practically immeasurable much less manageable by any conventional data management and retrieval methods that are available today. And that is only for our starters. Add to this the 300 million daily Facebook posts, 60 million daily Facebook updates, and 250 million daily tweets coming from more than 900 million combined Facebook and Tweeter users and for sure your imagination is going to go through the roof. Don't forget to include the voluminous data coming from over six billion smart phones currently in use today which continually access the internet to do business online, to post status updates on social media, send out tweets, and many other digital transactions. Remember, approximately one billion of these smart phones are GPS enabled which means they are constantly connected to the internet and therefore, they are continuously leaving behind their digital trails which is adding more data to the already burgeoning bulk of information already stored in millions of servers that span the internet.

Big Data deluge is due to Mobile Sensors, Social media, Smart grids, Video surveillance, Video rendering, Medical imaging, space exploration which will stand out from previous data systems in various attributes shown in figure 1.1 and stand out in defining following characteristics:

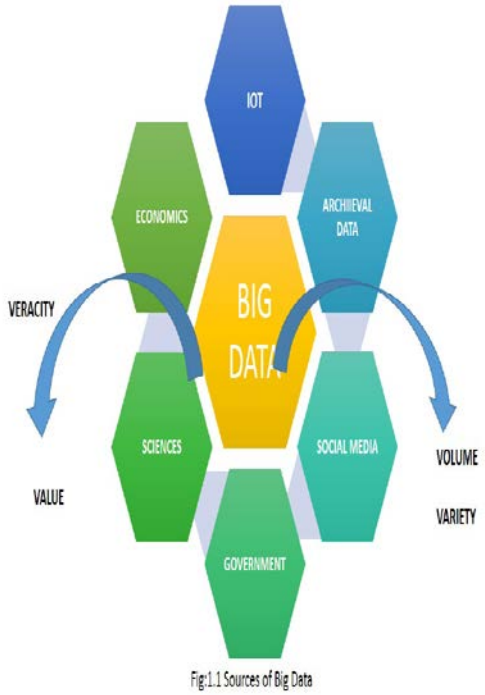

Fig:1.1 Sources of Big Data

ISSN 2348 – 7968

1. Huge volume of data: Rather than thousands or millions of rows, Big Data can be billions of rows and millions of columns.
2. Complexity of data types and structures: Big Data reflects the variety of new data sources, formats, and structures, including digital traces being left on the web and other digital repositories for subsequent analysis.
3. Speed of new data creation and growth: Big Data can describe high velocity data, with rapid data ingestion and near real time analysis.

Big data comprises various types of homogeneous and heterogeneous data which is difficult to store and process which includes

- Structured data: policy data, claim data, customer profile data and quote data.
- Unstructured data: social media data, insurance application documents, call center agent notes, claim adjuster notes and incident photographs.
- Semi-structured data: health records, customer profile data, weather reports, census data, webserver logs and emails [1].

In 2011, Mckinsey's report [2][4] defines *big data* as "datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze". This definition is subjective and does not define big data in terms of any particular metric. However, it incorporates an evolutionary aspect in the definition (over time or across sectors) of what a dataset must be to be considered as big data.

Architectural Definition: The National Institute of Standards and Technology (NIST) [3][4] suggests that, "Big data is where the data volume, acquisition velocity, or data representation limits the ability to perform effective analysis using traditional relational approaches or requires the use of significant horizontal scaling for efficient processing." In particular, big data can be further categorized into big data science and big data frameworks. Big data science is "the study of techniques covering the acquisition, conditioning, and evaluation of big data," whereas big data frameworks are "software libraries along with their associated algorithms that enable distributed processing and analysis of big data problems across clusters of computer units". An instantiation of one or more big data frameworks is known as big data infrastructure.

We also define Big data as huge flood data in various formats are generated by PC's, Smart Phones ,RFID readers and traffic CAMS which will be in order of magnitude larger *(volume)*; more diverse, including structured, semi structured, and unstructured data *(variety)*; and arriving faster *(velocity)* than you or your organization has had to deal with before.

Big-data system life cycle comprises four consecutive phases, including Data generation, Data acquisition, Data Storage, and Data Analytics.

Data generation concerns with how data are generated in huge volume which is termed as "big data" which contains diverse, and complex datasets from various heterogeneous and/or distributed data sources, including sensors, video, click streams, and other available digital sources. Normally, these datasets are associated with different levels of domain-specific values [5]. In current scenario the explosion of IOT will even add to the existing big data challenges. Datasets from three prominent domains, business, Internet, and scientific research, for which values are relatively easy to understand. However, there are overwhelming technical challenges in collecting, processing, and analyzing these datasets that demand new solutions to embrace the latest advances in the information and communications technology (ICT) domain.

In Data acquisition stage deals with the process of obtaining information which is further contains tasks of data collection, data transmission, and data pre-processing. Data collection refers to specific data collection technology that acquires raw data from a data production environments which comes from sources in this smart world in various formats and types. After collecting raw data, we need a high-speed transmission mechanism to transmit the data into the proper storage sustaining system like NAS, SAN etc., systems for various types of analytical applications and Finally, the collected datasets might contain many meaningless data, which unnecessarily increases the amount of storage space and affects the consequent data analysis. For instance, redundancy is common in most datasets collected from sensors deployed to monitor the environment, and we can use data compression technology to address this issue. Thus, we must perform data pre-processing operations for efficient storage and mining.

Data storage primarily concerns with persistent storage and managing of large-scale datasets. In data

ISSN 2348 – 7968

storage system we should consider both hardware infrastructure and data management. Hardware infrastructure consists of a pool of shared ICT resources organized in an elastic way for various tasks in response to their instantaneous demand like cloud computing technology which should be able to scale up and out and be able to be dynamically reconfigured to address different types of application environments. Data management software is deployed on top of the hardware infrastructure to maintain large-scale datasets. Additionally, to analyze or interact with the stored data, storage systems must provide several interface functions, fast querying and other programming models.

In the context of Big Data, data analysis plays vital role which is the base purpose of above stages of data we have studied, to inspect, transform, and model data to extract value. Many application fields leverage opportunities presented by abundant data and domain-specific analytical methods to derive the intended impact. Emerging analytics research can be classified into six critical technical areas: structured data analytics, text analytics, multimedia analytics, web analytics, network analytics, and mobile analytics. This classification is intended to highlight the key data characteristics of each area. In further sections of paper we discuss about various analytics related concepts in more detail as now-a-days these will play key role in there usage and its impact on the society.

2. Big Data Analytics

Data analytics is a discipline that includes the management of the complete data lifecycle, which encompasses collecting, cleansing, organizing, storing, analyzing and governing data. The term includes the development of analysis methods, scientific techniques and automated tools. In Big Data environments, it is challenge to develop and use highly scalable distributed technologies and frameworks that are capable of analyzing large volumes of data from different sources. Big data analysis differs from traditional data analysis primarily due to the volume, velocity and variety characteristics of the data being processes [1].

The Big Data analytics lifecycle [1] can be divided into the following nine stages, as shown in Figure 1.2:

1. Business Case Evaluation

2. Data Identification

3. Data Acquisition & Filtering

4. Data Extraction

5. Data Validation & Cleansing

6. Data Aggregation & Representation

7. Data Analysis

8. Data Visualization

9. Utilization of Analysis Results

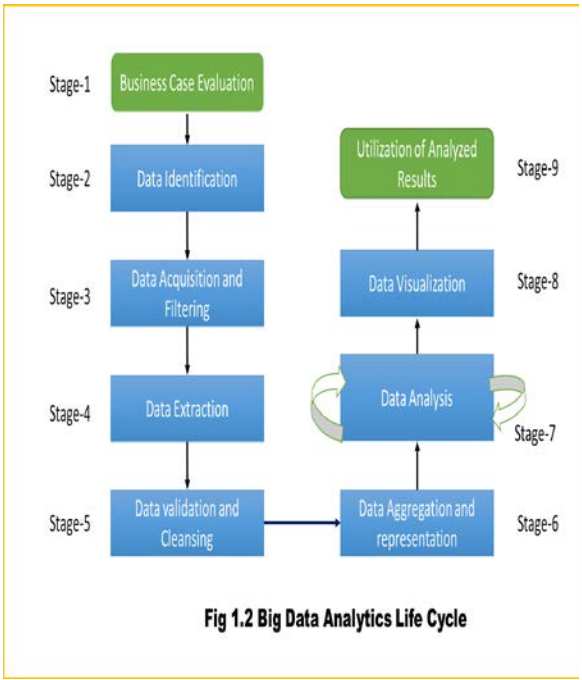

Fig 1.2 Big Data Analytics Life Cycle

1. An evaluation of a Big Data analytics business case helps decision-makers understand the business resources that will need to be utilized and which business challenges the analysis will tackle. The further identification of KPIs during this stage can help determine assessment criteria and guidance for the evaluation of the analytic results.

2. Stage -2 identifying the datasets required for the analysis project and their sources. Identifying a wider variety of data sources may increase the probability of finding hidden patterns and correlations. In the case of internal datasets, a list of available datasets from internal sources, such as data marts and operational systems, are typically compiled and matched against a predefined dataset specification. In the case of external datasets, a list of possible third-party data providers, such as data markets and publicly available datasets, are compiled. Some

forms of external data may be embedded within blogs or other types of content-based web sites, in which case they may need to be harvested via automated tools.

3. During the Data Acquisition and Filtering stage, the data is gathered from all of the data sources that were identified during the previous stage. The acquired data is then subjected to automated filtering for the removal of corrupt data or data that has been deemed to have no value to the analysis objectives. Data classified as "corrupt" can include records with missing or nonsensical values or invalid data types. Data that is filtered out for one analysis may possibly be valuable for a different type of analysis. Therefore, it is advisable to store a verbatim copy of the original dataset before proceeding with the filtering. To minimize the required storage space, the verbatim copy can be compressed.

4. Some of the data identified as input for the analysis may arrive in a format incompatible with the Big Data solution. The need to address disparate types of data is more likely with data from external sources. The Data Extraction is dedicated to extracting disparate data and transforming it into a format that the underlying Big Data solution can use for the purpose of the data analysis. The extent of extraction and transformation required depends on the types of analytics and capabilities of the Big Data solution. For example, extracting the required fields from delimited textual data, such as with webserver log files, may not be necessary if the underlying Big Data solution can already directly process those files.

5. Invalid data can skew and falsify analysis results. Unlike traditional enterprise data, where the data structure is pre-defined and data is pre-validated, data input into Big Data analyses can be unstructured without any indication of validity. Its complexity can further make it difficult to arrive at a set of suitable validation constraints. The Data Validation and Cleansing stage is dedicated to establishing often complex validation rules and removing any known invalid data. Big Data solutions often receive redundant data across different datasets. This redundancy can be exploited to explore interconnected datasets in order to assemble validation parameters and fill in missing valid data.

6. Data may be spread across multiple datasets, requiring that datasets be joined together via common fields, for example date or ID. In other cases, the same data fields may appear in multiple datasets, such as date of birth. Either way, a method of data reconciliation is required or the dataset representing the correct value needs to be determined. The Data Aggregation and Representation stage, is dedicated to integrating multiple datasets together to arrive at a unified view. Performing this stage can become complicated because of differences in:

• *Data Structure* – Although the data format may be the same, the data model may be different.

• *Semantics* – A value that is labeled differently in two different datasets may mean the same thing, for example "surname" and "last name."

The large volumes processed by Big Data solutions can make data aggregation a time and effort-intensive operation. Reconciling these differences can require complex logic that is executed automatically without the need for human intervention. Future data analysis requirements need to be considered during this stage to help foster data reusability. Whether data aggregation is required or not, it is important to understand that the same data can be stored in many different forms.

7. The Data Analysis stage is dedicated to carrying out the actual analysis task, which typically involves one or more types of analytics. This stage can be iterative in nature, especially if the data analysis is exploratory, in which case analysis is repeated until the appropriate pattern or correlation is uncovered. The exploratory analysis approach will be explained shortly, along with confirmatory analysis. Depending on the type of analytic result required, this stage can be as simple as querying a dataset to compute an aggregation for comparison. On the other hand, it can be as challenging as combining data mining and complex statistical analysis techniques to discover patterns and anomalies or to generate a statistical or mathematical model to depict relationships between variables.

8. The ability to analyze massive amounts of data and find useful insights carries little value if the only ones that can interpret the results are the analysts. The Data Visualization stage, is dedicated to using data visualization techniques and tools to graphically communicate the analysis results for effective interpretation by business users. The results of completing the Data Visualization stage provide users with the ability to perform visual analysis, allowing for the discovery of answers to questions that users have not yet even formulated. The same results may be presented in a number of different

ISSN 2348 – 7968

ways, which can influence the interpretation of the results

9. Based on the data analysis results, the underwriting and the claims settlement users have now developed an understanding of the nature of fraudulent claims. However, in order to realize tangible benefits from this data analysis exercise, a model based on a machine-learning technique is generated, which is then incorporated into the existing claim processing system to flag fraudulent claims.

2.1 Analytics Forms

Data analytics enable data-driven decision-making with scientific backing so that decisions can be based on factual data and not simply on past experience or intuition alone. There are four general categories of analytics that are distinguished by the results they produce[1]:

• descriptive analytics

• diagnostic analytics

• predictive analytics

• prescriptive analytics

2.1.1 Descriptive analytics

It is carried out to answer questions about events that have already occurred. This form of analytics contextualizes data to generate information. The following questions gives the requirements of descriptive analytics:

• What was the sales volume over the past 12 months?

• What is the number of support calls received as categorized by severity and geographic location?

• What is the monthly commission earned by each sales agent?

Descriptive analytics are often carried out via ad-hoc reporting or dashboards. The reports are generally static in nature and display historical data that is presented in the form of data grids or charts. Queries are executed on operational data stores from within an enterprise, for example a Customer Relationship Management system (CRM) or Enterprise Resource Planning (ERP) system.

2.1.2 Diagnostic Analytics

Diagnostic analytics aim to determine the cause of a phenomenon that occurred in the past using questions that focus on the reason behind the event. The goal of this type of analytics is to determine what information is related to the phenomenon in order to enable answering questions that seek to determine why something has occurred.

Such questions include:

• Why were Q2 sales less than Q1 sales?

• Why have there been more support calls originating from the Eastern region than from the Western region?

• Why was there an increase in patient re-admission rates over the past three months?

Diagnostic analytics results are viewed via interactive visualization tools that enable users to identify trends and patterns. The executed queries are more complex compared to those of descriptive analytics and are performed on multidimensional data held in analytic processing systems.

2.1.3 Predictive Analytics

Predictive analytics are carried out in an attempt to determine the outcome of an event that might occur in the future. With predictive analytics, information is enhanced with meaning to generate knowledge that conveys how that information is related. The strength and magnitude of the associations form the basis of models that are used to generate future predictions based upon past events. It is important to understand that the models used for predictive analytics have implicit dependencies on the conditions under which the past events occurred. If these underlying conditions change, then the models that make predictions need to be updated. Questions are usually formulated using a what-if rationale, such as the following:

• What are the chances that a customer will default on a loan if they have missed a monthly payment?

• What will be the patient survival rate if Drug B is administered instead of Drug A?

• If a customer has purchased Products A and B, what are the chances that they will also purchase Product C?

This kind of analytics involves the use of large datasets comprised of internal and external data and

ISSN 2348 – 7968

various data analysis techniques. It provides greater value and requires a more advanced skillset than both descriptive and diagnostic analytics. The tools used generally abstract underlying statistical intricacies by providing user-friendly front-end interfaces. Many of the fundamental algorithms for predictive analytics depend crucially on keeping the data in main memory with a single CPU to access it. Big Data breaks that condition. The data can't all be in memory at the same time, so it needs to be processed in a distributed fashion which should take help of cloud computing and its related technologies.

2.1.4 Prescriptive Analytics

Prescriptive analytics build upon the results of predictive analytics by prescribing actions that should be taken. The focus is not only on which prescribed option is best to follow, but why. In other words, prescriptive analytics provide results that can be reasoned about because they embed elements of situational understanding. Thus, this kind of analytics can be used to gain an advantage or mitigate a risk. Sample questions may include:

• Among three drugs, which one provides the best results?

• When is the best time to trade a particular stock?

This sort of analytics incorporates internal data with external data. Internal data might include current and historical sales data, customer information, product data and business rules. External data may include social media data, weather forecasts and government produced demographic data. Prescriptive analytics involve the use of business rules and large amounts of internal and external data to simulate outcomes and prescribe the best course of action.

Big data analytics uses various algorithms include C4.5, CART, AdaBoost, PageRank etc., along with various machine learning techniques to come to final decisions which is used at situations when demanded. Depending upon data streaming algorithm may changes accordingly.

3. Big Data Tools

3.1 Data Storage and Management Tools

An effective data storage provider should offer you an infrastructure on which to run all your other analytics tools as well as a place to store and query your data [6].

Hadoop

Hadoop is an ecosystem of open source components that fundamentally changes the way enterprises store, process, and analyze data. It is an open-source software framework for distributed storage of very large datasets on computer clusters by Apache Hadoop Project. All that means you can scale your data up and down without having to worry about hardware failures. Hadoop provides massive amounts of storage for any kind of data, enormous processing power and the ability to handle virtually limitless concurrent tasks or jobs by one of its module YARN. MapReduce based on YARN allows for parallel processing of large data sets stored on clusters.

Cloudera

Cloudera provides modern platform for data management and analytics for Apache Hadoop. It is essentially a brand name for Hadoop for building a data-driven business by securing your data in one secure place. They can help your business build an enterprise data hub, to allow people in your organization better access to the data you are storing. While it does have an open source element, Cloudera is mostly and enterprise solution to help businesses manage their Hadoop ecosystem. Essentially, they do a lot of the hard work of administering Hadoop for you. They will also deliver a certain amount of data security, which is highly important if you're storing any sensitive or personal data.

MongoDB

MongoDB is an open-source *document database* that provides high performance, high availability, and automatic scaling. MongoDB obviates the need for an Object Relational Mapping (ORM) to facilitate development. MongoDB is the modern, start-up approach to databases which is alternative to relational databases. It is good for managing data that changes frequently or data that is unstructured or semi-structured. Common use cases include storing data for mobile apps, product catalogs, real-time personalization, content management and applications delivering a single view across multiple systems. Again, MongoDB is not for the data newbie. One of the most important reasons for the popularity of MongoDB is that it is a JSON-friendly database. It means that documents are stored and retrieved from MongoDB as JavaScript objects

Talend

Talend is the leading open source integration software provider to data-driven enterprises. Talend is another great open source company that offers a number of data products. Talend's unified platform architecture meets all of our needs with regards to data integration and data governance. Facilities like Master Data Management (MDM) is offered, which combines real-time data, applications, and process integration with embedded data quality. Because it's open source, Talend is completely free making it a good option no matter what stage of business it is and helps us to build and maintain your own data management system – which is a tremendously complex and difficult task.

3.2 Data Cleaning Tools

Big Data solutions often receive redundant data across different datasets. This redundancy can be exploited to explore interconnected datasets in order to assemble validation parameters and fill in missing valid data. Data should be cleaned to be used for effective analytics. Data sets can come in all shapes and sizes (some good, some not so good!), especially when you're getting it from the web. The companies below will help you refine and reshape the data into a useable data set.

OpenRefine

OpenRefine (formerly GoogleRefine) is an open source tool that is dedicated to cleaning messy data. Huge data sets can be processed easily and quickly even if the data is a little unstructured. OpenRefine can be used in transforming it from one format into another; and extending it with web services and external data. OpenRefine can be used to link and extend your dataset with various webservices. Some services also allow OpenRefine to upload your cleaned data to a central database The nice thing about OpenRefine is that it has a huge community with lots of contributors meaning that the software is constantly getting better and better [7].

DataCleaner

DataCleaner gives impetus to data quality which provides completeness, and meaning for data to be used by business organizations. Data visualization tools can only read nicely structured, "clean" data sets. DataCleaner does the hard work for you and transforms messy semi-structured data sets into clean readable data sets that all of the visualization companies can read. The DataCleaner is a strong data profiling engine for discovering and analyzing the quality of your data. Find the patterns, missing values, character sets and other characteristics of your data values. DataCleaner is built to handle data both big and small. Give everything from CSV files, Excel spreadsheets to Relational Databases (RDBMs) and NoSQL databases [8].

3.3 Data Mining Tools

Data mining is the process of discovering insights within a database as opposed to extracting data from web pages into databases and then making sense of the data through IT strategies and tools, which can project future trends and behaviors. The tools used in data mining can provide answers to many business questions, which conventionally require too much time for resolution. Data miners can dig through databases for concealed patterns, searching for predictive information that specialists may miss because they are beyond the normal pattern.

RapidMiner

RapidMiner's modern enterprise platform, you can quickly and easily create analytic workflows called *processes* to determine who to target. RapidMiner is a fantastic tool for predictive analysis. It's powerful, easy to use and has a great open source community behind it. We can even integrate own specialized algorithms into RapidMiner through their APIs.

The IBM SPSS Modeler offers a whole suite of solutions dedicated to data mining. This includes text analysis, entity analytics, decision management and optimization. Their five products provide a range of advanced algorithms and techniques that include text analytics, entity analytics, decision management and optimization.

TERADATA having a long standing history provides end-to-end solutions and services in data warehousing, big data and analytics and marketing applications. Teradata Integrated Data Warehouse contains nearly every business department, including merchandising, marketing, customer service, credit and IT. Among number of tools Teradata gives new insights around customer experience through Teradata Aster n-Path™ analysis [9].

3.4 Data Analysis Tools

While data mining is all about sifting through your data in search of previously unrecognized patterns, data analysis is about breaking that data down and assessing the impact of those patterns overtime.

ISSN 2348 – 7968

Analytics is about asking specific questions and finding the answers in data. You can even ask questions about what will happen in the future!

Qubole

Qubole simplifies, speeds and scales big data analytics workloads against data stored on AWS, Google, or Azure clouds. They take the hassle out of infrastructure wrangling. Once the IT policies are in place, any number of data analysts can be set free to collaboratively "click to query" with the power of Hive, Spark, Presto and many others in a growing list of data processing engines.

BigML

BigML is attempting to simplify machine learning. They offer a powerful Machine Learning service with an easy-to-use interface for you to import your data and get predictions out of it. You can even use their models for predictive analytics. A good understanding of modeling is certainly helpful, but not essential, if you want to get the most from BigML. They have a free version of the tool that allows you to create tasks that are under 16mb as well as having a pay as you go plan and a virtual private cloud that meet enterprise-grade requirements.

Statwing

Statwing takes data analysis to a new level providing everything from beautiful visuals to complex analysis. This tool provides web-based statistical analysis software for business users, data analysts, and market researchers. Statwing makes data analysis easy. It encodes statistical best practices and into software, so nontechnical users can point and click to visualize and understand data like experts. Existing tools like SPSS, R, and SAS are highly technical and very difficult to use for the half of their users who do not have technical backgrounds.

3.5 Data Visualization Tools

Data visualization companies will make your data come to life. Main challenge for any data scientist is conveying the information of interest from that data to the rest of company in effective way. But some of databases like, MySQL databases and spreadsheets aren't sufficient to do it. So Visualizations are a bright and easy way to convey complex data insights.

Tableau

Tableau is a data visualization tool with a primary focus on business intelligence. You can create maps, bar charts, scatter plots and more without the need for programming. They recently released a web connector that allows you to connect to a database or API thus giving you the ability to get live data in a visualization.

Silk

Silk is a much simpler data visualization and analytical tool than Tableau. It allows to bring your data to life by building interactive maps and charts with just a few clicks of the mouse. Silk also allows you to collaborate on a visualization with as many people as required.

CartoDB

CartoDB is a data visualization tool that specialises in making maps. It makes easy for anyone to visualize location data – without the need for any coding. CartoDB can manage a myriad of data files and types, they even have sample datasets that you can play around with while you're getting the hang of it. It may not be the easiest system to use, but once you get the hang of it, it is incredibly powerful.

Chartio

Chartio allows you to combine data sources and execute queries in-browser. You can create powerful dashboards in just a few clicks. Chartio's visual query language allows anyone to grab data from anywhere without having to know SQL or other complicated model languages. It is simple to generate PDF reports to export and email your dashboard.

Pentaho

Pentaho offers big data integration with zero coding required. Using a simple drag and drop UI you can integrate a number of tools with minimal coding. They also offer embedded analytics and business analytics services too.

3.6 Data Languages

In earlier days a tool simply is not sufficient to use it but today's tools are becoming more powerful and easier to use, sometimes it is just better to code it yourself. Even if you're not a programmer, understanding the basics of how these languages work will give you a better understanding of how many of these tools function and how best to use them.

R

R is a language for statistical computing and graphics. If the data mining and statistical software listed above doesn't quite do what you want it to, learning R is the way forward. In fact, if you're planning on being a data scientist, knowing R is a requirement.It can run on Linux, Windows and MacOS and you can download R. There is a huge community of statisticians using R nowadays and its popularity is always growing.

Python

Another language that is gaining popularity in the data community is Python. Created in the 1980s and named from Monty Python's Flying Circus, it has consistently ranked in the top ten most popular programming languages in the world. Many journalists use Python to write custom scrapers if data collection tools fail to get the data that they need.

RegEx

RegEx or Regular Expressions are a set of characters that can manipulate and change data. It's used mainly for pattern matching with strings, or string matching. At Import.io, you can use RegEx while extracting data to delete parts of a string or keep particular parts of a string. It is an incredibly useful tool to use when doing data extraction as you can get exactly what you want when you extract data meaning you don't need to rely on those data manipulation companies mentioned above but expert in programming is necessary.

4. Conclusion

In this paper we have studied various concepts relating to Big Data life cycle, Big Data Analysis and Tools used in various stages. These study throws light into these areas, where researchers can further probe into these for better solutions to the issues which are challenging to IT world.

5. References

[1].Thomas Erl et.al , Big Data Fundamentals Concepts, Drivers & Techniques, Prentice Hall, First Edition 2016

[2] J. Manyika et al., Big data: The Next Frontier for Innovation, Competition, and Productivity. San Francisco, CA, USA: McKinsey Global Institute,2011, pp. 1 to137

[3] M. Cooper and P. Mell. (2012). Tackling Big Data[Online].Available:http://csrc.nist.gov /groups/SMA/forum/documents/june2012 presentations /%csm_june2012_cooper_mell.pdf

[4] Han Hu, Yonggang Wen et.al, Toward Scalable Systems for Big Data Analytics:A Technology Tutorial, IEEE Access, July 8, 2014.

[5] Big Data in the Cloud: Converging Technologies,Intel White Paper

[6]. https://www.import.io/post/all-the-best-big-data-tools-and-how-to-use-them/

[7] Available online: http://openrefine.org/

[8] http://datacleaner.org/resources/docs/3.0.3 /html_single/

[9] http://in.teradata.com/about-us