**International Journal of Advanced Research and Development**
**ISSN: 2455-4030, Impact Factor: RJIF 5.24**
**www.advancedjournal.com**
**Volume 2; Issue 3; May 2017; Page No. 147-149**

# Data analytics on big data

**[1] RR Karthikeyan, [2] Dr. B Raghu**

[1] Research-Scholar, Bharath University, Chennai, Tamil Nadu, India
[2] Principal, SVS Groups of Institutions, Warangal, Telangana, India

**Abstract**
Data analysis the combined process of data inspection, data cleaning, transformation and data modeling. Data analysis gives more useful results and mined data for decision making process in various business Industries like retail, banking and social domain.
Data integration is the pre work for the data analysis, data visualization and data dissemination are closely linked with data analysis.
Data analysis could be done on any of the structured, semi structured and an unstructured data.
In fast paced and dynamic market, Data analyzed from any of the Data analysis tools being used for the following benefits

- Market trend could be captured in the competitive market place.
- To understand the customer and their behavior.
- To minimize the operational inefficiency to zero level.
- Arrive at optimal business decision with confidence and avoid the bad decision.
- Identify the data and data sources which are clouding business horizon.

**Keywords:** big data, data analytics, preprocessing of input data, steps in big data analytics, data mining in big data analytics

## 1. Introduction

### Types of data can be analyzed by Big Data

### Structured data

Majority of the analytical platforms are running with structured data. These all are tables or other data structures of relational data base. But other sources are may be flat files such as record files, character-delimited rows yield predictable structures.

### Semi structured data

In particular, today at least half of analytical tools or platform running semi structured data (XML and similar standards) or complex data (hierarchical or legacy sources). These data types are driven up by increased use of industry standards (SWIFT, ACORD, and HL7) and XML applied to business-to-business data exchange (which tends to be modeled in hierarchies).

### Web Data

Web data is finally getting the attention it deserves. The skills and the platforms emerged for web data, used in exploring social media data (blogs, tweets, social networks) and Web logs and clickstreams.

### Real-time data

This types are little less at the moment. But they stand a good chance of becoming more common as real-time technologies continue to improve and to be adopted by user organizations. This includes event data, spatial data and machine-generated data (from sensors, RFID chips, robots, and various devices.
Data streams are infinite length of continuous data which either may be structured or unstructured. Data stream could be defined as class of data generated as text, audio or video.

## 2. Tools for Big Data Analytics

### Discovery Tools

This tools can be used anywhere in the information cycle for rapid and dynamic data analysis in any combination of structured and unstructured data. This tools can run along with BI systems and does not require any up front designing. User can draw new insights, meaningful conclusions and quick business decisions.

### BI Tools

This tools are important for analyzing, reporting and performance management in real time transactional data of data warehouse system. Various enterprise reporting provided in the form of scorecards, dashboard, and ad-hoc analysis to improve the comprehensive capabilities for business intelligence.

### In Memory -Database Analytics

When the data grows exponentially, the data movement between the data storage and processing engine may decrease the performance by slowing down the process speed. To overcome the above said problem, business intelligence technique are applied in the directly with in the data base to find the useful pattern and relationship.

### Hadoop

Hadoop is the frame works which includes the map reduce parallel processing and SQL enabled business intelligence system provide the fast accurate preprocessing system. Special inexpensive hardware called commodity hardware provide the support for fast growing voluminous data.

**Decision Management system**
The set of predictive modeling, business rules and self-learning nodes forms the decision management. This system maximizing the every customer input recommendations by passing it to multiple channels.

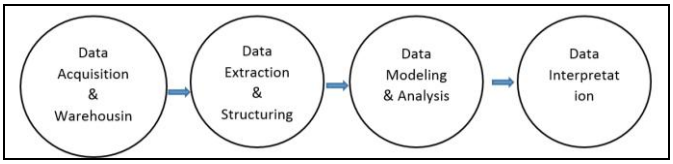


**Fig 1:** Phases of Data analysis in Big Data Analysis.

**3. Phases of Data analysis in Big Data**
There are some important point's needs to be considered on credibility, uncertainty and gap in identifying and picking the relevant data from the voluminous data. To manage such historic and real time voluminous data requires smart systems and also better human collaboration for user interaction.
The new system should not affect the fundamental basis of managing the data analysis. Instead, it should built on the existing Structures, warehouses and analytics to improve those existing capabilities in data quality, Master Data Management and data protection frameworks.

**Data Acquisition & Data Warehousing**
Data always come from the multiple sources and they can produce the millions terabytes of raw data daily from the sources like social media users, internal organization data engine, data analysis engine for private data and non-social media users.
The term data reduction is the intelligent science that extract all the piece of relevance information without missing a single piece of useful information. It is more expensive since it needs the raw real data and sore them first and reduce them in user-friendly size.
Good data warehouse platform should collect and consolidate the data across the various sources. Apart from creating the repository of master data it should provide the consistent information across the organization.

**Data Extraction & Structuring**
Even after data collection and filtering, it is not in ready format for analysis. Generally data may be in any of the content such as text, pictures, videos, multiple sources of data with different file formats.
For those above reasons it is mandatory to do the data extraction to integrate the data from diverse enterprise information system and transform in to the consumable format.
There are two categories of data, structured and unstructured. Structured data is available in row column format database and easy to enter, store and analysis.
Data is basically of two categories – structured and unstructured. Structured data is that which is available in a pre-set format such as row and column based databases. This type of data is mostly actual and transactional. Whereas unstructured data are free form and behavioral like tweets. These data may be in any form and heterogeneous. Unstructured data are growing at exponential with rapid speed.
Extract-Transform-Load (ETL) process ensuring the entire cycle of the data getting loaded in to the target data warehouse after proper data cleaning process.

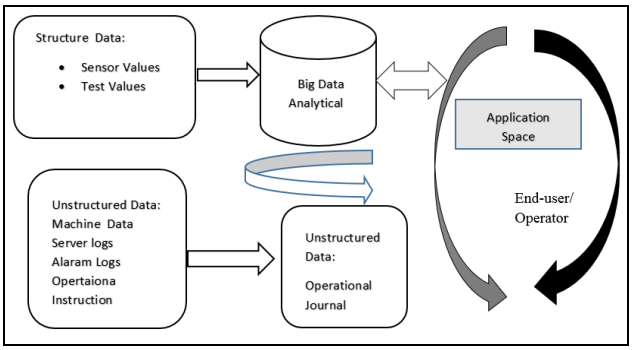


**Fig 2:** Architecture diagram for Big Data Analysis.

**Data Modeling & Data Analysis**
Once the data acquisition and extraction is completed, data repository is ready to use. Now the data has all the consolidated information without missing a single piece of relevance. Then complex procedure data analysis is applied. Data analysis is not only the process of locating, identifying and understanding the data, need complete automation which requires processing of different data structure(structure, unstructured), presenting complex semantics in an understandable and computer intelligent format.
Technological advancements easing analysis of unstructured data possible well and cost effective. In this heterogeneous environment all data base are easily scalable, many of them are non-relational and parallel. So there is a need for the redefining the data base. NoSQL database extending its help to grow horizontally, avoid joins and working without fixed table schemas.
Data analysis in Big Data emerged as the well-equipped science that collect the various techniques from statistics, mathematical modeling and visualization as follows.
Data manipulation & analytic applications addressing automation, application development and testing.
Data modelling covering key areas like experimental design, graphical models and path analysis.
Statistics and machine learning through classical and spatial statistics, simulation and optimization techniques Text data analysis through pattern analysis, text mining and NLP by developing.
Integrating solutions or deploying packaged solutions

**Data Interpretation**
Output of the big data analysis become successful only when the analyzed data presented in the user friendly way and reusable and intelligent format. Complexity of the analyzed data adding complexity to the presentation also.
Many times simple tabular presentation is not enough to present historical data need further explanation. Output from the predictive analysis supports and provide the critical inputs for decision making. Interactive Data Visualization is the next big thing to support the data presentation.
From static graphs and spreadsheets to using mobile devices and interacting with data in real time – the future of data interpretation is becoming more agile and responsive.

## 4. Conclusion

This paper has explained about the types of data those can be handled in Big Data analytics and the steps involved in it. Following use cases can be implemented using a good big data engine.

- An architecture for managing and processing Big Data using grid technologies.
- The machine learning computational models for Big Data.
- Cost-effective design on kernel-based machine learning and classification for Big Data learning applications.
- Approaches to analyze unstructured data like imagery, sensors, telemetry, video, documents, log files, and email data files.
- Investigated energy efficient architecture for Big Data application
- Big Data and Big compute by combining Hadoop clusters and MPI clusters.

## 5. References


1. Atzmueller M. Subgroup Discovery – Advanced Review. WIREs: Data Mining and Knowledge Discovery, 2015; 5(1):35-49.
2. Atzmueller M. Knowledge-Intensive Subgroup Mining -- Techniques for Automatic and Interactive Discovery. Dissertations in Artificial Intelligence-Infix (Diski), (307) IOS Press, 2007.
3. Hong B *et al.* (Eds.): DASFAA Workshops LNCS 7827, 2013. © Springer-Verlag Berlin Heidelberg 2013
4. Zadrozny P, Kodali R. Big Data Analytics using Splunk, Berkeley, CA, USA: Apress, 2013, 1-15.
5. Ohlhorst F. Big Data Analytics: Turning Big Data into Big Money, Hoboken, NJ, USA: Wiley, 2013.
6. Dean J, Ghemawat S. MapReduce: Simplified data processing on large clusters, Commun ACM, 2008; 51(1):107-113.
7. Apache Hadoop, http://hadoop.apache.org.