International Research Journal of Education and Technology
Peer Reviewed Journal
ISSN 2581-7795

BIG DATA ANALYTICS

Jayamani. R

B. Tech-Information Technology

Dr. Mahalingam College of engineering and technology

Pollachi, India

Harini.M

B. Tech-Information Technology

Dr. Mahalingam College of engineering and technology

Pollachi,India

Balusamy Nachiappan,

Technical lead Prologis, 1800 Wazee Street, Suite 500

Denver, CO 80202, United States

International Research Journal of Education and Technology
Peer Reviewed Journal
ISSN 2581-7795

Abstract- Big data is a new technological paradigm for data that is generated at high speed, high volume, and with great diversity. Big data is seen as a revolution that has the potential to revolutionize the way companies operate in many industries. This paper introduces big data and the dimensions of data quality where the challenges of the quality factors of big data quality are discussed, and the lifecycle of big data analytics is discussed.

Key Words- Big data, Big Data Quality, BigData Quality Dimensions, Big Data Analysis

I. Introduction

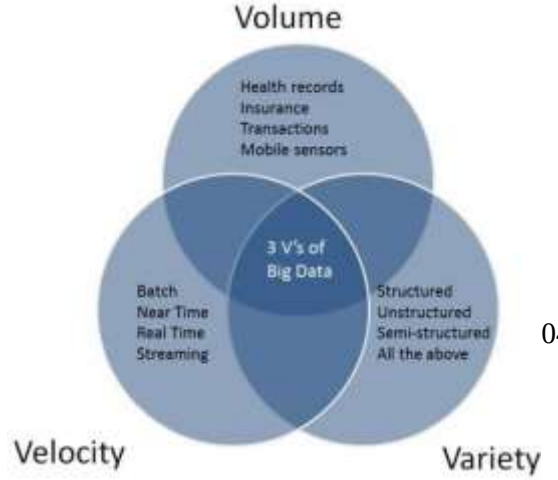
Big data refers to the concept of very large data sets involving three major dimensions or properties named (3Vs). First is a volume according to the amount of data located in the storage medium. Second is Variety which refers to the various heterogeneous and complex types of data. Data can be structured, unstructured, or semi structured generated either by humans or machines. The thirdis velocity which indicates the speed of data processing required to handle that large amount of data. Most definitions of big data focus on the size of data in storage. Size matters, but there are other important attributes of big data, namely data variety and data velocity. The three Vs of big data (volume, variety, and velocity) constitute a comprehensive definition, and they bust the myth that big data is only about data volume. In addition, each of the three Vs has its own ramifications for analytics.

[1].3 v‟s of Big Data

"Big data is a term that describes large volumes of high velocity, complex and variable data that require advanced techniques and technologies to enable the capture, storage, distribution, management, and analysis of the information." A large data volumes are daily generated at unprece-dented rate from heterogeneous sources (e.g., health, government, social networks, marketing, financial). This is due to many techno- logical trends, including the Internet of Things, the proliferation of the Some of the difficulties related to big data include capture, storage, search, sharing, analytics, and visualizing. Today, enterprises are exploring large volumes of highly detailed data so as to discover facts they didn‟t known before. Big data is data whose scale, distribution, diversity, and/or timeliness require the use of new technical architectures, analytics, and tools in order to enable insights that unlock new sources of business value.

II. Main Features

characterize big data: volume, variety, and velocity, or the three V‟s. The volume of the data is its size, Velocity refers to the rate with which data is changing, or variety includes the different formats and types of data, as well as the different kinds of uses and ways of analysing the data. Big Data Quality& Big Data Quality Dimensions Huge quantities of data does not automatically guarantee quality. And with larger volumes it becomes more important to focus on the quality in order to derive some meaningful insights out of the available data. In most contexts the worth of the data is determined by its „fitness for use‟, this criteria of data is still the central part to determine the necessity or mandate for organizations to invest in big data. Data quality



for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications.

[2].Big data quality dimension

Accessibility and availability are related to the ability of the user to access data from his or her culture, physical status/functions, and technologies available. Consistency, cohesion and coherence refer to the capability of data to comply without contradictions to all properties of the reality of interest, as specified in terms of integrity constraints, data edits, business rules and other formalisms

Completeness: It measures the degree with which a dataset is complete. It is evaluated by assessing the ratio between the amount of values currently available in the dataset and the expected amount of values. The expected amount of values considers both null values in available registrations and missing registrations. Note that, as regards data streams, missing registrations are easy to detect if data are sensed with a specific frequency. If data are not collected at a regular pace it is possible to rely on historical data to estimate the sampling frequencythat often varies over time.

Consistency: It refers to the violation of semantic rules defined over a 7 set of data items. Therefore, this dimension can be calculated only if there is the availability of a set of rules that represent dependencies between attributes. We have developed a module that detects functional dependencies and checks that values in the dataset respect them.

Timeliness: refers to the time expectation for accessibility and availability of information. Timeliness can be measured as the time between when information is expected and when it is readily available for use, this concept is of particular interest, because synchronization of data updates to application data with the centralized resource supports the concept of the common, shared, unique representation. The success of business applications relying on master data depends on consistent and timely information

Validity: Is the data presented in the correct and pre-defined format, type or range so as to be applicable to the given analytical task . The data set may be complete but does it tell the user what it purports to and is it valid for the current business context. Data quality not only depends on the completeness but also on the business environment and the business purpose it is supposed to serve. Only the data that conform to the business requirements can be considered valid.

III. Challenges

Big Data can bring cost saving, risk control, improvement of management efficiency, and increment of value into enterprise. In the meanwhile, Big Data brings some challenges Unevenness of Data Quality: Though the first step of processing data is to gather data, if the gather all data in spite of quality, it is possible to make wrong predictions and decisions. according to view of this condition, after gathering data, it is necessary to select relative data and clean conflicting data.

Lack of skills: Big data application requires enterprise to design new data analysis models. That‟s because traditional models are fit to process structured data not big data including multi-type data. The enterprise is short of talents who can design new data analysis models. The talents who not only can design new data analysis models but also know the financial management are fewer. Lack of talents is a severe and long-term issue. Through affecting the idea, function, mode, and method of financial management, it can bring cost saving, risk

control, improvement of management efficiency, and increment of value into enterprise.

IV. Big Data Features

The diversity of data sources brings abundant data types and complex data structures and increases the difficulty of data integration One data type is unstructured data, for example, documents, video, audio, etc. The second type is semi-structured data, including: software packages/modules, spreadsheets, and financial reports. The third is structured data. The quantity of unstructured data occupies more than 80% of the total amount of data in existence. - Data change very fast and the “timeliness” of data is very short, which necessitates higher requirements for processing technology – Due to the rapid changes in big data, the “timeliness” of some data is very short. If companies can‟t collect the required data in real time or deal with the data needs over a very long time, then they may obtain outdated and invalid information. In order to guarantee the product quality and improve benefits to enterprises, in 1987 the International for Standardization (ISO) published ISO9000

International Research Journal of Education and Technology
Peer Reviewed Journal
ISSN 2581-7795

standards. Nowadays, there are more than 100 countries .

V. Big Data Analysis

Big data analytics is differences from traditional analytics Because of the big increase in the volume of data and that led to Many researchers have suggested commercial DBMS and this not suitable with size of data. This type of data is impossible to handle using traditional relational database management systems. New innovative technologies were needed and Google found the solution by using a processing model called Map Reduce. There are more solutions to handle Big Data, but the most widely-used one is Hadoop, an open source project based on Google‟s Map Reduce and Google File System. Hadoop was founded by the Apache Software Foundation. The main contributors of the project are Yahoo, Facebook, Citrix, Google, Microsoft, IBM, HP, Cloudera and many others. Hadoop is a distributed batch processing infrastructure . Big Data Analytics refers to the process of collecting, organizing, analyzing large data sets to discover different patterns and other useful information. Big data analytics is a set of technologies and techniques that require new forms of integration to disclose large hidden values from large datasets that are different from the usual ones, more complex, and of a large enormous scale. It mainly focuses on solving new problems or old problems in better and effective ways.

VI. Types Of Analysis

TYPE	DESCRIPTION
Descriptive Analysis	It simplifies the data and summarises the past data into a readable form.
Prescriptive Analysis	It allows businesses to determine the best possible solution available to a problem.
Diagnostic Analysis	It gives detailed and in-depth insights into the root cause of the problem.
Predictive Analysis	This type of analytics makes use of historical and present data to predict future events.

VII. Stages

The Big Data Analytics Lifecycle defines analytics process best practices spanning discovery to project completion.

•Discovery

•Operationalize

• preparation

•Communicate

• Model planning

Stage 1 (Discovery): In Stage 1, the team learns the business domain, including relevant history such as

whether the organization or business unit has attempted similar projects in the past from which they can learn.

Stage 2 (Data preparation): Stage 2 requires the presence of an analytic sandbox, in which the team can work with data and perform analytics for the duration of the project.

Stage 3: Model planning: Stage 3 is model planning, where the team determines the methods, techniques, and workflow it intends to follow for the subsequent model building Stage.

Stage 4: Model building: In Stage 4,the team develops datasets for testing, training, and production purposes. In addition, in this Stage the

team builds and executes models based on the work done in the model planning Stage.

Stage 5: Communicate results: In Stage 5, the team, in collaboration with major stakeholders, determines if the results of the project are a success or a failure based on the criteria developed in Stage 1. The team should identify key findings, quantify the business value, and develop a narrative to summarize and convey findings to stakeholders.

Stage 6: Operationalize: In Stage 6, the team delivers final reports, briefings, code, and technical documents. In addition, the team may run a pilot project to implement the models in a production environment.

[3].Stages In Analysis

VIII. Conclusion

Big data refers to the set of numerical data produced by the use of new technologies for personal or professional Purposes. In this paper, we have studied Big Data characteristics and discussed the challenges of big data quality. might affect raised by Big Data. Also, Big Data analytics is the process of examining these data in order to uncover hidden patters. Big Data Analytics is a fast growing technology. But difficult degree of analysis of these data in the framework of the Big Data is a process that depended on kind of process which required.

IX. References

Textbook References:

• Too Big to Ignore: The Business Case for Big Data, by Phil Simon.

• The Data Revolution: Big Data, Open Data, Data Infrastructures, And Their Consequences, By Rob Kitchin.

• Big Data at Work: Dispelling the Myths, Uncovering the Opportunities, by T. H. Davenport.

• Big Data in Practice By Bernard Marr.

Web References:

• http://www.analyticsvidhya.com/

• http://www.dataversity.net/

• http://www.smartdatacollective.com/

• http://www.datasciencecentral.com/

• http://planetbigdata.com/