Big Data: A Review

Nadir Abdelrahman Ahmed Farah
Information Systems Department
University of Bisha
Alnamas, Saudi Arabia
Email: nadiraaf [AT] hotmail.com

Abstract— **One of the important issues in the Information Technology is the Big data. Combination of huge amount of data is generally called Big data. It becomes difficult to process data using database management tools. Big data are rapidly growing in all science and technological areas. The paper starts with the introduction of Big data then discuss its Definition, its Characteristics, Architecture & Platforms, Classification, Difference Between Big data and Traditional data, Life cycle, Analytics, Techniques, Business value, Data acquisition, Relationship between IOT and Big data, Data quality, and Data challenges.**

Keywords- **Big Data, Acquisition, Quality, Architecture ,Techniques, Analytics.**

Introduction

In today environment, data is generated from various sources. This data is in different forms. Capturing, retrieving, extracting, analyzing, manipulating and storing of this data is bit critical. This amount of massive data is considered as Big Data.

Big data is a popular term used to describe the fast growth and availability of data, both structured and unstructured**.** Structured data are numbers and words that can be easily categorized and analyzed. This data are generated by devices like network sensors embedded in electronic devices, smartphones, and global positioning system (GPS) devices. Structured data also include items like sales figures, account balances, and transaction data. Unstructured data includes more complex information, such as customer reviews from commercial websites, photos and other multimedia, and comments on social networking sites. This data cannot easily be divided into categories or analyzed numerically.

1. BIG DATA DEFINITION & CHARACTERISTICS

Big data is data that surpass the processing capacity of traditional database systems. The data is too big, moves too fast, or doesn't fit the limits of your database architectures. To gain value from this data, we must choose an alternative way to process it. [1]. Big data Characteristics includes the following[2][3]:

- **Volume:** Refers to the size of the data. Along with the growth of social media, the data volume is also growing very fast.
- **Velocity:** Refers to the speed at which data is generated.
- **Variety:** refers to the different formats in which data is generated. 70% of the data generated today is in an unstructured manner.
- **Value :** It is refers to the important feature of the data which is defined by the added-value that the collected data can bring to the intended process, activity or predictive analysis.
- **Veracity:** Refers to the accuracy or truth of the data.

2. BIG DATA ARCHITECTURE & PLATFORMS

The Big data is divided into four layers shown in Figure 1[4]

- **Infrastructure as a Service (IaaS):** This includes the storage, servers, network as the base, and inexpensive commodities of the big data stack.
- **Platform as a Service (PaaS):** The NoSQL data stores and distributed caches that logically queried using query languages which form the platform layer of big data.
- **Data as a Service (DaaS):** The entire array of tools available for integrating with the PaaS layer using search engines, integration adapters, and batch programs.
- **Big Data Business Functions as a Service (BFaaS):** Specific industries like health, retail, ecommerce, and banking can build packaged applications that serve a specific business need and leverage the DaaS layer for cross-cutting data functions.

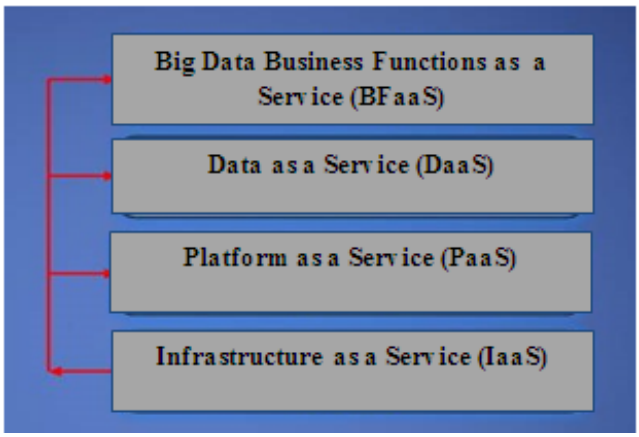

Figure 1. Big Data Architecture

3. CLASSIFICATION OF BIG DATA

Big data can split into classes. These classes are Data Sources, Content Format, Data Stores, Data Staging and Data Processing[2].

- **Data Sources:** Web & Social, Machine, Sensing, Transactions and IoT
- **Content Format:** Structured, Semi-Structured and Unstructured.
- **Data Stores:** Document-oriented, Column-oriented, Graph based and Key-value.
- **Data Staging:** Cleaning, Normalization and Transform .
- **Data Processing:** Batch and Real time resulted of the wide variety of data sources, the captured data differ in size with respect to idleness, consistency and noise, etc.

4. DIFFERENCE BETWEEN BIG DATA AND TRADITIONAL DATA

There are various differences between the traditional data and big data as you can see in the table 1 below [3].

Table 1 Difference between traditional and big data

FEATURES	RELATIONAL ATABASE	BIG DATA
Database Architecture	Centralized	Distributed
Data types	Structured	Semi-structured, Unstructured
Data volume range	Gigabytes to Terabytes	Terabytes, Petabytes and beyond
Data schema	Fixed or Static	Dynamic
Hardware / software cost	Higher	Lower

5. LIFE CYCLE OF BIG DATA

The life cycle of big data can be divided into four phases: (1) collection, (2) compilation and consolidation, (3) data mining and analytics, and (4) use.
As for the first step, not all data starts as big data. Rather, companies collect bits of data from a variety of sources. For example, as consumers browse the web online, companies can track and link their activities. Other times, techniques such as tracking cookies, browser or device fingerprinting, and even history sniffing identify who consumers are, and what they do. After collection, the next step in the life cycle of big data is compilation and consolidation. Commercial entities that compile data include online ad networks, social media companies, and large banks or retailers. One important category of commercial entities that compile and consolidate data is the data brokers. They combine data from various sources to build profiles about individual consumers. The third step is data analytics. One form of analytics is descriptive objective which is to uncover and summarize patterns or features that exist in data sets. By contrast, predictive data analytics refers to the use of statistical models to generate new data. The final step in the life cycle of big data is use [5].

6. BIG DATA ANALYTICS

Storing large data sets is worthless unless they become useful. Big data has effect on data analysis as well as data storage, processing, and management. As big data expands in all areas including science, industrial, business, healthcare, and military. Past data analysis techniques and architectures have become insufficient to analyze such an expansion. Analysis of big data is named in different ways such as BDA, advanced analytics, analytics, large-data-set analytics. BDA is the process of examining large data sets by using statistical models, datamining techniques, and computing technologies. This process includes discovering hidden business style and secret correlations, market orientation, customer preferences and other useful business information [6].

7. BIG DATA TECHNIQUES

Big data techniques include data mining, machine learning, neural networks, signal processing and visualization approaches[8].

- **Data Mining Technique:** Data mining involves a set of techniques to extract useful patterns from large amount data. Data mining is a process which involves number of methods such as clustering analysis, classification, regression and association rule learning.
- **Neural Networks:** Neural networks have a remarkable ability to extract complex styles and detect trends that are too complex to derive by either human beings or computer oriented techniques. Neural network is a network of continuous valued inputs and outputs.
- **Machine Learning:** is an important application of artificial intelligence which allows computers to explore behaviors on the basis of empirical data. The algorithms designed for machine learning have an obvious characteristic to discover knowledge and make intelligent decisions automatically.
- **Visualization Techniques:** are used to display the data in the form of tables, images, diagrams and other interactive display methods. When I think about large scale data visualization, many researchers use different techniques such as feature extraction to reduce the size of data before displaying the actual data.

- **Digital signal processing (DSP) :** is particularly motivated by the need to extend traditional signal processing techniques.

8. THE BUSINESS VALUE OF BIG DATA ANALYTICS

The main benefits of big data analytics are: i) to draw insight from data, ii) to make better decision based on the insight, and iii) to automate the decision and bake it into a business process, hence process automation.
In a more detailed level, each big data solution may address particular business problems the organizations face and the business value of the solution is further connected to the original business problems. When building a business case for big data analytics project, it is important to start with a business problem, not data or technology. Gathering data or purchasing technology without a clear business target is a losing planning [7].

9. BIG DATA ACQUISITION

Big data acquisition includes data collection, data transmission, and data pre-processing. During big data acquisition, once we collect the raw data, we shall utilize an efficient transmission mechanism to send it to a proper storage management system to support different analytical applications[9].

- **Data Collection:** Data collection is to use special data collection techniques to get raw data from a specific data generation environment.
- **Data Transportation:** after completion of raw data collection, data will be transferred to a data storage infrastructure for processing and analysis. Big data is mainly stored in a data center. The data layout should be adapting to improve computing efficiency or facilitate hardware maintenance.
- **Data Pre-Processing:** Because of the wide variety of data sources, the collected datasets vary with respect to noise, redundancy, and consistency, etc., and it is certainly a waste to store unuseful data. In addition, some analytical methods have important requirements on data quality. Therefore, in order to enable effective data analysis, we shall pre-process data under many conditions to integrate the data from different sources, which can not only reduce storage cost, but also improve analysis accuracy.

10. RELATIONSHIP BETWEEN IOT AND BIG DATA

The big data generated by IoT (internet of things) has different characteristics compared with general big data because of the different types of data collected, of which the most classical characteristics include heterogeneity, variety, unstructured feature, noise, and high redundancy. A report from Intel pointed out that big data in IoT has three features: **(i)** much terminals generating masses of data, **(ii)** data generated by IoT is usually semi-structured or unstructured, **(iii)** data of IoT is useful only when it is analyzed [9].

11. BIG DATA QUALITY

Data quality is not necessarily data that is devoid of errors. Incorrect data is only one part of the data quality equation. Most experts take a broader perspective.
"consistently meeting knowledge worker and end-customer expectations". Others say data quality is the fitness or suitability of data to meet business requirements[10]. High data quality is:

- **Complete:** All relevant data such as Purchases, addresses and relationships for a given customer is linked.
- **Accurate:** Common data problems like misspellings, typos, and random abbreviations have been cleaned up.
- **Available:** Required data are accessible on request, users do not need to search manually for the information.
- **Timely:** updated information is readily available to support decisions

12. BIG DATA CHALLENGES

- **Data Collection and Storage:** Data sets are growing in size since billion of data is created every day. There is a big storage requirement of databases, huge information storage and to store large output files. So, the major challenge is the requirement of more storage mediums and comparatively higher input output speed[8].
- **Data Inconsistence and Incompleteness:** Big data is a huge repository of various types of data sets consisting of structured, semi-structured and unstructured data. The most important challenge is that data should be complete and consistence from every view of accessibility.
- **Timeliness:** Timeliness means data should be available at the right time. The major challenge is to implement optimizing data access techniques, schema free databases to quickly modify the structure of data so that it do not need to rewrite tables.
- **Data Analysis:** Data analysis is a key challenge of big data. Various traditional techniques are invented but they cannot extract the core information from big data. To capture the useful patterns of big data, we need to develop modern techniques.
- **Data Visualization:** The main motive of visualization is to discover the hidden and complex knowledge in an understandable and interactive form. Current big data tools and techniques have poor response regarding data visualization.
- **Data Security:** Data security has great interest in field of information technology. Big data security is also a key

challenge. The size of big data is very large and it is stored in different data groups on remote computers.

CONCLUSION

Big data is data that exceeds the processing capacity of traditional database systems. In this paper I review the whole surroundings of Big data. it encompasses: definition , characteristics , architecture, classifications, difference between big data and traditional data, life cycle, big data analytics, data acquisition elements, data techniques elements, data quality characteristics, determine the relationship between IoT and big data, also the challenges that facing big data.
With Big data technologies, I will hopefully be able to provide better and accurate information to better understand of this subject.

REFERENCES

[1] A. Banik, S. Bandyopadhyay, " *Big Data- A Review on Analyzing 3Vs*", Journal of Scientific and Engineering Research (JSER), 2016
[2] R. Patil, O. Jadhav, "*Some Contribution of Statistical Techniques in Big Data: A Review*", International Journal on Recent and Innovation Trends in Computing and Communication (IJRITCC), 2016
[3] D. Singh, G. Singh, " *Big data – A Review*", International Research Journal of Engineering and Technology (IRJET), 2017
[4] M. Kataria, P. Mittal**,** " *BIG DATA: A Review*", International Journal of Computer Science and Mobile Computing, 2014
[5] E. Ramirez, J. Brill, M. Ohlhausen, T. McSweeny, Big Data A Tool for Inclusion or Exclusion?, Federal Trade Commission, 2016
[6] E. Sirin, H. Karacan, " *A Review on Business Intelligence and Big Data"* , International Journal of Intelligent Systems and Applications in Engineering (IJISAE), 2017
[7] X. Su, Introduction to Big Data, Kunnskap for en bedre verden NTNU, 2017
[8] B. Singh, S. Kumar, G. Kaur, M. Kaur, " *A Survey on Big Data: Challenges*, *Tools and Technique***"**, International Journal of Advanced Research in Computer Science (IJARCS), 2016
[9] M. Chen, S. Mao, Y. Liu , " *Big Data: A Survey*" , Mobile Networks and Applications , 2014
[10] N. Abdullah, S. Ismail , S. Sophiayati, " *Data Quality in Big Data: A Review"* , International Journal of Advances in Soft Computing and its Applications, 2015