



Significance of Big Data Analytics

Priyanka Jain¹, Sanjay Ojha²

^{1,2}School of IT, Centre for Development of Advanced Computing (CDAC),
B-30, Institutional Area, Sector 62, NOIDA-201307, Uttar Pradesh, INDIA

Abstract: With the development companies are realizing the importance of having more data in order to support decision for their strategies. The explosion of information systems, internet and web based applications, social networks, and new technologies have given rise to large worlds of information, often described as "Big Data." Big Data provides great promises for discovering small population patterns that are not possible with data that are small scale. Big Data's massive sample size and its dimensionality introduce unique computational and statistical challenges, measurement errors. This paper provides meaning, key technologies dealing with big data, uses in different area, issues and challenges in big data analytics.

Keywords: Big Data, Characteristics, Techniques, Tools, Architecture, Sectors, Issues

I. Introduction

With the high-speed development of the Internet and the informatization in various industries, the amount of data gathered in the society is at an explosive growth, and the whole society has realized the wave of big data has been rolling in [2]. But as the internet use increased at global and larger level, data went so big that it became difficult and complex to process them using data traditional data processing applications. This large data was termed as big data [1]. The concept of big data has been regional within computer science since the earliest days of computing. "Big Data" originally indicates the amount of data that could not be processed by traditional database methods and tools. [6]. Also SAP HANA can accelerate Big Data as SAP HANA is able to deliver an in-memory, read/write database that can simultaneous analyse data and support transactions-oriented applications [3] [4].

The original definition focused on structured data, but many researchers have realized that most of the world's information resides in massive, unstructured information in the form of text and imagery. We define "Big Data" as the amount of data just beyond technology's capability to store, manage and process efficiently [6].

II. Characteristics

Big Data embodies new data characteristics created by our digitized world:

1. **Volume: data at scale** - size from TB to PB and more. Too much volume is a storage issue, but too much data is also a Big Analytics issue [6].
2. **Velocity: data in motion** – analysis of streaming data, creating structured record, and availability for delivery and access. Velocity means both how quickly data is being produced and how quickly the data must be processed to meet demand [6].
3. **Variety: data in many formats/media types** – structured, unstructured, semi-structured, text, media.
4. **Veracity: uncertainty/quality** – managing the reliability and predictability of inherently imprecise data [6].

The first three V's have been introduced in (Gartner, 2011), in blog of Dwaine Snow's thoughts on Databases and Data Management, the V from Veracity has been added by Dwaine Snow [7] [12] [15]. IBM mentions the same step also in 2012 [7] [10] [14]. For Veracity of the data both Variety and Velocity are actually working against it [7].

5. **Visualization:** For insights of decision-making visual representations are needed (e.g., tag clouds, cluster grams, history flows, spatial information flows [13]) [7].

III. Analysing Big Data: Techniques

There are several techniques you should consider for this new type of analysis:

- **Discovery** – To figure out how different data sets relate to each other use process of exploration and discovery [8].
- **Iteration** – Sometimes iteration leads you down a path that can be a dead end. But experimentation is process's part. Many analysts and industry experts suggest that you start with small, well-defined projects and move on to the next idea or field of inquiry [8].
- **Flexible Capacity** – It means to be prepared for spending more time and utilizing more resources to solve problems [8].

- Mining and Predicting – Sometimes it is difficult to find out how the various data elements relate to each other. As mining the data starts to find out patterns and relationships, predictive analytics can give the insights [8].

IV. Tools Analysing Big Data

Five tools for analysing big data and generating insight:

- **Discovery tools:** Are useful for information lifecycle that provide rapid, intuitive exploration and analysis of information from any combination of sources that are structures or unstructured. These tools permit analysis alongside traditional BI source systems. Using these tool users can draw new insights, come to conclusions that are, and take informed decisions quickly [8].
- **BI tools:** Are used for generating reports, analysis and performance management. BI Tools provide comprehensive capabilities for BI and performance management which includes enterprise reporting, dashboards, ad-hoc analysis, scorecards [8].
- **In-Database Analytics:** It has techniques for searching patterns and relationships exist in data. As these techniques are applied directly in the database, it can eliminate data movement to and from analytical servers which increase information cycle times and decreases total cost of ownership [8].
- **Hadoop:** It is useful for pre-processing data that identity macro trends or used to find nuggets of information. It helps businesses to find out potential value from new data by having commodity servers which are inexpensive. Many organizations use Hadoop as a precursor to advanced forms of analytics [8].
- **Decision Management:** It includes predictive modelling, business rules, and self-learning to take informed action. This analysis provides individual recommendations across channels which are multiple, maximizing the value of every customer interaction. Oracle Advanced Analytics scores can be integrated to operationalize complex predictive analytic models and create real-time decision processes [8].

V. Conceptual Architecture

Framework for a big data analytics project is similar to that for a traditional business intelligence or analytics project. The main difference is how the processing is executed. In a regular analytics project, the analysis can be performed with the tools installed on a stand-alone system like desktop or laptop. For executing across multiple nodes processing is broken down. The availability of open-source platforms such as Hadoop/MapReduce on the cloud provides the application of big data analytics in many domains. Classical business analytics tools have become very user-friendly and transparent. [5]

As Figure 1 indicates, a primary component is the data. Sources of data are internal and external, most of times in multiple formats, residing at multiple locations. All data should be pooled together for analytics purposes. The data is still in a raw state so it should be transformed. For this several options are present. Service-oriented approach can be combined with web services. The data is in the same state and services are processed to call then retrieve and process. Data warehousing is other approach wherein all the data from the different sources are gathered and then ready for processing. As data is unavailable in real time through the steps of extract then transforming and load the data from diverse sources is cleansed and made ready. Several data formats are input to Hadoop/MapReduce platform whether the data is structured or unstructured .[5]

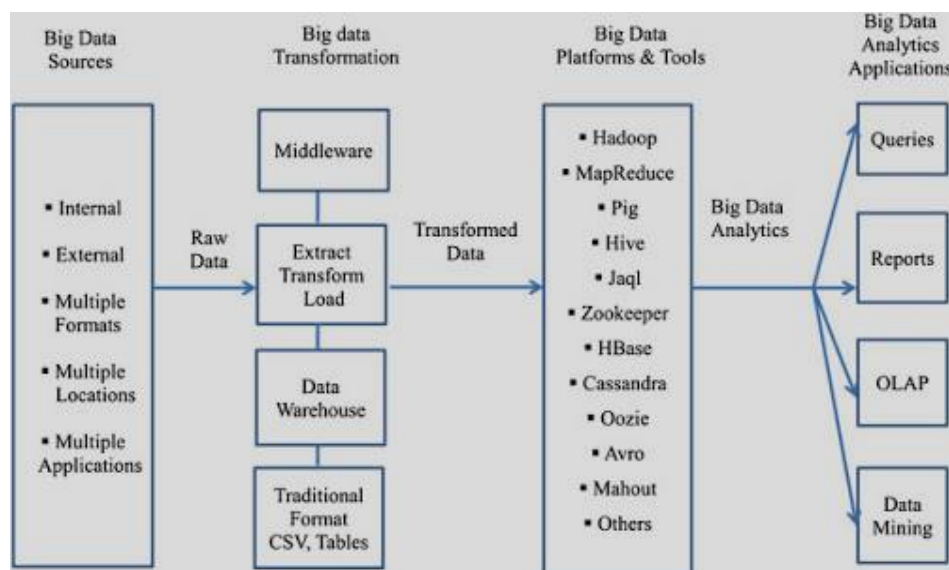


Fig. 1[5]: Conceptual Architecture

In this next stage several decisions are made like the data input approach, distributed design, selection of tool. Now, to the far right big data analytics applications are depicted. These include generation of queries, reports, online analytic processing and data mining. For aggregation, manipulation, analysis, and visualization of big data number of techniques and technologies are developed and adapted [5].

VI. Sectors in Big Data

1. Retail & E-Commerce

Retail is one of the high potential areas for big data. Companies are providing both online and offline data along with transactions information to understand factors that are driving the shoppers' behavioural traits. Many ONLINE BUSINESSES use algorithms which are advanced and allowing to have automatic recommendations based on purchase history data for customers [9].

2. Finance Services

Banking and financial management firms are rife with data that is transacted, having hundreds of millions records that are generated on a daily basis. With improvement in digital world, a variety of data sources – social media and web applications are increasing information to industry's existing data. Many of firms have challenges in terms of putting big data to work. As it is time consuming to set up and manage a cost effective analytics platform and also firms lack the talent pool of people for executing complex solution implementation. Even though some of the cloud services are providing on-demand big data infrastructure [9].

3. Telecommunications

Similar to other sectors, communication service providers all over the world are seeing huge data growth due to increased usage of smart phones, social media and having mobile networks. Many of firms are facing big data challenges for gaining more market share and increase profit margins. Big data can provide some of the key business objectives – better services for customers by having internal and external data, implementing innovative product services using segmentation techniques for generating new sources of revenue. Not only academia and industry are putting efforts for creating big data ready workforce but awareness is also there among young professionals that Big Data is one of the popular careers today [9].

4. Healthcare

By digitizing, combining and effectively using Big Data, health care organisations from single physician offices and large hospital networks realize many benefits. Potential advantages include detection of diseases at early stages where easily and effectively they can be treated; managing health of population, specific individual and detecting health care more quickly and efficiently [5].

VII. Big Data Issues

1. Storage and Transport Issue

The quantity of data has increased each time new storage medium is invented. What is different about the most recent increase – due largely to social media – is that no new storage medium has been there. Current disk technology has limits of about 4 terabytes per disk. So, 1 exabyte would require 25,000 disks. This means that if an exabyte of data could be processed on a single computer system, it would be not possible to directly attach the required number of disks. Access to data would overwhelm current communication networks [6].

2. Quality vs Quantity

An emerging challenge for big data users is “quantity vs. quality”. As users acquire and have access to more data and even more they want to access. For some users, the acquisition of data has become an addiction. Users believe that with more data, they can explain in which phenomenon they are interested in. A big data user have focus on quality which means not having all the data available, but having huge quantity of high quality data that can be used to draw conclusions that have high-value[6].

3. Data Ownership

Data ownership presents a critical and ongoing challenge, particularly in the social media. Petabytes of social media data that reside on the servers of Facebook, MySpace, and Twitter, are not really owned by users. Certainly, the “owners” of the pages or accounts believe that they own the data. With the advent of numerous social media sites, big data analytic includes mixing of first-party, reasonably verified data, with public and external data of third party, that is largely not been validated and verified by any formal methodology. Unverified data and compromises the fidelity of the dataset may introduce non-relevant entities and can give rise to erroneous linkages between entities. As a result, the accurate conclusions drawn from processing mixed data varies widely [6].

4. Timeliness

Larger the data set to be processed, the longer it will take to analyse. System design that effectively deals with size is also result in system that is processing size of data set faster. It is not just speed that is usually meant when one speaks of speed in the reference of Big Data. There are many situations in which the result of the analysis is required immediately. With new analyses desired using Big Data, there are new types of criteria specified, and a need to devise new index structures to support difficult criteria [11].

5. Human Collaboration

Big Data analysis system support input from many human experts and shared the results. When it is too expensive to assemble an entire team together in one room, multiple experts may be separated in space and time. System has to accept distributed expert input, and support their collaboration. However, there are individuals who have other motives and abilities – some may have a reason to provide false information in an intentional attempt to mislead. Framework is needed to use in analysis of such crowd-sourced data with conflicting statements. As humans, we can look at reviews of a restaurant, some of which are positive and negative, and come up with a process having summary assessment based on which we can decide whether to try eating there. We need computers to be able to do the equivalent [11].

VIII. Conclusion

We have entered in the world of Big Data. Big Data is “new era” of business. Information and knowledge that can be extracted from the digital universe is continuing to grow as users come up with new ways to the process data. Through better analysis and interpretation of the large volumes of data, there is a way for making faster development in many scientific disciplines and improving the profitability and success of many enterprises. The beginning of the “Big Data Revolution” brings us to a game changing and unique moment in history of data analysis that eclipses the traditions of the past. This revolution represents steps taking forward and a clear opportunity to realize enormous gains in efficiency, productivity, revenue, and profitability. The Era of Big Data Analytics is here.

IX. References

- [1] Vibha Bhardwaj (USICT GGSIPU, Dwarka, Delhi, India) vibhabhdwj@gmail.com, Rahul Johari (USICT GGSIPU, Dwarka, Delhi, India) rahul@ipu.ac.in, “Big Data Analysis: Issues and Challenge”, IEEE 2015.
- [2] Shangyi Chen (University of Chinese Academy of Sciences), Wei Li (Beihang University), Min Li (Baidu, Inc.), Xiaofei Zhang (Baidu, Inc.), Yue Min (Baidu, Inc.), “Latest Progress and Infrastructure Innovations of Big Data Technology”, 2014 International Conference on Cloud Computing and Big Data.
- [3] International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS) “Performance Analysis and Architectural Overview of SAP HANA”, Sanjay Ojha, Ayushi Chhabra, Shwetank Sharma, Centre for Development of Advanced Computing, Noida, India 2014.
- [4] International Journal of Software and Web Sciences (IJSWS) www.iasir.net, “FEED THE HUNGRY MOBILE APPLICATION USING SAP HANA AND CLOUDSHARE”, Sanjay Ojha, Nidhi Mathur, Anchal Singhal, School of Management, Centre for Development of Advanced Computing (CDAC), B-30, Sector-62, Institution Area, Noida-201307, INDIA 2014.
- [5] Wullianallur Raghupathi and Viju Raghupathi, “Big Data Analytics in healthcare: promise and potential”, health information science and systems 2014.
- [6] Stephen Kaisleri_SW Corporation skaisler1@comcast.net, Frank Armour American University fjarmour@gmail.com J. Alberto Espinosa American University, alberto@american.edu, William Money, George Washington University wmoney@gwu.edu, “Big Data: Issues and Challenges Moving Forward”, 2013 46th Hawaii International Conference on System Sciences.
- [7] Jaroslav Pokorný, “NEW DATABASE ARCHITECTURES: STEPS TOWARDS BIG DATA PROCESSING”, IADIS European Conference Data Mining 2013.
- [8] An Oracle White Paper March 2013, “Big Data Analytics, Advanced Analytics in Oracle Database”.
- [9] http://analyticstraining.com/2013/the-importance-of-big-data-training-and-the-hottest-sectors-using-big-data/
- [10] IBM Research, 2013. The Square Kilometer Array: The Ultimate Big Data Challenge. White Paper, IBM.
- [11] Divyakant Agrawal, UC Santa Barbara Philip Bernstein, Microsoft Elisa Bertino, Purdue Univ. Susan Davidson, Univ. of Pennsylvania Umeshwar Dayal, HP Michael Franklin, UC Berkeley Johannes Gehrke, Cornell Univ. Laura Haas, IBM Alon Halevy, Google Jiawei Han, UIUC H. V. Jagadish, Univ. of Michigan (Coordinator) Alexandros Labrinidis, Pittsburgh Univ. Sam Madden, MIT Yannis Papakonstantinou, UC San Diego Jignesh M. Patel, Univ. of Wisconsin Raghu Ramakrishnan, Yahoo! Kenneth Ross, Columbia Univ. Cyrus Shahabi, Univ. of Southern California Dan Suciu, Univ. of Washington Shiv Vaithyanathan, IBM Jennifer Widom, Stanford Univ, “Challenges and Opportunities with Big Data”, A community white paper developed by leading researchers across the United States 2012
- [12] Gartner, 2011. Pattern-Based Strategy: “Getting Value from Big Data”. Gartner Group http://www.gartner.com/it/page.jsp?id=1731916
- [13] Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C. and Byers, A.-H., 2011. “Big data: the next frontier for innovation, competition, and productivity”, McKinsey Global Inst.
- [14] Härder, T., 2005. “DBMS Architecture – the Layer Model and its Evolution”. In Datenbank-Spektrum 13, pp. 45-57.
- [15] Härder, T., Reuter, A., 1983. Concepts for Implementing and Centralized Database Management System. Proceedings of Int. Computing Symposium on Application Systems Development, Nürnberg, Germany, B.G., pp. 28-104.

Acknowledgements

It is a well-known fact that no task is achieved independently. It is always a collective task that is rewarded. Hence, in the present task too there are a number of people who are responsible in bringing out this paper in its current form. I would first of all like to extend my gratitude the entire management at CDAC Noida for supporting and providing me with the opportunity to utilize such a platform. I would also like to thank all the respective faculty members for their constant support and guidance.