An Overview of Big Data Characteristics and Enhancement of Big Data Application

[1]G.Venkatesh, [2]Dr. K. Arunesh
[1,2]Dept. of Computer Science, Sri S Ramasamy Naidu Memorial College
(Affiliated to Madurai Kamarajar University), Sattur, Virudhunagar District, Tamil Nadu, India

Abstract

Big Data has totally changed and revolutionized the businesses and organizations work. Here the major Big Data applications in various sectors and industries are given. In this era where every aspect of our day to day life has been technologized, there is a huge volume of data that has been emanating from various digital sources. We now feel it a necessity to have a tool to have this data in a systematic manner for applications in various fields including government, scientific research, industry, etc. This will help in a proper study, storage and processing of the same.in the traditional data processing tools there were a lot of challenges in the analysis and study of such a huge volume of data. To overcome these challenges, some big data solutions were introduced such as Hadoop. These big data tools really helped realize the applications of big data. Businesses are finding benefits which help them grow fast.

Keywords

Big Data, Industry, Application, Business, Characteristics, Data, Volume, Velocity.

I. Introduction

A Data is factual information used as a basis for reasoning, discussion and calculation. The Information output by a sensing device or organ that includes both useful and irrelevant or redundant information which is to be processed significantly in numerical form that can be digitally transmitted. Big data [11] is a term for large and complex unprocessed data (difficult and also time-consuming to process using the traditional processing methodologies). It can be characterized based on the volume, variety, velocity, variability, veracity and complexity.

In the attributes of big data, data is massive, comes at a speed and highly unstructured [3] that it doesn't fit conventional relational data base structures. With somuch insight hidden in this data, an alternative way to process this enormous data is necessary. Big corporations could be well resourced to handle this task but the amount of data being generated every day easily out grows this capacity. Cheaper hardware, [4] cloud computing and open source technologies have enabled processing big data at a much cheaper cost. Lot of data means lot of hidden insights. The ability to quickly analyze big data means the possibility to learn about customers, market trends, marketing and advertising drives, equipment monitoring and performance analysis and much more. It is an important reason that many big enterprises are in a need of robust big data analytics tools and technologies.

Big data tools mainly make use of in-memory data query principle. Queries are performed where the data is stored, unlike conventional Business Intelligence (BI) software that runs queries against data stored on server hard drive. In-memory data analytics has significantly improved data query performance. Big data [15] analytics not just helps enterprises make better decisions and gain an edge into real-time processing it has also inspired businesses to derive new metrics and gain new sources of revenue out of insights gained. Note that temporal data naturally leads to Big Data, as does spatial data. Early attempts to deal with large ware houses, including non-scalar data used so called ORDBMS i.e. object relations databases. Big Data out performs ORDBMS in various ways, including the need for more complicated backups, recovery and faster search algorithms, beyond RDBMS indexes.

`Industry influencers, academicians and other prominent stake holders certainly agree that big data has become a big game changer in most, if not all, types of modern industries over the last few years. As big data continues to permeate our day-to-day lives, there has been a significant shift of focus from the hype surrounding it to finding real value in its use. Generally, most organizations have several goals for adopting big data [9] projects. While the primary goal for most organizations is to enhance the customer experience, other goals include cost reduction, better targeted marketing and making existing processes more efficient. In recent times, data breaches have also made enhanced security an important goal that big data projects seek to incorporate.

II. Characteristics of Big Data

One can observe that several definitions coexist, accepted and used depending on the community. However in order to avoid potential confusions and to characterize better the concept of Big Data, a set of attributes were identified as defining characteristics. This set of attributes is referred to as the V's of Big Data, according to their common name initial. Considering that the Big Data ecosystem is highly dynamic, the set is expanding to include new V's that are continuously identified. Next we present the original set of the 3 V's as introduced and 2 additional ones which are widely accepted and used.

A. Volume

It represents perhaps the main feature that one associates with the concept of Big Data. This association with the magnitude of the data set arises naturally as all domains tend currently to collect and store massive amounts of data. This behavior is encouraged both by the low costs to store data and because having models which result from large data set tends to provide more accurate results from the data analytics point of view. In fact the size dimension of Big Data [13] represents the primary challenge to the existing data management systems. Accommodating the growing volumes calls for scalable storage solutions and distributed processing engines. Furthermore the size of the data can become big enough such that it has to be stored across multiple data centers, which requires high-performance solutions capable to operate in a geographically distributed manner. Processing across geographically distributed data centers and inter-site data management become a necessity in the Big Data era, due to the challenges raised by the data volumes.

B. Velocity

The high rates at which data are collected by organizations or flows into the processing engines make data velocity to gain importance alongside with volume. The common terminology used for fast-moving data is "streaming data". Initially, the velocity-related challenges were restricted to specific segments of industry, but it becomes a problem of a much broader setting with the Internet of Things. Financial tickers, stock market analysis, monitoring systems of large web services, network of sensors for wide-areas or scientific observatories are all concerned with the speed at which data is collected, streamed and processed in real time. In fact, it is expected that in the next year's most part of Big Data will be collected in real-time, which means that the speed to collect data over-passes the rate to artificially-produce them. Real-time data processing (i.e., stream processing) is necessary due to various reasons. For example, keeping the storage requirements practical can require pre-processing to filter out the useless parts in scenarios with high data rates. Additionally, many scenarios require the information to be extracted from data immediately or within a maximal (short) delay. However, depending on the locations of the data sources which produce the data, this task of real-time processing can be particularly challenging. The data sources can be geographically distant from the stream processing engine, which adds also the problem of latency to the challenges related to the streaming rate. Hence supporting the paradigm shift brought by Big Data calls for data management solutions which consider and address not only the size aspects of data, but also the challenges raised by streaming all the more in geographically distributed setups.

C. Variety

Collecting data from a variety of sources leads to a high heterogeneity. In fact, dealing with Big Data sets most often implies handling data without a predefined relational structure. Therefore, curating the data before storing and processing them becomes a critical tasks and a challenge on its own. However, pre-processing and determining a relational scheme before storing it is a complex task considering the large volumes. In the few cases when this phase is possible, the scalability limits of traditional databases can still arise as a prohibiting factor to address the variety aspect of Big Data via relational schemes. As a consequence, the common approach is to handle data in an unstructured fashion, e.g., storing data in large binary objects. On the one hand, such an approach provides scaling and performance advantages. On the other hand it amplifies the variety problem of Big Data sets, asmost of the contextual and self-describing information is lost. As a result, the process of extracting (co)relations and ordering the data is coupled with the [2] data mining itself, sometimes becoming the computation itself. A key direction enabled by such joint efforts is trying to correlate the heterogeneous data sets of the distinct disciplines in order to discover new ways of explaining the life-or universe-related observations. Such ambitious goals call for efficient, large-scale tools which can handle thevariety aspects of Big Data.

D. Veracity

The trust worthiness of data impacts theirs value. Therefore, one of the newly identified challenges when dealing with Big Data is veracity, which generically synthesizes the correctness, quality and accuracy of the data. The concerns related to the veracity apply both to the input as well as to the result harvested when mining it. Veracity becomes an important aspect due to the diversity of the sources and forms that Big Data takes, which provides less control over its correctness. Malfunctioning sensors typos in social media feeds, colloquial discourses in news media, systematic errors and heterogeneity of measuring devices, all need to be accounted for during the analysis phase. Sometimes, the volume can compensate for the lack of accuracy. But tackling veracity via volume needs to instrument the analytic models properly, which is achieved most often at the expense of extra computation. Therefore, providing tools which are able to scale the computation in order to ensure high and customizable confidence levels for the analysis is critical for the development of Big Data business and data-intensive sciences.

E. Value

Collecting, storing and analyzing Big Data is useless unless it produces value. Therefore, this aspect goes alongside and determines any previous or future challenge of Big Data. It can be safely stated that "Value" is the ultimate goal of Big Data, being the driven factor of this technological revolution. Dealing with Big Data is a complex task and it involves significant costs. Therefore the benefits gained whether financial or scientific, must compensate the resources and efforts which are invested. This observation raises a very important aspect to be considered when dealing with Big Data: the efficiency trade-off between cost and performance. The new management solutions need to provide mechanisms for estimating both the cost and performance of operations (e.g., for streaming, storing, processing, etc.). Based on these estimations the users can then choose the cost they are willing to pay for a certain performance level. In turn, the management systems need to optimize their resource usage according to the specified criterion to meet the budget/performance constraints. Designing such customizable trade-off mechanisms is needed because the density of the value in the data sets is neither uniform nor identical among different applications. This shows that the value aspect of Big Data needs to be considered not only from the point of view of the worthiness (or profitability) of the analysis, but also as a designing principle of the management and processing frameworks.

III. Big Data Applications

Data is everywhere, it is the amount of digital data that exists is growing at a rapid rate, doubling every two years and changing the way we live. Nowadays Big data has found many applications in various fields. Some of the major fields where big data is being used are shown in fig. 1.

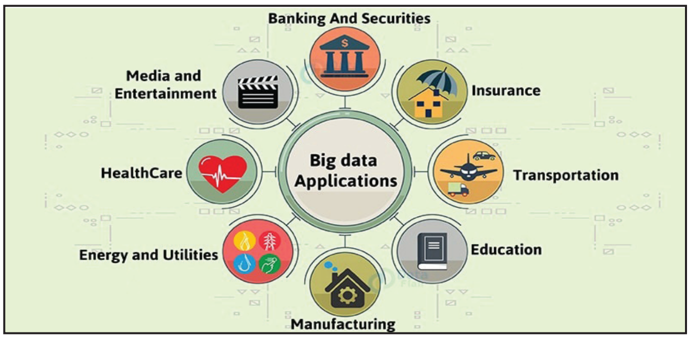


Fig. 1: Big Data in Industrial Applications

1. Banking and Securities Industry

In the banking sector proper study and analysis of the data can help detect any and all the illegal activities that are being carried out such as

- The misuse of credit cards
- Misuse of debit cards
- Venture credit hazard treatment
- Business clarity
- Customer statistics alteration
- Money laundering
- Risk Mitigation

The challenges in banking industry includes securities fraud early warning, tick analytics, card fraud detection, archival of audit trails, enterprise credit risk reporting, trade visibility, customer data transformation, social analytics for trading, IT operations analytics and IT policy compliance analytics. The Securities Exchange Commission (SEC) is using big data to monitor financial market activity. They are currently using network analytics and natural language processors to catch illegal trading activity in the financial markets.

2. Communications, Media and Entertainment Industry

With people having access to various digital gadgets the generation of large amount of data is inevitable and this is main cause of rise in big data in media and entertainment industry.Other than this, social media platforms are also another way in which huge amount of data is being generated. Although business in media and entertainment industry have realized the importance of the data and they have been able to leverage from it to help their businesses grow.Some of the benefits extracted from the big data in media and entertainment industry:

- Predicting the interests of audiences.
- Optimized or on-demand scheduling of media streams in digital media distribution platforms.
- Getting Insights into customer's reviews and pinpointing their animosities.
- Effective targeting of the advertisements for media

3. HealthCare Industry

Now healthcare is yet another industry which is bound to generate a huge amount of data. Following are some of the ways in which big data has contributed to healthcare

- Big data reduces costs of treatment since there is less chances of having to perform unnecessary diagnosis.
- It helps in predicting outbreaks of epidemics and also helps in deciding what preventive measures could be taken to minimize the effects of the same.
- It helps avoid preventable diseases by detecting diseases in early stages and prevents it from getting any worse which in turn makes the treatment easy and effective.
- Patients [1] can be provided with the evidence based medicine which is identified and prescribed after doing the research of past [3] medical results.

4. Energy and Utilities Industry

The Smart meter readers allow data to be collected almost every 15 minutes as opposed to once a day with the old meter readers. This granular data is being used to analyze consumption of utilities better which allows for improved customer feedback and better control of utilities use. In energy and utility companies the use of big data also allows for better asset and workforce management which is useful for recognizing errors and correcting them as soon as possible before complete failure is experienced.

5. Manufacturing and Natural Resources Industry

Increasing demand for natural resources including oil, agricultural products, minerals, gas, metals, and so on has led to an increase in the volume, complexity and velocity of data that is a challenge to handle. Similarly, large volumes of data from the manufacturing industry are untapped. The underutilization of this information prevents improved quality of products, energy efficiency, reliability and better profit margins.

In the natural resources industry, big data allows for predictive modeling to support decision making that has been utilized to ingest and integrate large amounts of data from geospatial data, graphical data, text and temporal data. Big data has also been used in solving today's manufacturing challenges and to gain competitive advantage among other benefits.

6. Education Industry

Education Industry is flooding with a huge amount of data related to students, faculties, courses, results and what not. It was not long before we realized that the proper study and analysis of this data can provide insights that can be used to improve the operational effectiveness and working of educational institutes. Following are some of the fields in education industry that have been transformed by big data motivated changes.

- **Customized and dynamic learning programs:** Customized programs and schemes for each individual can be created using the data collected on the bases of a student's learning history to benefit all students. This improves the overall student results
- **Reframing course material:** Reframing the course material according to the data that is collected on the basis of what student learns and to what extent by real time monitoring of what components of a course are easier to understand.
- **Grading Systems:** New advancements in grading systems have been introduced as a result of proper analysis of student data.
- **Career prediction:** Proper analysis and study of every student's records will help in understanding the student's progress, strengths, weaknesses, interests and more. It will help in determining which career would be most appropriate for the student in the future.

The applications [8] of big data have provided a solution to one of the biggest pitfalls in the education system, that is, the one-size-fits-all fashion of academic set up, by contributing in e-learning solutions.

7. Transportation Industry

Since the rise of big data, it has been used in various ways to make transportation more efficient and easy. Following are some of the areas where big data contributed to transportation.

- **Route planning:** Big data can be used to understand and estimate the user's needs on different routes and on multiple modes of transportation and then utilizing route planning to reduce the users wait times.
- **Congestion management and traffic control:** Using big data, real time estimation of congestion and traffic patterns [5] is now possible. For examples, people using Google Maps to locate the least traffic prone routes.
- **Safety level of traffic:** Using the real time processing of big data and predictive analysis to identify the traffic accidents prone areas can help reduce accidents and increase the safety level of traffic.

In recent times, huge amounts of data from location-based social networks and high speed data from telecoms have affected travel behavior. Regrettably, research to understand travel behavior has not progressed as quickly.

8. Insurance Industry

Big data has been used in the insuranceindustry to provide customer insights for transparent and simpler products, by analyzing and predicting the customer behavior through data derived from social media, GPS-enabled devices and CCTV footage. The big data also allows for better customer retention from insurance companies. When it comes to claims management, predictive analytics from big data has been used to offer faster service since massive amounts of data can be analyzed especially in the underwriting stage. Fraud detection has also been enhanced. Through massive data from digital channels and social media, real-time monitoring of claims throughout the claims cycle has been used to provide insights.

9. Agriculture Industry

A biotechnology firm uses sensor data to optimize crop efficiency. It plants test crops and runs simulations to measure how plants react to various changes in condition. Its data environment constantly adjusts to changes in the attributes of various data it collects, including temperature, water levels, soil composition, growth, output, and gene sequencing of each plant in the test bed. These simulations allow it to discover the optimal environmental conditions for specific gene types.

10. Marketing Industry

Marketers have begun to use facial recognition software to learn how well their advertising succeeds or fails at stimulating interest in their products. A recent study published in the Harvard Business Review looked at what kinds of advertisements compelled viewers to continue watching and what turned viewers off. Among their tools was “a system that analyses facial expressions to reveal what viewers are feeling”. The research was designed to discover what kinds of promotions induced watchers to share the ads with their social network, helping marketers create ads most likely to “go viral” and improve sales.

11. Smart Phones Industry

Perhaps more impressive, people now carry facial recognition technology in their pockets. Users of I Phone and Android smartphones have applications at their fingertips that use facial recognition technology for various tasks. For example, Android users with the remember app can snap a photo of someone and then bring up stored information about that person based on their image when their own memory lets them down a potential boon for salespeople.

12. Telecom Industry

In telecom also it plays a very good role. Operators face an uphill challenge when they need to deliver new, compelling, revenue-generating services without overloading their networks and keeping their running costs under control. The market demands new set of data management and analysis capabilities that can help service providers make accurate decisions by taking into account customer, network context and other critical aspects of their businesses. Most of these decisions must be made in real time, placing additional pressure on the operators. Real-time predictive analytics can help leverage the data that resides in their multitude systems, make it immediately accessible and help correlate that data to generate insight that can help them drive their business forward.

13. Weather patterns Industry

There are weather sensors and satellites deployed all around the globe. A huge amount of data is collected from them and then this data is used to monitor the weather and environmental conditions. All of the data collected from these sensors and satellites contributes to big data and can be used in different ways such as:

- In weather forecast.
- To study global warming.
- Understanding the patterns of natural disasters.
- To make necessary preparations in case of crisis.
- To predict the availability of usable water around the world.

IV. Conclusion

Applications of big data in real world are discussed. No wonder there is so much hype for big data, given all of its applications. The importance of big data lies in how an organization is using the collected data and not in how much data they have been able to collect. To use big data efficiently, there areBig Data solutions that make the analysis of Big Data easy. This is where the applications of Big Data start showing up, when big data solutions are used to gain benefits of big data. Big Data is a powerful tool that makes things ease in various fields as said above. Big data used in so many applications they are banking, agriculture, chemistry, data mining, cloud computing, finance, marketing, stocks, healthcare etc. an overview is presented especially to project the idea of Big Data.

References

[1] Shouman, M., Turner, T., Stocker, R.,"Applying k-Nearest Neighbor in Diagnosing Heart Disease Patients", International Conference on Knowledge Discovery, 2012.

[2] Candelieri, A., Dolce, G., Riganello, F., Sannita, W. G.,"Data Mining in Neurology", In Knowledge Oriented Applications in Data Mining, pp. 261-276, 2011.

[3] Bushinak, H., AbdelGaber, S., AlSharif, F. K.,"Recognizing the Electronic Medical Record Data from Unstructured Medical Data Using Visual Text Mining Techniques, 2011.

[4] R. Vrbić,“Data mining and cloud computing,” Journal of Information Technology & Applications, Vol. 2, No. 2, pp. 75-87, 2012.

[5] Sethukkarasi, R., Keerthika, U. Kannan, A.,"A self-Learning Rough Fuzzy Neural Network Classifier for Mining Temporal Patterns", Proceeding of International Conference on Advances in Computing, Communications and Informatics, 2012.

[6] Sau A, Bhakta I. Artificial Neural Network (ANN) model to predict depression among geriatric population at a slum in Kolkata, India. J ClinDiagn Res 2017 May; 11(5):VC01.

[7] Orabi AH, Buddhitha P, Orabi MH, Inkpen D.,"Deep learning for depression detection of twitter users", In Proceedings of the Fifth workshop on computational linguistics and clinical psychology: from keyboard to clinic. pp. 88–97, 2018.

[8] G. Dzemyda, O. Kurasova, J. Zilinskas,"Multidimensional Data Visualization: Methods and Applications", Springer Optimization and Its Applications, Springer, 2013.

[9] M Nagalakshmi, Dr. I Surya Prabha, K Anil,“Big Data Map Reducing Technique Based Apriori in Distributed Mining”, International Journal of Advanced Research in Engineering and Technology, 8(5), pp. 19 – 28, 2017.

[10] Rahmani, Mostafa, George Atia,"Randomized Robust Subspace Recovery and Outlier Detection for High Dimensional Data Matrices", IEEE Transactions on Signal Processing (2016).

[11] Peng, Sancheng, Guojun Wang, DongqingXie,"Social influence analysis in social networking big data: Opportunities and challenges." IEEE Network, 2016.

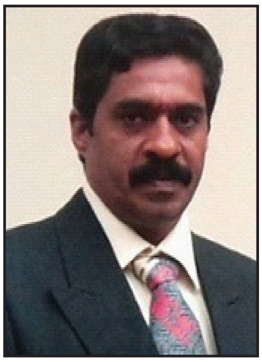
[12] Qiao, Yuanyuan, Yihang Cheng, Jie Yang, Jiajia Liu, Nei Kato,"A Mobility Analytical Framework for Big Mobile Data in Densely Populated Area", IEEE Transactions on Vehicular Technology, 2016.

[13] Lo'ai, A. Tawalbeh, Rashid Mehmood, ElhadjBenkhlifa, Houbing Song,"Mobile cloud computing model and big data analysis for healthcare applications", IEEE Access 4: pp. 6171-6180, 2016.

[14] Shi, Weiwei, Yongxin Zhu, S. Yu Philip, Tian Huang, Chang Wang, Yishu Mao, Yufeng Chen,"Temporal Dynamic Matrix Factorization for Missing Data Prediction in Large Scale Coevolving Time Series." IEEE Access 4, pp. 6719-6732, 2016.

[15] Dhanalakshmi, R., Mohamed Jakkariya, S., Mangaiarkarasi, S,"Aggregation methodology on map reduce for big data applications by using traffic-aware partition algorithm", Int. J. Innov. Res. Comput. Commun. Eng. 4(2), 2016.

G.Venkatesh has completed his under graduation B.C.A. from Urumu Dhanalakshmi College, Bharathidasan University, Trichy, his post-graduation M.C.A. from Urumu Dhanalakshmi College, Bharathidasan University, Trichy, B.Ed. from Oxford College of Education, Teacher Education University, Trichy and M.Ed. from Jeevan College of Education, Teacher Education University, Trichy, Second Class M.Phil. (Computer Science) from Jamal Mohamed College (Autonomous), Bharathidasan University, Trichy. He is presently research scholar at Sri S Ramasamy Naidu Memorial College (Affiliated to Madurai Kamarajar University), Sattur, Virudhunagar District. His current area of interest includes big data analysis and Image processing.



Dr. K. Arunesh has received Ph.D. degree in Computer Science from the Bharathidasan University, Tiruchirappalli, Tamil Nadu, and India. He is currently an Associate Professor of Computer Science and Dean, Academic Affairs at Sri. S.R.N.M College, Sattur. His current research interests include Knowledge Discovery in Databases, Big Data Analytics, Data Mining and Recommender Systems. He has published widely in leading journals and conference proceedings, and served program committees as conference Chairperson and Convener. He is currently on the editorial boards and serves as reviewer of leading journals and conferences.