**UNIT – I: Introduction to Big Data: Big Data and its Importance – Four V’s of Big Data – Drivers for Big Data – Introduction to Big Data Analytics – Big Data Analytics applications.**

1. **INTRODUCTION TO BIG DATA:**

Big Data is a collection of data that is huge in volume, yet growing exponentially with time. It is a data with so large size and complexity that none of traditional data management tools can store it or process it efficiently. Big data is also a data but with huge size.

***Example of Big Data:***

* **New York Stock Exchange** is an example of Big Data that generates about one terabyte of new trade data per day.
* **Social Media**

The statistic shows that 500+terabytes of new data get ingested into the databases of social media site Facebook, every day. This data is mainly generated in terms of photo and video uploads, message exchanges, putting comments etc.

* A **single Jet engine** can generate 10+terabytes of data in 30 minutes of flight time. With many thousand flights per day, generation of data reaches up to many Petabytes.

***Types of Big Data***

Following are the types of Big Data:

1. Structured
2. Unstructured
3. Semi-structured
4. ***Structured***

Any data that can be stored, accessed and processed in the form of fixed format is termed as a ‘structured’ data. Over the period of time, talent in computer science has achieved greater success in developing techniques for working with such kind of data (where the format is well known in advance) and also deriving value out of it.

*Examples of Structured Data:* An ‘Employee’ table in a database is an example of Structured Data

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Employee\_ID** | **Employee\_Name** | **Gender** | **Department** | **Salary\_In\_lacs** |
| 2365 | Rajesh Kulkarni | Male | Finance | 650000 |
| 3398 | Pratibha Joshi | Female | Admin | 650000 |
| 7465 | Shushil Roy | Male | Admin | 500000 |
| 7500 | Shubhojit Das | Male | Finance | 500000 |
| 7699 | Priya Sane | Female | Finance | 550000 |

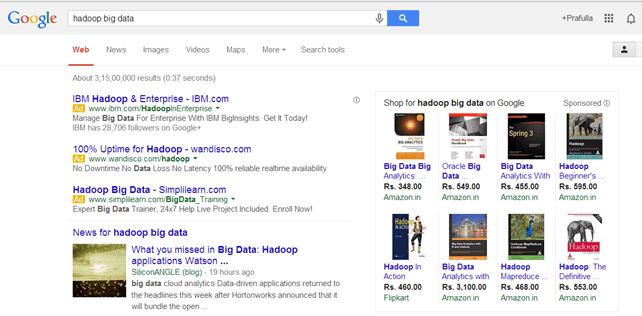
However, nowadays, we are foreseeing issues when a size of such data grows to a huge extent, typical sizes are being in the rage of multiple zettabytes. Looking at these figures one can easily understand why the name Big Data is given and imagine the challenges involved in its storage and processing.

1. ***Unstructured***

Any data with unknown form or the structure is classified as unstructured data. In addition to the size being huge, un-structured data poses multiple challenges in terms of its processing for deriving value out of it. A typical example of unstructured data is a heterogeneous data source containing a combination of simple text files, images, videos etc.

Now a days, organizations have wealth of data available with them but unfortunately, they don’t know how to derive value out of it since this data is in its raw form or unstructured format.

*Examples of Un-structured Data:* The output returned by ‘Google Search’



1. ***Semi-structured***

Semi-structured data can contain both the forms of data. We can see semi-structured data as a structured in form but it is actually not defined with e.g. a table definition in relational DBMS. Example of semi-structured data is a data represented in an XML file.

*Examples of Semi-structured Data:* Personal data stored in an XML file-

|  |
| --- |
| <rec><name>Prashant Rao</name><sex>Male</sex><age>35</age></rec>  <rec><name>Seema R.</name><sex>Female</sex><age>41</age></rec>  <rec><name>Satish Mane</name><sex>Male</sex><age>29</age></rec>  <rec><name>Subrato Roy</name><sex>Male</sex><age>26</age></rec>  <rec><name>Jeremiah J.</name><sex>Male</sex><age>35</age></rec> |

1. **BIG DATA AND ITS IMPORTANCE:**

Big Data initiatives were rated as “extremely important” to 93% of companies. Leveraging a Big Data analytics solution helps organizations to unlock the strategic values and take full advantage of their assets.

It helps organizations:

* To understand Where, When and Why their customers buy
* Protect the company’s client base with improved loyalty programs
* Seizing cross-selling and upselling opportunities
* Provide targeted promotional information
* Optimize Workforce planning and operations
* Improve inefficiencies in the company’s supply chain
* Predict market trends
* Predict future needs
* Make companies more innovative and competitive
* It helps companies to discover new sources of revenue

Companies are using Big Data to know what their customers want, who are their best customers, why people choose different products. The more a company knows about its customers, the more competitive it becomes.

We can use it with Machine Learning for creating market strategies based on predictions about customers. Leveraging big data makes companies customer-centric.

Companies can use Historical and real-time data to assess evolving consumers’ preferences. This consequently enables businesses to improve and update their marketing strategies which make companies more responsive to customer needs.

***Importance of Big data***

Big Data importance doesn’t revolve around the amount of data a company has. Its importance lies in the fact that how the company utilizes the gathered data. Every company uses its collected data in its own way. More effectively the company uses its data, more rapidly it grows. The companies in the present market need to collect it and analyze it because:

**1. Cost Savings:** Big Data tools like Apache Hadoop, Spark, etc. bring cost-saving benefits to businesses when they have to store large amounts of data. These tools help organizations in identifying more effective ways of doing business.

**2. Time-Saving:** Real-time in-memory analytics helps companies to collect data from various sources. Tools like Hadoop help them to analyze data immediately thus helping in making quick decisions based on the learning.

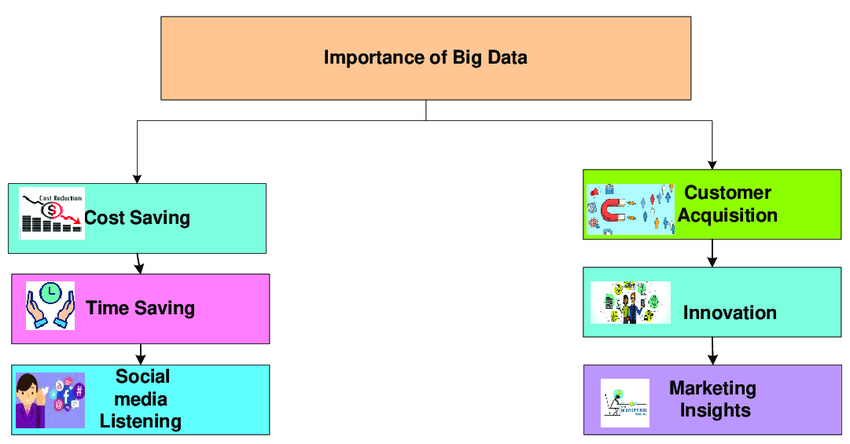
**3. Understand the market conditions:** Big Data analysis helps businesses to get a better understanding of market situations. For example, analysis of customer purchasing behavior helps companies to identify the products sold most and thus produces those products accordingly. This helps companies to get ahead of their competitors.

**4. Social Media Listening:** Companies can perform sentiment analysis using Big Data tools. These enable them to get feedback about their company, that is, who is saying what about the company. Companies can use Big data tools to improve their online presence.

**5. Boost Customer Acquisition and Retention:** Customers are a vital asset on which any business depends on. No single business can achieve its success without building a robust customer base. But even with a solid customer base, the companies can’t ignore the competition in the market. If we don’t know what our customers want then it will degrade companies’ success. It will result in the loss of clientele which creates an adverse effect on business growth. Big data analytics helps businesses to identify customer related trends and patterns. Customer behavior analysis leads to a profitable business.

**6. Solve Advertisers Problem and Offer Marketing Insights:** Big data analytics shapes all business operations. It enables companies to fulfill customer expectations. Big data analytics helps in changing the company’s product line. It ensures powerful marketing campaigns.

**7. The driver of Innovations and Product Development:** Big data makes companies capable to innovate and redevelop their products.

****

1. **FOUR V’S OF BIG DATA (CHARACTERISTICS):**

Big data can be described by the following characteristics:

1. Volume
2. Variety
3. Velocity
4. Variability

***(i) Volume –*** The name Big Data itself is related to a size which is enormous. Size of data plays a very crucial role in determining value out of data. Also, whether a particular data can actually be considered as a Big Data or not, is dependent upon the volume of data. Hence, **‘Volume’** is one characteristic which needs to be considered while dealing with Big Data solutions.

***(ii) Variety –*** The next aspect of Big Data is its **variety**. Variety refers to heterogeneous sources and the nature of data, both structured and unstructured. During earlier days, spreadsheets and databases were the only sources of data considered by most of the applications. Nowadays, data in the form of emails, photos, videos, monitoring devices, PDFs, audio, etc. are also being considered in the analysis applications. This variety of unstructured data poses certain issues for storage, mining and analyzing data.

***(iii) Velocity –*** The term **‘velocity’** refers to the speed of generation of data. How fast the data is generated and processed to meet the demands, determines real potential in the data. Big Data Velocity deals with the speed at which data flows in from sources like business processes, application logs, networks, and social media sites, sensors,[Mobile](https://www.guru99.com/mobile-testing.html)devices, etc. The flow of data is massive and continuous.

***(iv) Variability –*** This refers to the inconsistency which can be shown by the data at times, thus hampering the process of being able to handle and manage the data effectively.

***Advantages of Big Data Processing:***

Ability to process Big Data in DBMS brings in multiple benefits, such as-

1. Businesses can utilize outside intelligence while taking decisions

Access to social data from search engines and sites like facebook, twitter are enabling organizations to fine tune their business strategies.

1. Improved customer service

Traditional customer feedback systems are getting replaced by new systems designed with Big Data technologies. In these new systems, Big Data and natural language processing technologies are being used to read and evaluate consumer responses.

1. Early identification of risk to the product/services, if any.
2. Better operational efficiency

Big Data technologies can be used for creating a staging area or landing zone for new data before identifying what data should be moved to the data warehouse. In addition, such integration of Big Data technologies and data warehouse helps an organization to offload infrequently accessed data.

1. **DRIVERS FOR BIG DATA:**

Big Data has quickly risen to become one of the most desired topics in the industry. The main business drivers for such rising demand for Big Data Analytics are:

* 1. The digitization of society
  2. The drop in technology costs
  3. Connectivity through cloud computing
  4. Increased knowledge about data science
  5. Social media applications
  6. The rise of Internet-of-Things(IoT)

Example: A number of companies that have Big Data at the core of their strategy like: Apple, Amazon, Facebook and Netflix have become very successful at the beginning of the 21st century.

1. **INTRODUCTION TO BIG DATA ANALYTICS:**

Big Data analytics is a process used to extract meaningful insights, such as hidden patterns, unknown correlations, market trends, and customer preferences. Big Data analytics provides various advantages—it can be used for better decision making, preventing fraudulent activities, among other things.

***Importance:***

In today’s world, Big Data analytics is fueling everything we do online—in every industry.

Take the music streaming platform **Spotify**, for example. The company has nearly 96 million users that generate a tremendous amount of data every day. Through this information, the cloud-based platform automatically generates suggested songs—through a smart recommendation engine—based on likes, shares, search history, and more. What enables this is the techniques, tools, and frameworks that are a result of Big Data analytics.

If you are a Spotify user, then you must have come across the top recommendation section, which is based on your likes, past history, and other things. Utilizing a recommendation engine that leverages data filtering tools that collect data and then filter it using algorithms works. This is what Spotify does.

***Uses and Examples of Big Data Analytics:***

There are many different ways that Big Data analytics can be used in order to improve businesses and organizations. Here are some examples:

* Using analytics to understand customer behavior in order to optimize the customer experience
* Predicting future trends in order to make better business decisions
* Improving marketing campaigns by understanding what works and what doesn't
* Increasing operational efficiency by understanding where bottlenecks are and how to fix them
* Detecting fraud and other forms of misuse sooner

***Advantages of Big Data Analytics:***

**1. Risk Management:**

Use Case: Banco de Oro, a Phillippine banking company, uses Big Data analytics to identify fraudulent activities and discrepancies. The organization leverages it to narrow down a list of suspects or root causes of problems.

**2. Product Development and Innovations**

Use Case: Rolls-Royce, one of the largest manufacturers of jet engines for airlines and armed forces across the globe, uses Big Data analytics to analyze how efficient the engine designs are and if there is any need for improvements.

**3. Quicker and Better Decision Making Within Organizations**

Use Case: Starbucks uses Big Data analytics to make strategic decisions. For example, the company leverages it to decide if a particular location would be suitable for a new outlet or not. They will analyze several different factors, such as population, demographics, accessibility of the location, and more.

**4. Improve Customer Experience**

Use Case: Delta Air Lines uses Big Data analysis to improve customer experiences. They monitor tweets to find out their customers’ experience regarding their journeys, delays, and so on. The airline identifies negative tweets and does what’s necessary to remedy the situation. By publicly addressing these issues and offering solutions, it helps the airline build good customer relations.

1. **BIG DATA ANALYTICS APPLICATIONS:**

Here are some examples of the applications of big data analytics:

* ***Customer Acquisition and Retention:*** Customer information helps tremendously in marketing trends, through data-driven actions, to increase customer satisfaction. For example, personalization engines for Netflix, Amazon, and Spotify help with improved customer experiences and gaining customer loyalty.
* ***Targeted Ads:*** Personalized data about interaction patterns, order history, and product page viewing history can help immensely to create targeted ad campaigns for customers on a larger scale and at the individual level.
* ***Product Development:*** It can generate insights on development decisions, product viability, performance measurements, etc., and direct improvements that positively serve the customers.
* ***Price Optimization:*** Pricing models can be modeled and used by retailers with the help of diverse data sources to maximize revenues.
* ***Supply Chain and Channel Analytics:*** Predictive analytical models help with B2B supplier networks, preemptive replenishment, route optimizations, inventory management, and notification of potential delays in deliveries.
* ***Risk Management:*** It helps in the identification of new risks with the help of data patterns for the purpose of developing effective risk management strategies.
* ***Improved Decision-making:*** The insights that are extracted from the data can help enterprises make sound and quick decisions

**Big Data Analytics Implementation in Major Sectors:**

* ***Retail:*** The retail industry is actively deploying big data analytics. It is applying the techniques of data analytics to understand what the customers are buying and then offering products and services that are tailor-made for them. Today, it is all about having an omnichannel experience. Customers may make contact with a brand on one channel and then finally buy the product(s) through another channel, meanwhile going through more intermediary channels. The retailers will have to keep track of these customer journeys, and they must deploy their marketing and advertising campaigns based on that, to improve the chances of increasing sales and lowering costs.
* ***Technology:*** Technology companies are heavily deploying big data analytics. They are finding out more about how customers interact with websites or apps and gather key information. Based on this, technology companies can optimize their sales, customer service, customer satisfaction, etc. This also helps them launch new products and services since we are living in a knowledge-intensive economy, and the companies in the technology sector are reaping the benefits of big data analytics.
* ***Healthcare:*** Healthcare is another industry that can benefit from big data analytics tools, techniques, and processes. Healthcare personnel can diagnose the health of their patients through various tests, run them through the computers, and look for telltale signs of anomalies, maladies, etc. It also helps in healthcare to improve patient care and increase the efficiency of the treatment and medication processes. Some diseases can be diagnosed before their onset so that measures can be taken in a preventive manner rather than a remedial manner.
* ***Manufacturing:*** Manufacturing is an industrial sector that is involved with developing physical goods. The life cycle of a manufacturing process can vary from product to product. Manufacturing systems are involved within the industry setup and across the manufacturing floor. There are a lot of technologies that are involved in manufacturing such as the Internet of Things (IoT), robotics, etc., but the backbone of all of these is firmly based on big data analytics. By using this, manufacturers can improve their yield, reduce the time to market, enhance the quality, optimize the supply chain and logistics processes, and build prototypes before the launch of products. It can help manufacturers through all these steps.
* ***Energy:*** Most oil and gas companies, which come under the energy sector, are extensive users of big data analytics. It is deployed when it comes to discovering oil and other natural resources. Tremendous amounts of big data go into finding out what the price of a barrel of oil will be, what the output should be, and if an oil well will be profitable or not. It is also deployed in finding out equipment failures, deploying predictive maintenance, and optimally using resources in order to reduce capital expenditure.



***Big Data Analytics Tools:***

1. ***Apache Spark:*** Spark is a framework for real-time data analytics, which is a part of the Hadoop ecosystem.
2. ***Python***: Python is one of the most versatile programming languages that is rapidly being deployed for various applications including machine learning.
3. ***SAS***: SAS is an advanced analytical tool that is used for working with large volumes of data and deriving valuable insights from it.
4. ***Hadoop***: Hadoop is the most popular big data framework that is deployed by a wide range of organizations from around the world for making sense of big data.
5. ***SQL***: SQL is used for working with relational database management systems.
6. ***Tableau***: Tableau is the most popular business intelligence tool that is deployed for the purpose of data visualization and business analytics.
7. ***Splunk***: Splunk is the tool of choice for parsing machine-generated data and deriving valuable business insights out of it.
8. ***R:*** R is the no. 1 programming language that is being used by data scientists for statistical computing and graphical applications alike.

***Big Data Analytics Challenges:*** Big data analytics does not just come with wide-reaching benefits, it also comes with its own challenges:

1. ***Accessibility of Data:*** With larger volumes of data, storage and processing become a challenge. Big data should be maintained in such a way that it can be used by less-experienced data scientists and data analysts as well.
2. ***Data Quality Maintenance:*** With high volumes of data from disparate sources and in different formats, the proper management of data quality requires considerable time, effort, and resources.
3. ***Data Security:*** The complexity of big data systems poses unique challenges when it comes to security. It can be a complex undertaking to properly address such security concerns within complicated big data ecosystems.
4. ***Choosing the Right Tools:*** Choosing big data analytics tools from the wide range that is available in the market can be quite confusing. One should know how to select the best tool that aligns with user requirements and organizational infrastructure.
5. ***Supply-demand Gap in Skills:*** With a lack of data analytics skills in addition to the high cost of hiring experienced professionals, enterprises are finding it hard to meet the demand for skilled big data analytics professionals.