UNIT 1

INTRODUCTION TO BIG DATA

BIG DATA ANALYTICS

Big data analytics examines large amounts of data to uncover hidden patterns, correlations and other insights. With today's technology, it's possible to analyze your data and get answers from it almost immediately an effort that's slower and less efficient with more traditional business intelligence solutions.

Importance of Big Data Analytics

Big data analytics helps organizations harness their data and use it to identify new opportunities. That, in turn, leads to smarter business moves, more efficient operations, higher profits and happier customers. Businesses that use big data with advanced analytics gain value in many ways, such as:

1. **Reducing Cost :** Big data technologies like cloud-based analytics can significantly reduce costs when it comes to storing large amounts of data (for example, a data lake). Plus, big data analytics helps organizations find more efficient ways of doing business.
2. **Making Faster, Better Decisions :** The speed of in-memory analytics - combined with the ability to analyze new sources of data, such as streaming data from IoT - helps businesses analyze information immediately and make fast, informed decisions.
3. **Developing and Marketing New Products and Services :** Being able to gauge customer needs and customer satisfaction through analytics empowers businesses to give customers what they want, when they want it. With big data analytics, more companies have an opportunity to develop innovative new products to meet customers' changing needs.

1.1 BIG DATA

Introduction

Data can be defined as a representation of facts, concepts, or instructions in a formalized manner. Big data is extremely large data sets that may be analysed computationally to reveal patterns, trends, and associations, especially relating to human behaviour and interactions.

Big data is data that exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or doesn't fit the strictures of your database architectures. To gain value from this data, you must choose an alternative way to process it. Big Data has to deal with large and complex datasets that can be structured, Semi-structured, or unstructured and will typically not fit into memory to be processed.

According to Wikipedia : Big data is a field that treats ways to analyze, systematically extract information from, or otherwise deal with data sets that are too large or complex to be dealt with by traditional data processing application software.

3V's of Big Data

1. **Velocity**

   The data is increasing at a very fast rate. It is estimated that the volume of data will double in every 2 years.

2. **Variety**

   Now a days data are not stored in rows and column. Data is structured as well as unstructured. Log file, CCTV footage is unstructured data. Data which can be saved in tables are structured data like the transaction data of the bank.

3. **Volume**

   The amount of data which we deal with is of very large size of of Peta bytes.

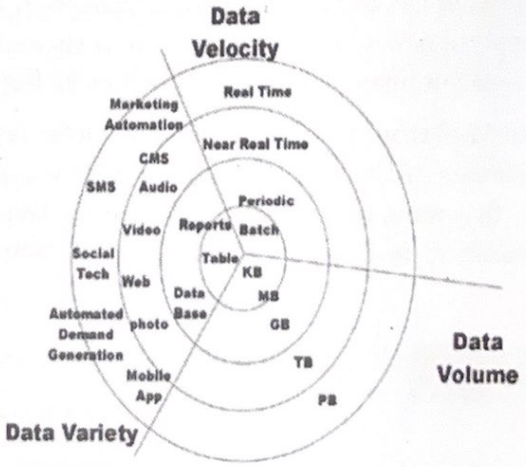

Fig. 3Vs of Big Data

Examples of Big Data

The New York Stock Exchange generates about one terabyte of new trade data per day. 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.

How Big Data Works

Big data gives you new insights that open up new opportunities and business models.

Getting started involves three key actions:

1. **Integrate**

   Big data brings together data from many disparate sources and applications. Traditional data integration mechanisms, such as extract, transform, and load (ETL) generally aren't up to the task. It requires new strategies and technologies to analyze big data sets at terabyte, or even petabyte, scale.

   During integration, you need to bring in the data, process it, and make sure it's formatted and available in a form that your business analysts can get started with.

2. **Manage**

   Big data requires storage. Your storage solution can be in the cloud, on premises, or both. You can store your data in any form you want and bring your desired processing requirements and necessary process engines to those data sets on an on-demand basis. Many people choose their storage solution according to where their data is currently residing. The cloud is gradually gaining popularity because it supports your current compute requirements and enables you to spin up resources as needed.

3. **Analyze**

   Your investment in big data pays off when you analyze and act on your data. Get new clarity with a visual analysis of your varied data sets. Explore the data further to make new discoveries. Share your findings with others. Build data models with machine learning and artificial intelligence. Put your data to work.

Use case

An e-commerce site XYZ (having 100 million users) wants to offer a gift voucher of 100$ to its top 10 customers who have spent the most in the previous year.Moreover, they want to find the buying trend of these customers so that company can suggest more items related to them.

Issues

Huge amount of unstructured data which needs to be stored, processed and analyzed

Solution :

1. **Storage:** This huge amount of data, Hadoop uses HDFS (Hadoop Distributed File System) which uses commodity hardware to form clusters and store data in a distributed fashion. It works on Write once, read many times principle.
2. **Processing:** Map Reduce paradigm is applied to data distributed over network to find the required output.
3. **Analyze:** Pig, Hive can be used to analyze the data.
4. **Cost:** Hadoop is open source so the cost is no more an issue.

1.2 WHAT IS BIG DATA AND WHY IS IT IMPORTANT ?

Big Data is the next generation of data warehousing and business analytics and is poised to deliver top line revenues cost efficiently for enterprises.

Big Data is defined as large or voluminous data that is difficult to store and also cannot be handled manually with traditional database systems. It is a collection of structured as well as unstructured data. Big data is a very vast field for anyone who is looking to make a career in the IT industry

Sources of data in Big Data

Big data can be of various formats of data either in structured as well as unstructured form, and comes from various different sources. The main sources of big data can be of the following types:

1. **Social Media**

   Data is collected from various social media platforms such as Facebook, Twitter, Instagram, Whatsapp, etc. Although data collected from these platforms can be anything like text, audio, video, etc., the biggest challenge is to store, manage and organize these data in an efficient way.

2. **Online cloud platforms**

   There are various online cloud platforms, such as Amazon AWS, Google Cloud, IBM cloud, etc., that are also used as a source of big data for machine learning.

3. **Internet of things**

The Internet of Things (IoT) is a platform that offers cloud facilities, including data storage and processing through IoT. Recently, cloud-based ML models are getting popular. It starts with invoking input data from the client end and processing machine learning algorithms using an artificial neural network (ANN) over cloud servers and then returning with output to the client again.

4. **Online Web pages**

Nowadays, every second, thousands of web pages are created and uploaded over the internet. These web pages can be in the form of text, images, videos, etc. Hence, these web pages are also a source of big data. Data which are very large in size is called Big Data. Normally we work on data of size MB(WordDoc ,Excel) or maximum GB(Movies, Codes) but data in Peta bytes i.e. 10^15 byte size is called Big Data. It is stated that almost 90% of today's data has been generated in the past 3 years.

Why Is It Important?

Here 's our standard answer in three parts:

1. **Computing perfect storm**

Big Data analytics are the natural result of four major global trends: Moore 's Law (which basically says that technology always gets cheaper), mobile computing (that smart phone or mobile tablet in your hand), social networking (Facebook, Foursquare, Pinterest, etc.), and cloud computing .

2. **Data perfect storm**

Volumes of transactional data have been around for decades for most big firms, but the flood gates have now opened with more volume , and the velocity and variety—the three Vs—of data that has arrived in unprecedented ways. This perfect storm of the three Vs makes it extremely complex and cumbersome with the current data management and analytics technology and practices.

3. **Convergence perfect storm**

Another perfect storm is happening, too. Traditional data management and analytics software and hardware technologies, open-source technology, and commodity hardware are merging to create new alternatives for IT and business executives to address Big Data analytics.

Misha Ghosh, who is known to be an innovator with several patents under his belt. Ghosh is currently an executive at MasterCard Advisors and before that he spent 11 years

at Bank of America solving business issues by using data. Ghosh explains, "Aside from the changes in the actual hardware and software technology, there has also been a massive change in the actual evolution of data systems. Using Misha 's analogy, let 's breakdown the three pinnacle stages in the Evolution of data systems:

1. **Dependent** (Early Days). Data systems were fairly new and users didn 't know quite know what they wanted. IT assumed that "Build it and they shall come."
2. **Independent** (Recent Years). Users understood what an analytical platform was and worked together with IT to define the business needs and approach for deriving insights for their fi rm.
3. **Interdependent** (Big Data Era). Interactional stage between various companies, creating more social collaboration beyond your firm 's walls.

1.3 A FLOOD OF MYTHIC "START-UP" PROPORTIONS

The global economy now generates unprecedented quantities of data. People who compare the amount of data produced daily to a deluge of mythic proportions are entirely correct. This fl ood of data represents something we 've never seen before. It 's new, it 's powerful, and yes, it 's scary but extremely exciting.

The best way to predict the future is to create it! —*Peter F. Drucker*

The influential writer and management consultant Drucker reminds us that the future is up to us to create. This is something that every entrepreneur takes to heart as they evangelize their start-up 's big idea that they know will impact the world! This is also true with Big Data and the new technology and approaches that have arrived at our doorstep.

Over the past decade companies like Facebook, Google, LinkedIn, and eBay have created treasured fi rms that rely on the skills of new data scientists, who are breaking the traditional barriers by leveraging new technology and approaches to capture and analyze data that drives their business. Time is flying and we have to remember that these fi rms were once start-ups. In fact, most of today 's start-ups are applying similar Big Data methods and technologies while they 're growing their businesses. The question is how.

This is why it is critical that organizations ensure that they have a mechanism to change with the times and not get caught up appeasing the ghost from data warehousing and business intelligence (BI) analytics of the past! At the end of the day, legacy data warehousing and BI analytics are not going away anytime soon. It 's all about fi nding the right home for the new approaches and making them work for you!

According to a recent study by the McKinsey Global Institute, organizations capture trillions of bytes of information about their customers, suppliers, and operations through digital systems. Millions of networked sensors embedded in mobile phones, automobiles, and other products are continually sensing, creating, and communicating data.

The result is a 40 percent projected annual growth in the volume of data generated. As the study notes, 15 out of 17 sectors in the U.S. economy already "have more data stored per company than the U.S. Library of Congress." The Library of Congress itself has collected more than 235 terabytes of data. That 's Big Data.

1.4 BIG DATA IS MORE THAN MERELY BIG

Edd Dumbill, founding chair of O 'Reilly 's Strata Conference and chair of the O 'Reilly Open Source Convention, defines Big Data as "data that becomes large enough that it cannot be processed using conventional methods."

Here is how the McKinsey study defines Big Data:

Big data refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze. This definition is intentionally subjective. . . . We assume that, as technology advances over time, the size of datasets that qualify as big data will also increase. Also note that the definition can vary by sector, depending on what kinds of software tools are commonly available and what sizes of datasets are common in a particular industry. With those caveats, big data in many sectors today will range from a few dozen terabytes to multiple petabytes (thousands of terabytes).

Big Data isn 't just a description of raw volume. "The real issue is usability," according to industry renowned blogger David Smith. From his perspective, big datasets aren 't even the problem. The real challenge is identifying or developing most cost-effective and reliable methods for extracting value from all the terabytes and petabytes of data now available. That 's where Big Data analytics become necessary.

Comparing traditional analytics to Big Data analytics is like comparing a horse-drawn cart to a tractor–trailer rig. The differences in speed, scale, and complexity are tremendous.

1.5 WHY NOW ?

On some level, we all understand that history has no narrative and no particular direction. But that doesn't stop us from inventing narratives and writing timelines complete with "important milestones." Keeping those thoughts in mind, Figure shows a timeline of recent technology developments.

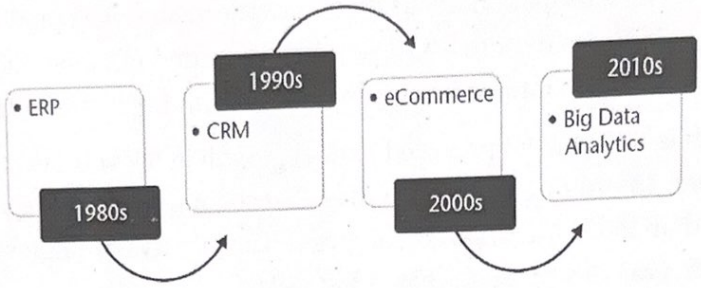

Figure : Timeline of Recent Technology Developments

If you believe that it 's possible to learn from past mistakes, then one mistake we certainly do not want to repeat is investing in new technologies that didn't fit into existing business frameworks. During the customer relationship management (CRM) era of the 1990s, many companies made substantial investments in customer-facing technologies that subsequently failed to deliver expected value. The reason for most of those failures was fairly straightforward: Management either forgot (or just didn't know) that big projects require a synchronized transformation of people, process, and technology.

All three must be marching in step or the project is doomed.

We can avoid those kinds of mistakes if we keep our attention focused on the outcomes we want to achieve. The technology of Big Data is the easy part the hard part is figuring out what you are going to do with the output generated by your Big Data analytics. As the ancient Greek philosophers said, "Action is character." It 's what you do that counts. Putting it bluntly, make sure that you have the people and process pieces ready before you commit to buying the technology.

1.6 A CONVERGENCE OF KEY TRENDS

Steve Lucas, is the Global Executive Vice President and General Manager, SAP Database & Technology at SAP. He 's an experienced player in the Big Data analytics space, and we 're delighted that he agreed to share some of his insights with us.

First of all, according to Lucas, it 's important to remember that big companies have been collecting and storing large amounts of data for a long time. From his perspective, the difference between "Old Big Data" and "New Big Data" is accessibility. Here 's a brief summary of our interview:

Companies have always kept large amounts of information. But until recently, they stored most of that information on tape. While it 's true that the amount of data in the world keeps growing, the real change has been in the ways that we access that data and use it to create value. Today, you have technologies like Hadoop, for example, that make it functionally practical to access a tremendous amount of data, and then extract value from it. The availability of lower-cost hardware makes it easier and more feasible to retrieve and process information, quickly and at lower costs than ever before.

So it 's the convergence of several trends—more data and less expensive, faster hardware—that 's driving this transformation. Today, we 've got raw speed at an affordable price. That cost/benefi t has really been a game changer for us.

That 's first and foremost—raw horsepower. Next is the ability to do that real-time analysis on very complex sets of data and models, so it 's not just let me look at my financials or let me look at marketing information. And finally, we now have the ability to find solutions for very complex problems in real time.

We asked Steve Lucas to offer some examples of scenarios in which the ability to analyze Big Data in real time is making an impact. Here 's what he told us:

A perfect example would be insurance companies. They need to know the answers to questions like this: As people age, what kinds of different services will they need from us?

In the past, the companies would have been forced to settle for general answers. Today, they can use their data to fi nd answers that are more specific and significantly more useful. Here are some examples that Lucas shared with us from the insurance and retail industries:

You don 't have to guess. You can look at actual data, from real customers. You can extract and analyze every policy they 've ever held. The answers to your questions are buried in this kind of massive mound of data—potentially petabytes worth of data if you consider all of your insurance customers across the lifespan of their policies. It 's unbelievable how much information exists.

But now you 've got to go from the level of petabytes and terabytes down to the level of a byte. That 's a very complex process. But today you can do it—you can compare one individual to all the other people in an age bracket and perform an analysis, in real time. That 's pretty powerful stuff. Imagine if a customer service rep had access to that kind of

information in real time. Think of all the opportunities and advantages there would be, for the company and for the customer.

Here 's another example: You go into a store to buy a pair of pants. You take the pants up to the cash register and the clerk asks you if you would like to save 10 percent off your purchase by signing up for the store 's credit card.

99.9 percent of the time, you 're going to say "no." But now let 's imagine if the store could automatically look at all of my past purchases and see what other items I bought when I came in to buy a pair of pants—and then offer me 50 percent off a similar purchase? Now that would be relevant to me. The store isn 't offering me another lame credit card—it 's offering me something that I probably want, at an attractive price.

The two scenarios described by Lucas are not fantasies. Yesterday, the cost of real-time data analysis was prohibitive. Today, real-time analytics have become affordable. As a result, market-leading companies are already using Big Data Analytics to improve sales revenue, increase profits, and do a better job of serving customers.

Before moving on, it 's worth repeating that not all new Big Data technology is open source. For example, SAP successfully entered the Big Data market with SAP HANA, an in-memory database platform for real-time analytics and applications. Products like SAP HANA are reminders that suppliers of proprietary solutions, such as SAP, SAS, Oracle, IBM, and Teradata, are playing—and will obviously continue to play—significant roles in the evolution of Big Data analytics.

1.7 RELATIVELY SPEAKING

Big Data, as you might expect, is a relative term. Although many people defi ne Big Data by volume, definitions of Big Data that are based on volume can be troublesome since some people define volume by the number of occurrences (in database terminology by the rows in a table or in analytics terminology known as the number of observations).

Some people define volume based on the number of interesting pieces of information for each occurrence (or in database terminology, the columns in a table or in analytics terminology the features or dimensions) and some people defi ne volume by the combination of depth and width.

If you 're a midmarket consumer packaged goods (CPG) company, you might consider 10 terabytes as Big Data. But if you 're a multinational pharmaceutical corporation, then

you would probably consider 500 terabytes as Big Data. If you 're a three-letter government agency, anything less than a petabyte is considered small.

The industry has an evolving definition around Big Data that is currently defined by three dimensions:

1. Volume
2. Variety
3. Velocity

These are reasonable dimensions to quantify Big Data and take into account the typical measures around volume and variety plus introduce the velocity dimension, which is a key compounding factor.

1. Data Volume

Data volume can be measured by the sheer quantity of transactions, events, or amount of history that creates the data volume, but the volume is often further exacerbated by the attributes, dimensions, or predictive variables. Typically, analytics have used smaller data sets called samples to create predictive models. Oftentimes, the business use case or predictive insight has been severely blunted since the data volume has purposely been limited due to storage or computational processing constraints. It 's similar to seeing the iceberg that sits above the waterline but not seeing the huge iceberg that lies beneath the surface.

By removing the data volume constraint and using larger data sets, enterprises can discover subtle patterns that can lead to targeted actionable microdecisions, or they can factor in more observations or variables into predictions that increase the accuracy of the predictive models. Additionally, by releasing the bonds on data, enterprises can look at data over a longer period of time to create more accurate forecasts that mirror real-world complexities of interrelated bits of information.

2. Data variety is the Assortment of Data

Traditionally data, especially operational data, is "structured" as it is put into a database based on the type of data (i.e., character, numeric, floating point, etc.). Over the past couple of decades, data has increasingly become "unstructured" as the sources of data have proliferated beyond operational applications.

Oftentimes, text, audio, video, image, geospatial, and Internet data (including click streams and log fi les) are considered unstructured data . However, since many of the sources of this data are programs the data is in actuality "semi structured." Semi-structured data is often a combination of different types of data that has some pattern or structure that is not as strictly defined as structured data.

For example, call center logs may contain customer name + date of call + complaint where the complaint information is unstructured and not easily synthesized into a data store.

3. **Data velocity**

Data velocity is about the speed at which data is created, accumulated, ingested, and processed. The increasing pace of the world has put demands on businesses to process information in real-time or with near real-time responses. This may mean that data is processed on the fl y or while "streaming" by to make quick, real-time decisions or it may be that monthly batch processes are run interday to produce more timely decisions.

1.8 A WIDER VARIETY OF DATA

The variety of data sources continues to increase. Traditionally, internally focused operational systems, such as ERP (enterprise resource planning) and CRM applications, were the major source of data used in analytic processing. However, in order to increase knowledge and awareness, the complexity of data sources that feed into the analytics processes is rapidly growing to include a wider variety of data sources such as:

- Internet data (i.e., clickstream, social media, social networking links)
- Primary research (i.e., surveys, experiments, observations)
- Secondary research (i.e., competitive and marketplace data, industry reports, consumer data, business data)
- Location data (i.e., mobile device data, geospatial data)
- Image data (i.e., video, satellite image, surveillance)
- Supply chain data (i.e., EDI, vendor catalogs and pricing, quality information)
- Device data (i.e., sensors, PLCs, RF devices, LIMs, telemetry)

The wide variety of data leads to complexities in ingesting the data into data storage. The variety of data also complicates the transformation (or the changing of data into a form that can be used in analytics processing) and analytic computation of the processing of the data.

1.9 THE EXPANDING UNIVERSE OF UNSTRUCTURED DATA

We spoke with Misha Ghosh to get a "level set" on the relationship between structured data (the kind that is easy to define, store, and analyze) and unstructured data (the kind that tends to defy easy definition, takes up lots of storage capacity, and is typically more difficult to analyze).

Unstructured data is basically information that either does not have a predefined data model and/or does not fi t well into a relational database. Unstructured information is typically text heavy, but may contain data such as dates, numbers, and facts as well. The term semi-structured data is used to describe structured data that doesn't fit into a formal structure of data models. However, semi-structured data does contain tags that separate semantic elements, which includes the capability to enforce hierarchies within the data.

At this point, it 's fair to ask: If unstructured data is such a pain in the neck, why bother? Here 's where Ghosh 's insight is priceless. Our conversation with him was long and wide-ranging, but here are the main takeaways that we would like to share with you:

- The amount of data (all data, everywhere) is doubling every two years.
- Our world is becoming more transparent. We, in turn, are beginning to accept this as we become more comfortable with parting with data that we used to consider sacred and private.
- Most new data is unstructured. Specifically, unstructured data represents almost 95 percent of new data, while structured data represents only 5 percent.
- Unstructured data tends to grow exponentially, unlike structured data, which tends to grow in a more linear fashion.
- Unstructured data is vastly underutilized. Imagine huge deposits of oil or other natural resources that are just sitting there, waiting to be used. That 's the current state of unstructured data as of today. Tomorrow will be a different story because there 's a lot of money to be made for smart individuals and companies that can mine unstructured data successfully.

The implosion of data is happening as we begin to embrace more open and transparent societies. "Résumés used to be considered private information," says Ghosh. "Not anymore with the advent of LinkedIn." We have similar stories with Instagram and Flickr for pictures,

	Improve Operational Efficiencies	Increase Revenues	Achieve Competitive Differentiation
Mature Analytic Applications	■ Supply chain optimization ■ Marketing campaign optimization	■ Algorithmic trading	■ In-house custom analytic applications
Maturing Analytic Applications	■ Portfolio optimization ■ Risk management ■ Next best offer	■ Ad targeting optimization ■ Yield optimization	■ In-house custom analytic applications
Emerging Analytic Applications	■ Chronic disease prediction and prevention	■ Customer churn prevention	■ Product design optimization ■ Design of experiments optimization ■ Brand ■ Product Market

Table : Enabling Big Data Analytic Applications

While Big Data analytics may not be the "Final Frontier," it certainly represents an enormous opportunity for businesses to exploit their data assets to realize substantial bottom line results for their enterprise. The key to success for organizations seeking to take advantage of this opportunity is:

- Leverage all your current data and enrich it with new data sources
- Enforce data quality policies and leverage today' s best technology and people to support the policies
- Relentlessly seek opportunities to imbue your enterprise with fact based decision making
- Embed your analytic insights throughout your organization.

IMPORTANT QUESTIONS

1. What is big data and why is it important ?
2. Explain about a flood of mythic "start-up" proportions.
3. Discuss briefly about big data is more than merely big.
4. Why big data now?
5. Explain about a convergence of key trends and relatively speaking.
6. Discuss briefly about a wider variety of data and the expanding universe of unstructured data.