Introduction to Big Data: Types of digital data, history of Big Data innovation, introduction to Big Data platform, drivers for Big Data, Big Data architecture and characteristics, 5 Vs of Big Data, Big Data technology components, Big Data importance and applications, Big Data features – security, compliance, auditing and protection, Big Data privacy and ethics, Big Data Analytics, Challenges of conventional systems, intelligent data analysis, nature of data, analytic processes and tools, analysis vs reporting, modern data analytic tools.

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. Data with many fields (columns) offer greater statistical power, while data with higher complexity (more attributes or columns) may lead to a higher false discovery rate. Big data analysis challenges include capturing data, data storage, data analysis,
search, sharing, transfer, visualization, querying, updating, information privacy, and data source. Big data was originally associated with three key concepts: volume, variety, and velocity. The analysis of big data presents challenges in sampling, and thus previously allowing for only observations and sampling. Therefore, big data often includes data with sizes that exceed the capacity of traditional software to process within an acceptable time and value.

Current usage of the term big data tends to refer to the use of predictive analytics, user behavior analytics, or certain other advanced data analytics methods that extract value from big data, and seldom to a particular size of data set.

Types of digital data

Unstructured data: This is the data which does not conform to a data model or is not in a form which can be used easily by a computer program. About 80—90% data of an organization is in this format; for example, memos, chat rooms, PowerPoint presentations, images, videos, letters, researches, white papers, body of an email, etc.

Semi-structured data: This is the data which does not conform to a data model but has some structure structure. However, However, it is not in a form which can be used easily by a computer program; for example, emails, XML, markup languages like HTML, etc. Metadata for this data is available but is not sufficient.

Structured data: This is the data which is in an organized form (e.g., in rows and columns) and can be easily used by a computer program. Relationships exist between entities of data, such as classes and their objects. Data stored in databases is an example of structured data

Big data characteristics

Big data can be described by the following characteristics:

Volume

The quantity of generated and stored data. The size of the data determines the value and potential insight, and whether it can be considered big data or not. The size of big data is usually larger than terabytes and petabytes.

Variety

The type and nature of the data. The earlier technologies like RDBMSs were capable to handle structured data efficiently and effectively. However, the change in type and nature from structured to semi-structured or unstructured challenged the existing tools and technologies. The big data technologies evolved with the prime intention to capture, store, and process the semi-structured and unstructured (variety) data generated with high speed (velocity), and huge in size (volume). Later, these tools and technologies were explored and used for handling structured data also but preferable for storage. Eventually, the processing of structured data was still kept as optional, either using big data or traditional

RDBMSs. This helps in analyzing data towards effective usage of the hidden insights exposed from the data collected via social media, log files, sensors, etc. Big data draws from text, images, audio, video; plus it completes missing pieces through data fusion.

Velocity

The speed at which the data is generated and processed to meet the demands and challenges that lie in the path of growth and development. Big data is often available in real-time. Compared to small data, big data is produced more continually. Two kinds of velocity related to big data are the frequency of generation and the frequency of handling, recording, and publishing.

Veracity

The truthfulness or reliability of the data, which refers to the data quality and the data value. Big data must not only be large in size, but also must be reliable in order to achieve value in the analysis of it. The data quality of captured data can vary greatly, affecting an accurate analysis.

Value

The worth in information that can be achieved by the processing and analysis of large datasets. Value also can be measured by an assessment of the other qualities of big data. Value may also represent the profitability of information that is retrieved from the analysis of big data.

Big Data architecture

Big data solutions typically involve one or more of the following types of workload:

- Batch processing of big data sources at rest.
- Real-time processing of big data in motion.
- Interactive exploration of big data.
- Predictive analytics and machine learning.

Most big data architectures include all of the following components:

- **Data sources**: All big data solutions start with one or more data sources. Examples include:
  - Application data stores, such as relational databases.
  - Static files produced by applications, such as web server log files.
  - Real-time data sources, such as IoT devices.
- **Data storage**: Data for batch processing operations is typically stored in a distributed file store that can hold high volumes of large files in various formats. This kind of store is often called a *data lake*.
- **Batch processing**: Because the data sets are so large, often a big data solution must process data files using long-running batch jobs to filter, aggregate, and otherwise prepare the data for analysis. Usually these jobs involve reading source files, processing them, and writing the output to new files.
- **Real-time message ingestion**: If the solution includes real-time sources, the architecture must include a way to capture and store real-time messages for stream processing. This might be a simple data store, where incoming messages are dropped into a folder for processing. However, many solutions need a message ingestion store to act as a buffer for messages, and to support scale-out processing, reliable delivery, and other message queuing semantics.
- **Stream processing**: After capturing real-time messages, the solution must process them by filtering, aggregating, and otherwise preparing the data for analysis. The processed stream data is then written to an output sink.

- **Analytical data store**: Many big data solutions prepare data for analysis and then serve the processed data in a structured format that can be queried using analytical tools.
- **Analysis and reporting**: The goal of most big data solutions is to provide insights into the data through analysis and reporting. To empower users to analyze the data, the architecture may include a data modeling layer, such as a multidimensional OLAP cube or tabular data model
- **Orchestration**: Most big data solutions consist of repeated data processing operations, encapsulated in workflows, that transform source data, move data between multiple sources and sinks, load the processed data into an analytical data store, or push the results straight to a report or dashboard. To automate these workflows, you can use an orchestration technology such Azure Data Factory or Apache Oozie and Sqoop.

Big Data technology components

1. Machine Learning

It is the science of making computers learn stuff by themselves. In machine learning, a computer is expected to use algorithms and statistical models to perform specific tasks without any explicit instructions. Machine learning applications provide results based on past experience. For example, these days, there are some mobile applications that will give you a summary of your finances, bills, will remind you of your bill payments, and also may give you suggestions to go for some saving plans. These functions are done by reading your emails and text messages.

2. Natural Language Processing (NLP)

It is the ability of a computer to understand human language as spoken. The most obvious examples that people can relate to these days are google home and Amazon Alexa. Both use NLP and other technologies to give us a virtual assistant experience. NLP is all around us without us even realizing it. When writing a mail, while making any mistakes, it automatically corrects itself, and these days it gives auto-suggests for completing the mails and automatically intimidates us when we try to send an email without the attachment that we referenced in the text of the email, this is part of Natural Language Processing Applications which are running at the backend.

3. Business Intelligence

Business Intelligence (BI) is a method or process that is technology-driven to gain insights by analyzing data and presenting it in a way that the end-users (usually high-level executives) like managers and corporate leaders can gain some actionable insights from it and make informed business decisions on it.

4. Cloud Computing

If we go by the name, it should be computing done on clouds; well, it is true, just here we are not talking about real clouds, cloud here is a reference for the Internet. So we can define cloud computing as the delivery of computing services—servers, storage, databases, networking, software, analytics, intelligence, and moreover the Internet ("the cloud") to offer faster innovation, flexible resources, and economies of scale.

Big Data importance and applications

Travel and tourism are the users of Big Data. It enables us to forecast travel facilities requirements at multiple locations, improve business through dynamic pricing, and many more.

The financial and banking sectors use big data technology extensively. Big data analytics help banks and customer behaviour on the basis of investment patterns, shopping trends, motivation to invest, and inputs that are obtained from personal or financial backgrounds.

Big data has started making a massive difference in the **healthcare sector**, with the help of predictive analytics, medical professionals, and health care personnel. It can produce personalized healthcare and solo patients also.

Telecommunications and the multimedia sector are the main users of Big Data. There are zettabytes to be generated every day and handling large-scale data that require big data technologies.

The government and military also used technology at high rates. We see the figures that the government makes on the record. In the military, a fighter plane requires to process petabytes of data.Government agencies use Big Data and run many agencies, managing utilities, dealing with traffic jams, and the effect of crime like hacking and online fraud.

Aadhar Card: The government has a record of 1.21 billion citizens. This vast data is analyzed and store to find things like the number of youth in the country. Some schemes are built to target the maximum population. Big data cannot store in a traditional database, so it stores and analyze data by using the Big Data Analytics tools.

E-commerce is also an application of Big data. It maintains relationships with customers that is essential for the e-commerce industry. E-commerce websites have many marketing ideas to retail merchandise customers, manage transactions, and implement better strategies of innovative ideas to improve businesses with Big data.

Social Media is the largest data generator. The statistics have shown that around 500+ terabytes of fresh data generated from social media daily, particularly on Facebook. The data mainly contains videos, photos, message exchanges, etc. A single activity on the social media site generates many stored data and gets processed when required. The data stored is in terabytes (TB); it takes a lot of time for processing. Big Data is a solution to the problem.

Big Data features – security, compliance, auditing and protection

Enterprises are using big data analytics to identify business opportunities, improve performance, and drive decision-making. Many big data tools are open source and not designed with security in mind. The huge increase in data consumption leads to many data security concerns.Big data security is an umbrella term that includes all security measures and tools applied to analytics and data processes. Attacks on big data systems – information theft, DDoS attacks(Distributed Denial of Service Attack), ransomware, or other malicious activities – can originate either from offline or online spheres and can crash a system.The consequences of information theft can be even worse when organizations store sensitive or confidential information like credit card numbers or customer information. They may face fines because they failed to meet basic data security measures to be in compliance with data loss protection and privacy mandates like the General Data Protection Regulation (GDPR).

Data Security Challenges

Distributed Data

Most big data frameworks distribute data processing tasks throughout many systems for faster analysis. Hadoop, for example, is a popular open-source framework for distributed data processing and storage. Hadoop was originally designed without any security in mind.

Cybercriminals can force the MapReduce mapper to show incorrect lists of values or key pairs, making the MapReduce process worthless. Distributed processing may reduce the workload on a system, but eventually more systems mean more security issues.

Non-Relational Databases

Traditional relational databases use tabular schema of rows and columns. As a result, they cannot handle big data because it is highly scalable and diverse in structure. Non-relational databases, also known as NoSQL databases, are designed to overcome the limitations of relational databases.

Non-relational databases do not use the tabular schema of rows and columns. Instead, NoSQL databases optimize storage models according to data type. As a result, NoSQL databases are more flexible and scalable than their relational alternatives.

NoSQL databases favor performance and flexibility over security. Organizations that adopt NoSQL databases have to set up the database in a trusted environment with additional security measures.

Endpoint Vulnerabilities

Cybercriminals can manipulate data on endpoint devices and transmit the false data. Security solutions that analyze logs from endpoints need to validate the authenticity of those endpoints.

For example, hackers can access manufacturing systems that use sensors to detect malfunctions in the processes. After gaining access, hackers make the sensors show fake results. Challenges like that are usually solved with fraud detection technologies.

Access Controls

Companies sometimes prefer to restrict access to sensitive data like medical records that include personal information. But people that do not have access permission, such as medical researchers, still need to use this data. The solution in many organizations is to grant granular access. This means that individuals can access and see only the information they need to see.

Big data technologies are not designed for granular access. A solution is to copy required data to a separate big data warehouse. For example, only the medical information is copied for medical research without patient names and addresses.

Addressing Big Data Security Threats

Security tools for big data are not new. They simply have more scalability and the ability to secure many data types. The list below explains common security techniques for big data.

Encryption

Big data encryption tools need to secure data-at-rest and in-transit across large data volumes. Companies also need to encrypt both user and machine-generated data. As a result, encryption tools have to operate on multiple big data storage formats like NoSQL databases and distributed file systems like Hadoop.

User Access Control

User access control is a basic network security tool. The lack of proper access control measures can be disastrous for big data systems. A robust user control policy has to be based on automated role-based settings and policies. Policy-driven access control protects big data platforms against insider threats by automatically managing complex user control levels, like multiple administrator settings.

Intrusion Detection and Prevention

The distributed architecture of big data is a plus for intrusion attempts. An Intrusion Prevention System (IPS) enables security teams to protect big data platforms from vulnerability exploits by examining network traffic. The IPS often sits directly behind the firewall and isolates the intrusion before it does actual damage.

Centralized Key Management

Key management is the process of protecting cryptographic keys from loss or misuse. Centralized key management offers more efficiency as opposed to distributed or application-specific management. Centralized management systems use a single point to secure keys and access audit logs and policies. A reliable key management system is essential for companies handling sensitive information.

Granular auditing is a must in Big Data security, particularly after an attack on your system. Organizations create a cohesive audit view following any attack, and be sure to provide a full audit trail while ensuring there's easy access to that data in order to cut down incident response time.

Audit information integrity and confidentiality are also essential. Audit information should be stored separately and protected with granular user access controls and regular monitoring. Make sure to keep your Big Data and audit data separate, and enable all required logging when you're setting up auditing (in order to collect and process the most detailed information possible)..

Compliance is always a headache for enterprises, and even more so when you're dealing with a constant deluge of data. It's best to tackle it head-on with real-time analytics and security at every level of the stack. Organizations apply Big Data analytics by using tools such as Kerberos(Kerberos is a computer network security protocol that authenticates service requests between two or more trusted hosts across an untrusted network, like the internet. It uses secret-key cryptography and a trusted third party for authenticating client-server applications and verifying users' identities.), secure shell (SSH also known as Secure Shell or Secure Socket Shell, is a network protocol that gives users, particularly system administrators, a secure way to access a computer over an unsecured network.), and internet protocol security (IPsec) to get a handle on real-time data.

Once you're doing that, you can mine logging events, deploy front-end security systems such as routers and application-level firewalls, and begin implementing security controls throughout the stack at the cloud, cluster, and application levels.

Big Data privacy and ethics

Big data ethics also known as simply data ethics refers to systemizing, defending, and recommending concepts of right and wrong conduct in relation to data, in particular personal data

Data ethics is concerned with the following principles

1. Ownership - Individuals own their own data.
2. Transaction transparency - If an individuals personal data is used, they should have transparent access to the algorithm design used to generate aggregate data sets
3. Consent - If an individual or legal entity would like to use personal data, one needs informed and explicitly expressed consent of what personal data moves to whom, when, and for what purpose from the owner of the data.
4. Privacy - If data transactions occur all reasonable effort needs to be made to preserve privacy.
5. Currency - Individuals should be aware of financial transactions resulting from the use of their personal data and the scale of these transactions.
6. Openness - Aggregate data sets should be freely available

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.

Why is big data analytics important?

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.

1. Cost reduction. Big data technologies such as Hadoop and cloud-based analytics bring significant cost advantages when it comes to storing large amounts of data – plus they can identify more efficient ways of doing business.
2. Faster, better decision making. With the speed of Hadoop and in-memory analytics, combined with the ability to analyze new sources of data, businesses are able to analyze information immediately – and make decisions based on what they've learned.
3. New products and services. With the ability to gauge customer needs and satisfaction through analytics comes the power to give customers what they want. Davenport points out that with big data analytics, more companies are creating new products to meet customers' needs.

Challenges of conventional systems

Storage and analysis of large data sets

Complex data sets that can be both structured or unstructured.

They are so large that it is not possible to work on them with traditional analytical tools.

One of the major challenges of conventional systems was the uncertainty of the Data Management Landscape.

A big challenge for companies is to find out which technology works bests for them without the introduction of new risks and problems.

Intelligent Data Analysis (IDA)

Intelligent Data Analysis (IDA) is one of the hot issues in the field of artificial intelligence and information. Intelligent data analysis reveals implicit, previously unknown and potentially valuable information or knowledge from large amounts of data. Intelligent data analysis is also a kind of decision support process. Based on artificial intelligence, machine learning, pattern recognition, statistics, database and visualization technology mainly, IDA automatically extracts useful information, necessary knowledge and interesting models from a lot of online data in order to help decision makers make the right choices.

The process of IDA generally consists of the following three stages: (1) data preparation; (2) rule finding or data mining; (3) result validation and explanation. Data preparation involves selecting the required data from the relevant data source and integrating this into a data set to be used for data mining. Rule finding is working out rules contained in the data set by means of certain methods or algorithms. Result validation requires examining these rules, and result explanation is giving intuitive, reasonable and understandable descriptions using logical reasoning.

As the goal of intelligent data analysis is to extract useful knowledge, the process demands a combination of extraction, analysis, conversion, classification, organization, reasoning, and so on. It is challenging and fun working out how to choose appropriate methods to resolve the difficulties encountered in the process. Intelligent data analysis methods and tools, as well as the authenticity of obtained results pose us continued challenges.

Modern Big Data analytics tool

Hadoop - helps in storing and analyzing data
MongoDB - used on datasets that change frequently
Talend - used for data integration and management
Cassandra - a distributed database used to handle chunks of data
Spark - used for real-time processing and analyzing large amounts of data
STORM - an open-source real-time computational system
Kafka - a distributed streaming platform that is used for fault-tolerant storage

Analytics Vs Reporting in Big Data

Here are five differences between reporting and analysis:

1. Purpose

Reporting helps companies monitor their data even before digital technology boomed. Various organizations have been dependent on the information it brings to their business, as reporting extracts that and makes it easier to understand.

Analysis interprets data at a deeper level. While reporting can link between cross-channels of data, provide comparison, and make understand information easier (think of a dashboard, charts, and graphs, which are reporting *tools* and not analysis reports), analysis interprets this information and provides recommendations on actions.

2. Tasks

As reporting and analysis have a very fine line dividing them, sometimes it’s easy to confuse tasks that have analysis labeled on top of them when all it does is reporting. Hence, ensure that your analytics team has a healthy balance doing both.

Here’s a great differentiator to keep in mind if what you’re doing is reporting or analysis:

Reporting includes building, configuring, consolidating, organizing, formatting, and summarizing. It’s very similar to the above mentioned like turning data into charts, graphs, and linking data across multiple channels.

Analysis consists of questioning, examining, interpreting, comparing, and confirming. With big data, predicting is possible as well.

3. Outputs

Reporting and analysis have the push and pull effect from its users through their outputs. Reporting has a push approach, as it pushes information to users and outputs come in the forms of canned reports, dashboards, and alerts.

Analysis has a pull approach, where a data analyst draws information to further probe and to answer business questions. Outputs from such can be in the form of ad hoc responses and analysis presentations. Analysis presentations are comprised of insights, recommended actions, and a forecast of its impact on the company—all in a language that's easy to understand at the level of the user who'll be reading and deciding on it.

This is important for organizations to realize truly the value of data, such that a standard report is not similar to a meaningful analytics.

4. Delivery

Considering that reporting involves repetitive tasks—often with truckloads of data, automation has been a lifesaver, especially now with big data. It's not surprising that the first thing outsourced are data entry services since outsourcing companies are perceived as data reporting experts.

Analysis requires a more custom approach, with human minds doing superior reasoning and analytical thinking to extract insights, and technical skills to provide efficient steps towards accomplishing a specific goal. This is why data analysts and scientists are demanded these days, as organizations depend on them to come up with recommendations for leaders or business executives make decisions about their businesses.

5. Value

This isn't about identifying which one brings more value, rather understanding that both are indispensable when looking at the big picture. It should help businesses grow, expand, move forward, and make more profit or increase their value.

This Path to Value diagram illustrates how data converts into value by reporting and analysis such that it's not achievable without the other.