Introduction to Big Data

- By Shraddha Nikam

Data can be defined as figures or facts that can be stored in or can be used by a computer.

Now, what is Big Data?

Big Data is a term that is used for denoting a collection of datasets that is large and complex, making it very difficult to process using traditional data processing applications.

What is Big Data Analytics?

Big Data Analytics examines large and different types of data to uncover hidden patterns, insights, and correlations. Big Data Analytics is helping large companies facilitate their growth and development. And it majorly includes applying various data mining algorithms on a certain dataset.

Types of Big Data

Big Data is essentially classified into three types:

- Structured Data
- Unstructured Data
- Semi-structured Data

Structured Data

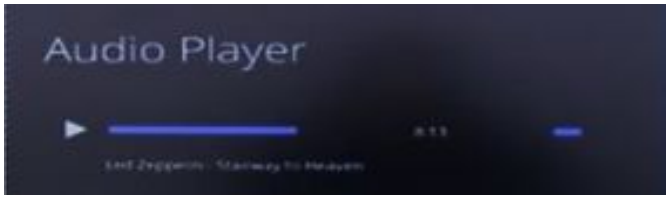
- Structured data are those type of data which are stored already in an order. There are nearly 20% of the total existing data are structured data.
- All the data generated from Sensors, weblogs, these are all Machine Generated Structured Data. The human- generated structured data are those which are taken as information from a human. Like their names, addresses etc.
- The example of Structured Data is Database.

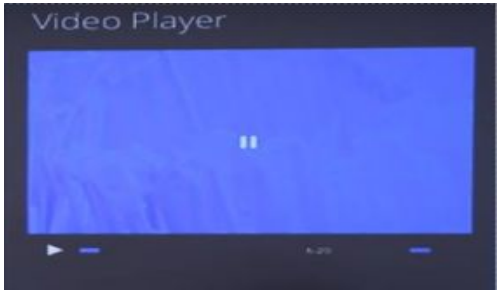
Emp_id	Emp_name	Job_name	Salary	Mobile_no	Dep_id	Project_id
AfterA001	John	Engineer	100000	9111037890	2	99
AfterA002	Adam	Analyst	50000	9587569214	3	100
AfterA003	Kande	Manager	890000	7895212355	2	65

EMPLOYEE TABLE

Unstructured Data

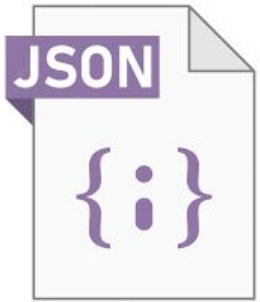
- The Unstructured data have no clear format in storage. We can store structured data in the row-column database, but unstructured data cannot be stored like that. At least 80% of data are unstructured.
- All satellite-generated images, scientific data or images are categorized as machine-generated unstructured data. There are various types of human-generated unstructured data. These are images, videos, social media data etc.
- The examples of Unstructured data are text documents, PDFs, Images, videos etc.





Semi-Structured Data

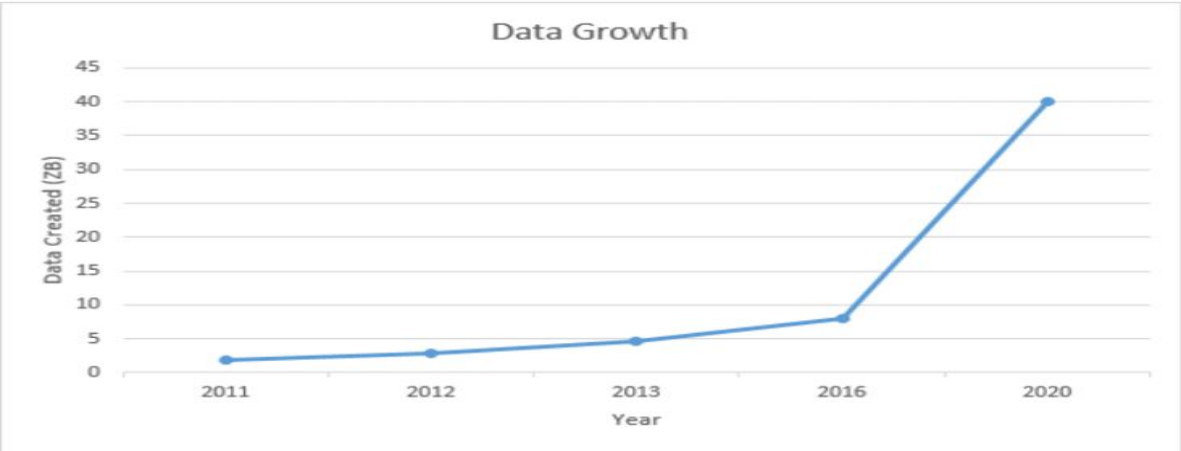
- It is very difficult to categorize this type of data. Sometimes they look structured, or sometimes unstructured. So that's why these data are known as semi-structured data. We cannot store these type of data using traditional database format, but it contains some organizational properties.
- The examples of Semi-Structured Data are Spreadsheet files, XML or JSON documents, NoSQL database data items etc.



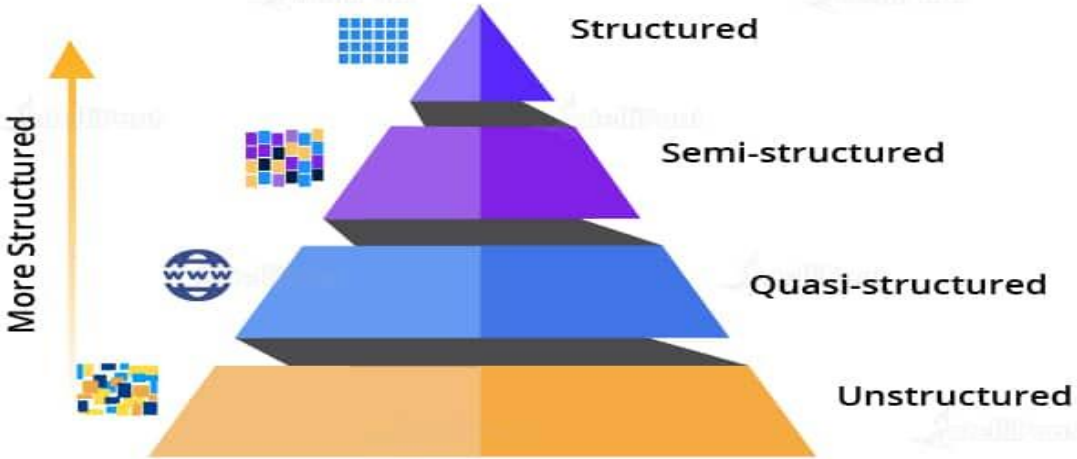
Characteristics of Big Data

Big Data has the following distinct characteristics:

1. Volume: This refers to tremendously large data. As you can see from the image, the volume of data is rising exponentially. In 2016, the data created was only 8 ZB; it is expected that, by 2020, the data would rise to 40 ZB, which is extremely large.



2. Variety: A reason for this rapid growth of data volume is that data is coming from different sources in various formats. We have already discussed how data is categorized into different types. Let us take another glimpse at it with more examples.



a) Structured Data: Here, data is present in a structured schema along with all the required columns. It is in a structured or tabular format. Data that is stored in a relational database management system is an example of structured data. For example, in the below-given employee table, which is present in a database, the data is in a structured format.

Emp. ID	Emp. Name	Gender	Department	Salary (INR)
2383	ABC	Male	Finance	650,000
4623	XYZ	Male	Admin	5,000,000

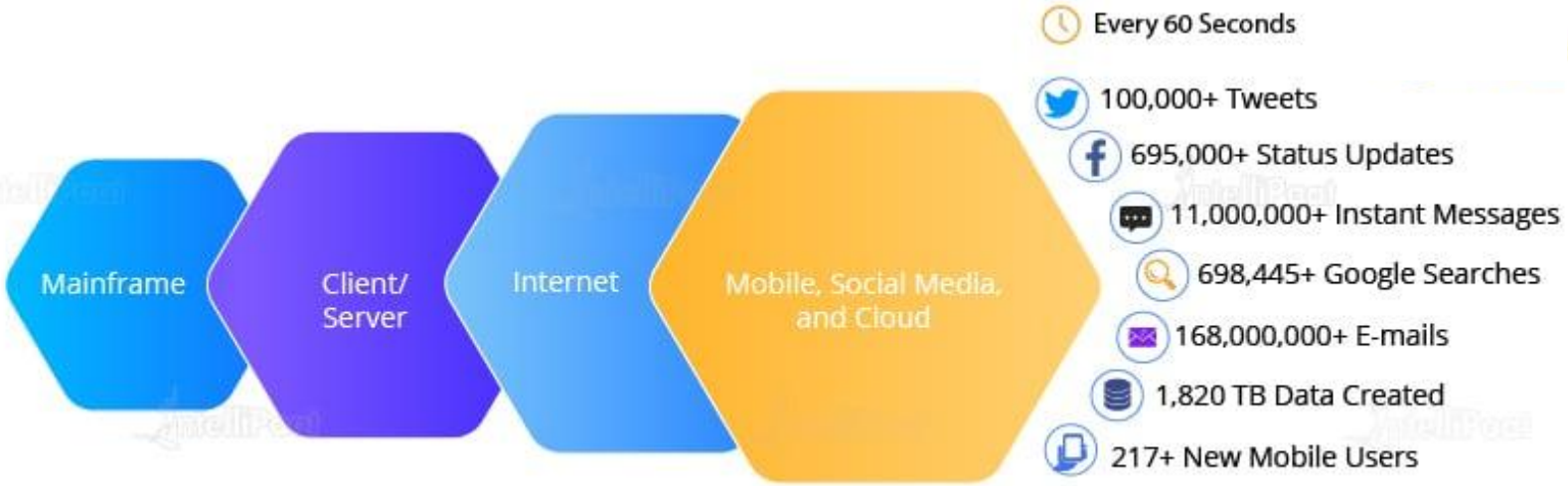
b) Semi-structured Data: In this form of data, the schema is not properly defined, i.e., both forms of data are present. So, semi-structured data has a structured form but it is not defined; for example, JSON, XML, CSV, TSV, and email. The web application data that is unstructured contains transaction history files, log files, etc. Online Transaction Processing (OLTP) systems are built to work with structured data, and this data is stored in relations, i.e., tables.

c) Unstructured Data: This data format includes all unstructured files such as video files, log files, audio files, and image files. Any data that has an unfamiliar model or structure is categorized as unstructured data. Since its size is large, unstructured data possesses various challenges in terms of processing for deriving value out of it.

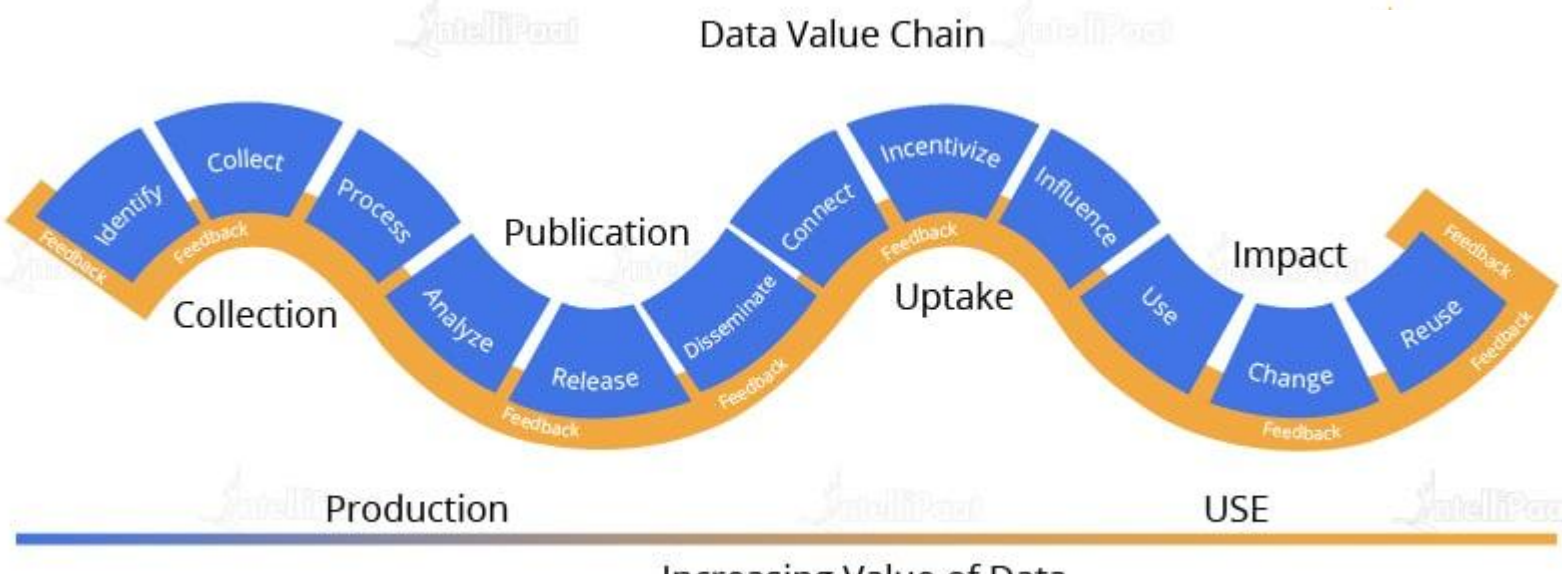
An example of this is a complex data source that contains a blend of text files, videos, and images. Several organizations have a lot of data available with them but they don't know how to derive value out of it since the data is in its raw form.

d) Quasi-structured Data: This data format consists of textual data with inconsistent data formats that can be formatted with effort, time, and with the help of several tools. For example, web server logs, i.e., a log file that is automatically created and maintained by a server that contains a list of activities.

3. Velocity: The speed of data accumulation also plays a role in determining whether the data is big data or normal data. As can be seen in the image below, mainframes were initially used when fewer people were using computers. As computers evolved, the client/server model came into existence. Later, web applications came into the picture and their popularity extended to more and more devices such as mobiles, which led to the creation of a lot of data!



4. Value: How will the extraction of data work? Here, our fourth V comes in; it deals with a mechanism to bring out the correct meaning of data. First of all, you need to mine data, i.e., the process to turn raw data into useful data. Then, an analysis is done on the data that you have cleaned or retrieved from the raw data. Then, you need to make sure whatever analysis you have done benefits your business, such as in finding out insights, results, etc., in a way that was not possible earlier.



You need to make sure to clean up whatever raw data you are given for deriving business insights. After you have cleaned the data, a challenge pops up, i.e., during the process of dumping a large amount of data, some packages might be lost.

So, to resolve this issue, our next V comes into the picture.

5. Veracity: Since packages get lost during execution, we need to start again from the stage of mining raw data to convert it into valuable data. And this process goes on. There will also be uncertainties and inconsistencies in the data that can be overcome by veracity. Veracity means the trustworthiness and quality of data. The veracity of data must be maintained. For example, think about Facebook posts, hashtags, abbreviations, images, videos, etc., which make the posts unreliable and hamper the quality of their content. Collecting loads and loads of data is of no use if the quality and trustworthiness of the data are not up to the mark.

Applications of big data

Banking

The amount of data in the banking sector is skyrocketing every second. Proper study and analysis of this data can help detect any illegal activities that are being carried out such as:

- Misuse of credit/debit cards
- Venture credit hazard treatment
- Business clarity
- Customer statistics alteration
- Money laundering
- Risk mitigation


EMAIL
CALL CENTER
SHOPPING BASKET
ATM
POS
MOBILE BANKING
BRANCH
CRM
INTERNET BANKING
WATCH LISTS
CREDIT RATING
ACCOUNT LEDGER

Government

Government agencies utilize Big Data and have devised a lot of running agencies, managing utilities, dealing with traffic jams, or limiting the effects of crime. However, apart from its benefits in Big Data, the government also addresses the concerns of transparency and privacy.

- **Aadhar Card:** The Indian government has a record of all 1.21 billion citizens. This huge data is stored and analyzed to find out several things, such as the number of youth in the country. According to which several schemes are made to target the maximum population. All this big data can’t be stored in some traditional database, so it is left for storing and analyzing using several Big Data Analytics tools


Building and Constructions
Political Decisions
Unemployment
Health Welfare
Electricity
Tax Evaders
TAX
0%
Agriculture
Natural Disasters
Insurance

Education

Education concerning Big Data produces a vital impact on students, school systems, and curriculums. By interpreting big data, people can ensure students' growth, identify at-risk students, and achieve an improvised system for the evaluation and assistance of principals and teachers.


High-speed
Connectivity to Schools
Data Privacy
and Security
Quality Digital
Content and Resources
Digitalized Records
of Students
Internet Access
Provided at Home
High-quality
Low-cost Devices
Online Portals Sharing
Digital Course Material

Example: The education sector holds a lot of information concerning curriculum, students, and faculty. The information is analyzed to get insights that can enhance the operational adequacy of the educational organization. Collecting and analyzing information about a student such as attendance, test scores, grades, and other issues take up a lot of data. So, big data approaches a progressive framework wherein this data can be stored and analyzed making it easier for the institutes to work with.

Big Data in Healthcare

When it comes to what Big Data is in Healthcare, we can see that it is being used enormously. It includes collecting data, analyzing it, leveraging it for customers. Also, patients' clinical data is too complex to be solved or understood by traditional systems. Since big data is processed by Machine Learning algorithms and Data Scientists, tackling such huge data becomes manageable.

Example: Nowadays, doctors rely mostly on patients' clinical records, which means that a lot of data needs to be gathered, that too for different patients. It is not possible for old or traditional data storage methods to store this data. Since there is a large amount of data coming from different sources, in various formats, the need to handle this large amount of data is increased, and that is why the Big Data approach is needed.

Big Data Contributions to Healthcare

E-commerce

Maintaining customer relationships is the most important in the e-commerce industry. E-commerce websites have different marketing ideas to retail their merchandise to their customers, manage transactions, and implement better tactics of using innovative ideas with Big Data to improve businesses.

Flipkart: Flipkart is a huge e-commerce website dealing with lots of traffic daily. But, when there is a pre-announced sale on Flipkart, traffic grows exponentially that crashes the website. So, to handle this kind of traffic and data, Flipkart uses Big Data. Big Data can help in organizing and analyzing the data for further use.

Social Media

Social media in the current scenario is considered the largest data generator. The stats have shown that around 500+ terabytes of new data get generated into the databases of social media every day, particularly in the case of Facebook. The data generated mainly consist of videos, photos, message exchanges, etc. A single activity on any social media site generates a lot of data which is again stored and gets processed whenever required. Since the data stored is in terabytes, it would take a lot of time for processing if it is done by our legacy systems. Big Data is a solution to this problem.

Challenges with Big Data

Since there is so much of big data sometimes it is hard to find out what the real valuable data is and what the noise in it is.

The second issue is with regard to data that is in silos. Since data is coming in from various sources most of the data is not compatible with each other and there is no uniformity and hence this issue needs to be taken care of.

Sometimes there is too much inaccurate data and all this should be taken into consideration before deploying it for applications in the real world.

Storage

With vast amounts of data generated daily, the greatest challenge is storage (especially when the data is in different formats) within legacy systems. Unstructured data cannot be stored in traditional databases.

Processing

Processing big data refers to the reading, transforming, extraction, and formatting of useful information from raw information. The input and output of information in unified formats continue to present difficulties.

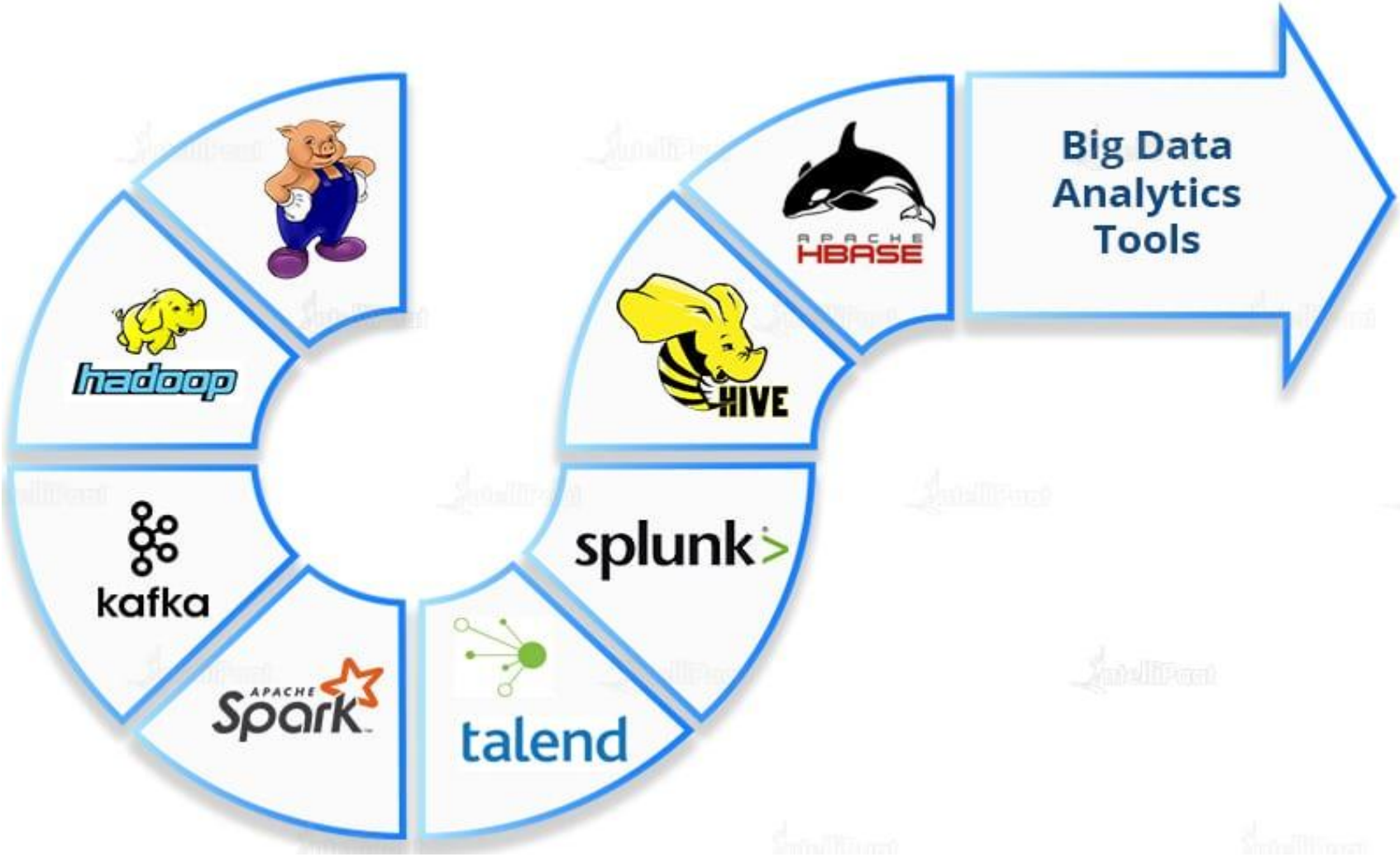
Security

Security is a big concern for organizations. Non-encrypted information is at risk of theft or damage by cyber-criminals. Therefore, data security professionals must balance access to data against maintaining strict security protocols.

Finding and Fixing Data Quality Issues

Many of you are probably dealing with challenges related to poor data quality, but solutions are available. The following are four approaches to fixing data problems:

- Correct information in the original database.
- Repairing the original data source is necessary to resolve any data inaccuracies.
- You must use highly accurate methods of determining who someone is.


Big Data
Analytics
Tools
APACHE
HBASE
HIVE
splunk>
talend
APACHE
Spark
kafka
hadoop

- **Apache Hadoop**

  Big Data Hadoop is a framework that allows you to store big data in a distributed environment for parallel processing.
- **Apache Pig**

  Apache Pig is a platform that is used for analyzing large datasets by representing them as data flows. Pig is designed to provide an abstraction over MapReduce which reduces the complexities of writing a MapReduce program.
- **Apache HBase**

  Apache HBase is a multidimensional, distributed, open-source, and NoSQL database written in Java. It runs on top of HDFS providing Bigtable-like capabilities for Hadoop.
- **Apache Spark**

  Apache Spark is an open-source general-purpose cluster-computing framework. It provides an interface for programming all clusters with implicit data parallelism and fault tolerance.

- **Talend**

  Talend is an open-source data integration platform. It provides many services for enterprise application integration, data integration, data management, cloud storage, data quality, and Big Data.
- **Splunk**

  Splunk is an American company that produces software for monitoring, searching, and analyzing machine-generated data using a Web-style interface.
- **Apache Hive**

  Apache Hive is a data warehouse system developed on top of Hadoop and is used for interpreting structured and semi-structured data.
- **Kafka**

  Apache Kafka is a distributed messaging system that was initially developed at LinkedIn and later became part of the Apache project. Kafka is agile, fast, scalable, and distributed by design.

Enabling technologies play a crucial role in the successful implementation and management of big data solutions. These technologies help organizations collect, store, process, analyze, and visualize large volumes of data. Here are some key enabling technologies in the big data landscape:

1. Distributed Storage Systems:

- Hadoop Distributed File System (HDFS):A scalable and distributed file system designed to store vast amounts of data across multiple nodes in a Hadoop cluster.

2. Processing Frameworks:

- Apache Hadoop MapReduce: A programming model and processing engine for distributed processing of large data sets.

- Apache Spark: A fast and general-purpose cluster-computing framework that supports in-memory data processing for iterative algorithms and interactive data analysis.

3. Data Integration:

- Apache Kafka: A distributed streaming platform that enables the creation of real-time data pipelines and streaming applications.

- Apache NiFi:An open-source data integration tool that provides an intuitive interface for designing data flows and automating data movement between systems.

4. NoSQL Databases:

- MongoDB, Cassandra, Couchbase:NoSQL databases are often used for handling unstructured or semi-structured data and provide horizontal scalability.

5. Data Warehousing:

- Amazon Redshift, Google BigQuery, Snowflake: Cloud-based data warehousing solutions that allow organizations to store and analyze large volumes of structured data.

6. Machine Learning and Analytics:

- Apache Flink, TensorFlow, PyTorch: Frameworks for implementing machine learning models and analytics on large datasets.

7. Data Governance and Security:

- Apache Ranger, Apache Atlas:Tools for managing and securing data access within a big data ecosystem.

8. Cloud Computing:

- Amazon Web Services (AWS), Microsoft Azure, Google Cloud Platform (GCP):Cloud providers offer scalable infrastructure and managed services that facilitate the deployment of big data solutions without the need for significant upfront investments.

9. Containerization and Orchestration:

- Docker, Kubernetes: Containers help in packaging and deploying applications, while orchestration tools like Kubernetes assist in managing containerized applications at scale.

10. Data Visualization and BI Tools:

- Tableau, Power BI, QlikView: Tools that enable organizations to create interactive visualizations and derive insights from big data.

11. SQL-on-Hadoop:

- Apache Hive, Apache Impala: Tools that allow SQL queries to be executed on Hadoop data, making it easier for users familiar with SQL to interact with big data.

12. Edge Computing:

- Apache Edgent, Microsoft Azure IoT Edge:Technologies that bring computing resources closer to the data source, reducing latency and improving real-time analytics for IoT and other edge use cases.

These enabling technologies work together to create a comprehensive big data ecosystem, providing the tools and infrastructure needed to handle the challenges posed by large and diverse datasets. The choice of specific technologies depends on the requirements and goals of individual organizations.