**[Big Data](https://www.simplilearn.com/tutorials/big-data-tutorial)**

“[Big Data](https://www.simplilearn.com/tutorials/big-data-tutorial) are high volume, high velocity, or high-variety information that require new forms of processing to enable enhanced decision making, insight discovery, and process optimization.”

In simple words a collection of huge amounts of complex data that need advanced computing techniques for processing. It comes from variety of sources including online transactions, sensors, digital devices, and social media platforms.

**Characteristics Of Big Data**

Big data can be described by the following characteristics:

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

**1.Volume:**

We consider any data as Big Data or not, depending upon the volume of data. Hence, ‘Volume’

is one characteristic that needs to be considered while dealing with Big Data solutions.

**2.Variety:**

Data used to be given in a single format from a single source. Previously given in database files like excel, csv, and access files, it is now being delivered through tech like wearable devices and social media in non-traditional formats, including video, text, pdf, and graphics. Although this data is helpful to us, it demands more labor and analytical abilities to interpret it, manage it, and make it function.

**3.Velocity:**

The term ‘velocity’ refers to the speed of the generation of 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 devices, etc. That data will come in batches or it may arrive in fits and starts.

**4. Veracity:**

The correctness of your data is referred to as veracity. The accuracy of your findings can be severely harmed by poor veracity, making it one of the most crucial big data qualities. It specifies the level of data reliability. It is vital to remove the information that is not essential and use the remaining data for processing because most of the data you encounter is unstructured.

**5.Value:**

Value is one`s of the most crucial fundamentals of big data. After data collection it is cleaned to make it valuable and insightful for decision making purposes.

**How Does Big Data Work?**

**1.Data Collecting**

Every company collect data from variety of sources, including cloud storage, mobile apps, in-store IoT sensors, and more.

**2. Organize the Data**

For analytical queries to yield correct answers, data must be appropriately organized once gathered and stored.

**3. Clean Data**

At this stage all data must be clean, through removing Duplicate or unnecessary data and structured it appropriately.

**4. Analysis of Data**

At his stage huge amounts of data is transformed into a usable form to get significant insights by applying different analytics techniques.

**Types Of Big Data**:

**1.Structured Data**

Structured data is in the form of tabular schema and is easy to organize.

**2.Unstructured data**

Unstructured data lacks predetermined conceptual meanings and is difficult for conventional databases to comprehend or analyze. Example; Video and audio files, mobile activities, satellite photos, and other types of big data

**3.Semi-Structured Data**

A combination of unstructured and structured data is semi-structured data having characteristics of both. semi-structured data is frequently seen in JSON and XM.

**Structured vs Unstructured Data:**

|  |  |
| --- | --- |
| Structured Data | Unstructured Data |
| Structured data is highly specific. | Unstructured data is a conglomeration of many varied types of data e.g., images, audio, video, etc. |
| It is stored in a predefined format in the form of table having rows and columns.e.g.csv file, text fi les, spreadsheets, etc. | It is stored in their native formats |
| It employs schema-on-write. | Unstructured data employs schema-on-read. |

**Challenges of Big Data**

1. One of the issues with Big data is the exponential growth of data, organizations often find it difficult to rightly store this huge amount of data.
2. Other challenge is choosing the right Big Data tool.
3. Next challenge of Big Data is securing it. Often organizations are too busy understanding and analyzing the data, that they leave it and hackers take benefit from this.

**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
2. Improved customer service
3. Early identification of risk to the product/services, if any
4. Better operational efficiency

**Database:**

A collection of logically related information organized so that it can be easily accessible, managed, and updated. It has predefined schemas on write.

**Types of DBMS**

There are four different types of DBMS which are mentioned below

1. Hierarchical database
2. Network Database
3. Relational Database
4. Object Oriented Database

**Examples of Database Management System**

Following are some of the popular database management systems in use:

1. MySQL
2. Oracle
3. MS-Access
4. SQLite

**Objectives of Database Management System:**

Database management system has the following objectives:

1. Elimination of redundant data
2. Enabling data access easy for the user
3. Providing a source of mass storage of data
4. Allowing multiple users to access the database at the same time
5. Providing appropriate and quick response to user queries

**Functions of Database Management System:**

Following are the functions of the database management system:

1. Provide a mechanism for storing, organizing and retrieving data
2. Reducing redundancy of data
3. Providing facility for preventing unrestricted access to database by means of passwords
4. Creating data and program independence such that one can be changed without changing the other

**Data Warehouse:**

 Data warehouses serve as a central repository for storing, analyzing and interpreting information to make better informed decisions. An organization's data warehouse receives data from a variety of sources, typically on a regular basis, including transactional systems, relational databases, and other sources.

**Data Warehousing Tools:**

Some popular data warehouse tools are Xplenty, Amazon Redshift, Teradata, Oracle 12c, Informatica, IBM Infosphere, Cloudera, and Panoply.

**How Data Warehouse Works?**

Data Warehousing integrates data and information collected from various sources into one comprehensive database. For example, A company wants to analyze How many employees over achieved their sales target within 1 year of joining the company?

The first thing needed is employee details, secondly the sales which would be generated by that employee. A data warehouse combines employee information getting from the HR database, with the sales information getting from the sales database to make a performance dataset where we combine all those data to create a summary report of the top performer of the company which can be further reported to the higher management for evaluation.

**Key Characteristics of Data Warehouse:**

The main characteristics of a data warehouse are as follows:

**1.Subject-Oriented:**

A data warehouse is subject-oriented means it provides specifically provide the information for which it was made. For example, if you want to analyze your company’s sales data, you need to build a data warehouse that concentrates on sales. Such a warehouse would provide valuable information like ‘who was your best customer last year?’ or ‘who is likely to be your best customer in the coming year?’

**2.Integrated:**

A data warehouse is developed by integrating data from varied sources into a consistent and structured format for effective data analysis.

**3.Non-Volatile:**

Data once entered into a data warehouse must remain unchanged. All data is read-only. Previous data is not erased when current data is entered. This helps you to analyze what has happened and when.

4.**Time-Variant:**

 Mean that whatever data you store in the data warehouse varies with time means you can do analysis based on time study, e.g., you can do the analysis of your employee’s performance for the past 10 years then all that analysis can be done within data warehouse.    
**Functions of Data Warehouse:**

 It works as a collection of data and here is organized by various communities that endures the features to recover the data functions. It has stocked facts about the tables which have high transaction levels which are observed so as to define the data warehousing techniques and major functions which are involved in this are mentioned below:

**Data Consolidation:**

The process of combining multiple data sources into a single data repository in a data warehouse. This ensures a consistent and accurate view of the data.

**Data Cleaning:**

The process of identifying and removing errors, inconsistencies, and irrelevant data from the data sources before they are integrated into the data warehouse.

**Data Integration:**

The process of combining data from multiple sources into a single, unified data repository in a data warehouse for analysis. It ensures that the data is accurate and usable for analysis.

**Data Storage:**

A data warehouse can store large amounts of historical data and make it easily accessible for analysis.

**Data Transformation:**

Data can be transformed and cleaned to remove inconsistencies, duplicate data, or irrelevant information.

**Data Analysis:**

Data can be analyzed and visualized in various ways to gain insights and make informed decisions.

**Data Reporting:**

A data warehouse can provide various reports and dashboards for different departments and stakeholders.

**Data Mining:** Data can be mined for patterns and trends to support decision-making and strategic planning.

**Performance Optimization:** Data warehouse systems are optimized for fast querying and analysis, providing quick access to data.

**Why do you Need Data Warehouses?**

Data warehouse is important because of the following reasons:

**1.Consistency:**

It Ensure consistency which makes it easier for corporate decision-makers to analyze and share data insights with their colleagues around the globe. It also reduces the risk of error in interpretation and improves overall accuracy

**2.Centralization:**

All the data is stored, managed and maintained in one single centralized repository eliminating data quality, inconsistent data, inaccurate data reports and low query performance issues.

**3.Make better decisions, faster:**

warehousing improves the speed and efficiency of accessing different data sets and makes it easier for corporate decision-makers to derive insights from that and make better business decisions.

**4.Improves Organization`s Progress:**

Data warehouse keep the records of organization's historical activities, therefore, helping executives to evaluate their initiatives that have been successful or unsuccessful in the past and adjust their strategy according to the situation to decrease costs, maximize efficiency, and increase sales to improve their bottom line.

**Benefits of Data Warehouse:**

There are several [benefits of data warehouse](https://www.simplilearn.com/benefits-of-data-warehousing-article) for end users.

1. Improved data consistency
2. Better business decisions
3. Easier and faster access to enterprise data for end-users
4. Data centralization
5. higher productivity

**Drawbacks of Data Warehouse:**

**1.Hidden problems with source systems**

Sometimes hidden problems associated with the source systems feeding the data warehouse may be identified after years of being undetected. For example, when entering the details of a new property, certain fields may allow nulls which may result in staff entering incomplete property data, even when available and applicable.

**2.Required data not captured:**

In some cases, the required data is not captured by the source systems which may be very important for the data warehouse purpose. For example, the date of registration for the property may be not used in the source system but it may be very important for analysis purpose.

**Data Lake:**

A data lake is a centralized repository for storing, and processing large amounts of data in its raw form, including structured, unstructured, and semi-structured data at any scale.

**Why do you need a data lake?**

We need data lake for storing all types of data like log files, data from click-streams, social media, and internet-connected devices. Then perform different analytics on that data to generate business value from data. This helped organizations to identify and act upon opportunities for business growth faster by attracting and retaining customers, boosting productivity, proactively maintaining devices, and making informed decisions.

**Benefits:**

1. Increase operational efficiencies
2. A data lake makes it easy to store, and run analytics on machine-generated IoT data to discover ways to reduce operational costs, and increase quality.
3. It Helps for fast ingestion of new data.

**The Challenges of Data Lakes:**

The main challenge with a data lake architecture is that raw data is stored with no oversight of the contents. For a data lake to make data usable, it needs to have defined mechanisms to catalog, and secure data. Without these elements, data cannot be found, or trusted resulting in a “data swamp." Meeting the needs of wider audiences require data lakes to have governance, semantic consistency, and access controls.