

Sheroze Rehman

TASK 1

~~Sheroze R~~

BIG DATA:

Data is collection of facts & figure and big data can be referred as large and complex data that can't be easily processed. These data are mainly generated from social media, IoT devices.

each day Peta Byte of data is generated. They can only be analyze through machine learning tools as they are not in schema format.

It is characterized into 3V's

volume (large volume of data)

variety (data type text, video, audio, image)

velocity (fast moving & access)

DATA LAKE:

Lake refers to small minerals of ^{alone ph} water in the same way data lake refers as a central repository that stores raw & unprocessed. They are mainly from, social media & IoT devices.

The data stored in data lake are stored in original format without being worry of how it been used in future

The main goal is to provide a single source of truth for all data in the organization

They are scalable as they can accommodate growing amount of data over time

Data lake store data in big data they need specialized tool to analyze the data and make analysis

Currently Data lake is offered by IBM, Google, AWS

DATABASE:

Database is a structured collection of data that is stored in either computer system or other database platform on web server.

They are mainly used in inventory, management, user record, financial record. They can be relational or non-relational.

Mostly SQL is used to access and manipulate the database.

DATA WAREHOUSE:

A data warehouse is a centralized repository that store data from various sources in an organized format. It uses schema structure to store data.

They are used mostly is transaction related databases.

It contains historical data in form of table so that they can be easily queried and analyze using SQL

When loading data in to data warehouse it have to pass through ETL model (extract, transform, load) to populate warehouse and define security measure