DATA ENGINEERING

Data Engineer:

They build and scale the platforms that enable data collection, processing and

storage for data science/business analytics use.

Data Scientist:

They use linear algebra and multivariable calculus to create new insight from

existing data.

DATA CLASSIFICATION:

Raw Data:

Unprocessed data in format used on source e.g JSON

No schema applied

Processed data:

Raw data with schema applied

Stored in event tables/destinations in pipelines

Cooked data:

Processed data that has been summarized.

BIG DATA PROPERTIES:(4 V's)

Volume:

How much data you have

Velocity:

How fast data is getting to you

Variety:

How different your data is

Veracity:

How reliable your data isHow much data you have

DATA PROCESSING METHODS

BATCH PROCESSING:

Batch processing is a data handling method where tasks are processed in groups at scheduled intervals rather than in real-time, optimizing efficiency for large-scale data processing.

STREAM PROCESSING:

Process data on the fly, as it comes in  
  
  
STREAMING METHODS:  
At Least Once:

In this delivery guarantee, a message or operation is ensured to be processed and delivered at least once to the intended recipient.

At Most Once:

This guarantee ensures that a message or operation is processed and delivered at most once.

Exactly Once:

This is the most stringent guarantee, ensuring that a message or operation is processed and delivered exactly once, without duplication or loss.

Data warehousing

Data Warehouse (DW) is a Subject oriented, integrated, time variant, non-volatile collection of data in support of management’s system.

It is a collection of data designed to support management decision making by presenting a coherent picture of business conditions at a single point of time.

Features of Data Warehouse:-

Time-variant:

The data warehouse encompasses a longer time span compared to operational systems, offering historical insights over an extended period, typically ranging from the past 5 to 10 years.

Non-volatile:

Once data is incorporated into the data warehouse, updates are prohibited, ensuring permanence. The warehouse preserves the company's history, continually appending operational data reflecting recent events.

Subject-oriented:

Data are organized according to the subject instead of application.

It mainly focuses on modeling and analysis of data for decision makers, not on daily operations or transaction processing.

Integrated:

Constructed by integrating multiple, heterogeneous data sources like relational databases, flat files, on-line transaction records.

Ensure consistency in naming conventions, encoding structures, attribute measures, etc. among different data sources.

DSS architectural styles:

OLTP (Online Transaction Processing)

-used by traditional operational systems (RDBMS).

OLAP (Online Analytical Processing)

-used by Data Warehouse.

An Operational Database is accessed by an Operational System to perform routine organizational operations. Typically employing an OLTP architecture for swift transaction processing, these databases interact with data through operations such as insertion, deletion, and updating.

OLTP:

OLTP is a method designed to grant end users access to extensive data, operating in an intuitive and speedy manner to aid deductions through investigative reasoning. It denotes a category of systems that support and oversee transaction-oriented applications, usually for the processing of data entry and retrieval transactions.

Benefits of OLTP:

1.Simplicity & Efficiency

Examples of how OLTP streamlines business processes include minimizing paperwork and improving the accuracy and speed of revenue and expense forecasts. OLTP systems ensure data integrity and deliver rapid query processing, particularly in environments with multiple access points.

Pitfalls of OLTP:-

OLTP requires instant update.The data what we get from OLTP is not suitable for data analysis.To perform one simple transaction even with the normalized structure, we need to query multiple tables by using joins.