Data Engineering -

* Design and scaling the data for future analytics

ETL: -

* Extract- data gathered from various sources
* Transform-Removing or scaling the data
* Load- The transformed data is loaded into a data warehouse
* Examples: - Getting data from IOT devices.

Data Science Platform Blueprint:

connect -- API, dataflow apps, external Data Warehouse, SQL DB

Store -- BIG DATA, SQLDB -- processing framework-- stream, batch Buffer – Cache, messages

Visualize -- web Ui's, mobile apps, BI tools

Classifications Of Data:

RAW DATA

The raw data is the data which has all information including the unprocessed ones.

PROCESSED DATA

* Processed data is data where schema is applied like tables and pipelines

COOKED DATA

* processed data when summarized used for analytics

BIG DATA:

To store all the Kind of Datas we need BIG DATA and in very large format like 100s of tbs

Important Properties in Big Data:

* VOLUME-how much data you have
* VELOCITY-how fast the data getting to you
* VARIETY-how different is your data
* VERACITY-how reliable is your data

DATA PROCESSING METHODS

Batch Processing:

These frameworks are designed to process large volumes of data in batches.

Data are stored in a Storage and later to measure the data we need to Analysis.

After analysis the Insight of the required data that will be in dashboard, table etc.

Stream Processing:

stream processing frameworks handle data in real-time as it is generated.

Example: YouTube streaming

Data Storage:

* Relational and non-relational databases (SQL/NoSQL)

PROCESSING FRAMEWORKS:

Big Data Works on Map Reduce

* Maps reduce key-value pairing
* organize the data into key and values sort the key
* combine the data with the matching key
* repeat the process you have the final key value outcome.

Tools:

Hadoop, Azure data bricks, Apache spark, Samza , beam

An introduction to Data Warehousing:

* It is a Subject oriented, integrated, time variant, non-volatile collection of data in support of management’s system.
* Data Warehousing is used for data reporting & analysis.
* The data stored in the warehouse is uploaded from the operational system.

Purpose of Data Warehouse:

* Subject-oriented:
  + Data is organized according to the subject instead of application.
  + focuses on modeling and analysis of data.
* Integrated:
  + Constructed by integrating multiple, heterogeneous data sources like relational databases, flat files, on-line transaction records.
* Time-variant:
  + The time horizon for the data warehouse is significantly longer than that of operational systems. i.e. provide information from a historical perspective.
* Non-volatile:
  + No updates are allowed.

DSS:

* Decision support system.
* works by compiling useful information from a combination of raw data

DSS Architectural Styles:

* OLTP- Online Transaction Processing (OLTP)-(RDBMS).
* OLAP- Online Analytical Processing (OLAP)-Data Warehouse.

Data Warehouse Architecture:

Operational Data Store

* The data that arrived at data warehouse are first passed to operational data store
* Data is integrated from multiple sources for more operations on the data.  
   OLTP Vs Warehouse Applications

Data Marts:

* The data in the data warehouse is stored in the form of Data marts.
* It allows the user to access the data in terms of a specific business line or team.

Data marts Vs Data Warehouses

* The data mart is a *subset* of the data warehouse that is usually oriented to a specific business line.
* Data Warehouse is a database used for data reporting and analysis.

Eg: products for sale in store and the products in warehouse