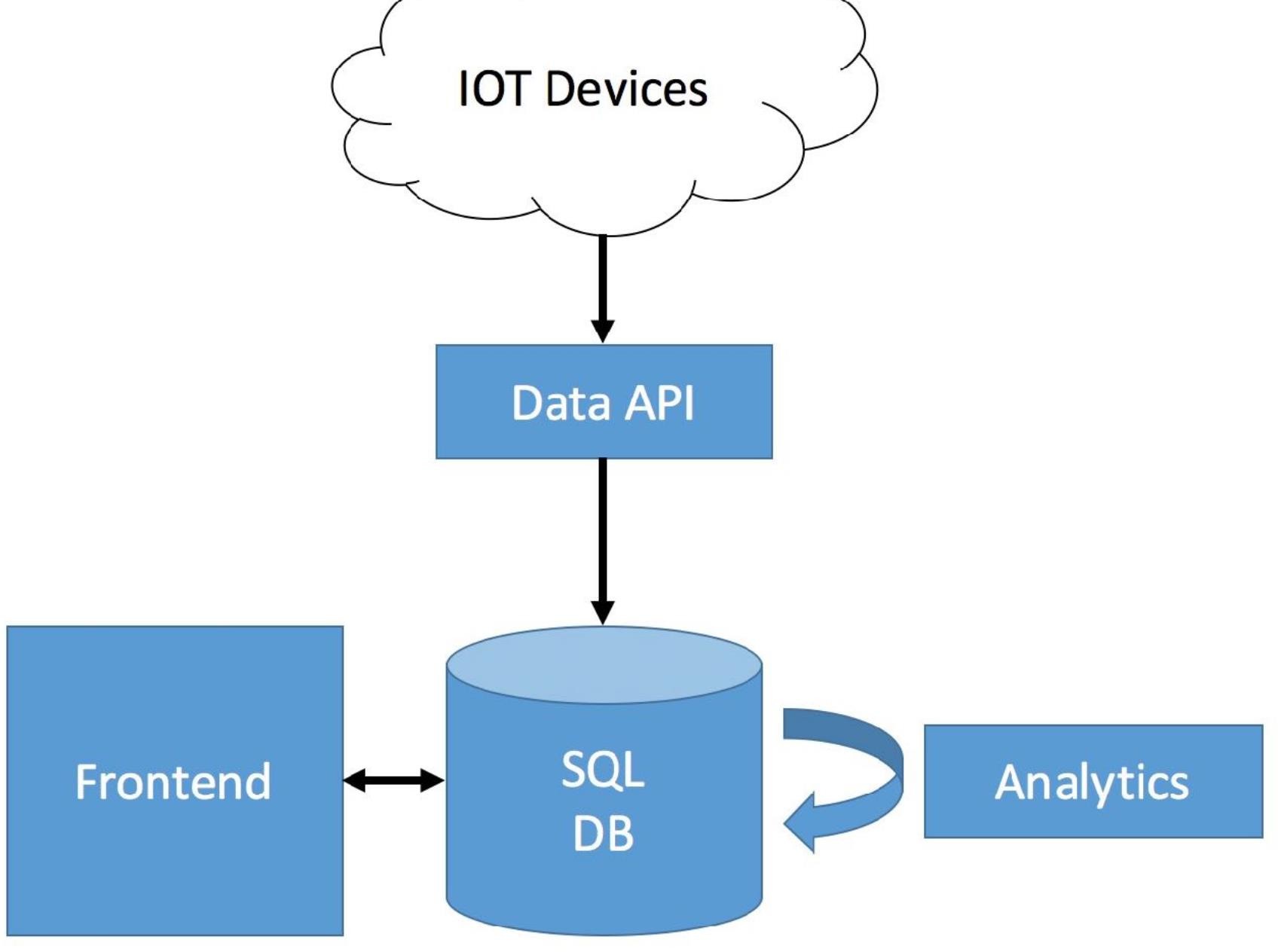
Day 1 Assessment - Vamsi Viswanadham

**Data Engineering:** Data Engineering is defined as the process of **designing, building, and scaling** systems that organize data for analytics (measuring the data).

for that, ETL (extract-transform-load) tools work.

**ETL**: ETL is a process that extracts, transforms, and loads data from multiple sources to a data warehouse or other unified data repository.

**The architecture of ETL**



**DataScience platform blueprint:**

**Connect** (APIS, DataFlow Apps), **External** (Enterprise Data Warehouses, SQL DBs) <-> **Store** (Big Data, SQL DB/Warehouse), **Processing Framework** (Stream, Batch), **Buffer** (Cache, Message Queues) <-> **Visualize** (Web uIs, BI Tools, Mobile Apps)

**Data Classification:**

Below are the three types of Data.

-> **Raw Data** (just as it is, it comes from the data sources): Raw data can be defined as Unprocessed data in the format used on the source, eg: JSON

Here there is no Schema applied.

-> **Processed Data**: Processed data is the Raw data with schema (JSON schema, SQL schema, etc.) applied (unwanted, replicated data will be deleted here)

-> **Cooked Data** (Only necessary data with summarized): Processed data that is summarised.

* providing solutions to store, manage, and scale huge data, data engineering comes into the picture

**Big Data (store all the types of data) Properties**:

**-> 4 V's**

1. Volume: How much data you have (the amount of data we have)
2. Velocity: How fast data is getting to you (the speed of accessing data)
3. Variety: How different your data is (the types of data you have)
4. Veracity: How reliable (how consistent or stable) your data is

**Data Processing Methods**

* **Batch processing:**

Data -> Storage -> Analytics (measuring the data - like how many viewers, etc.) -> Insight (Visualizations, results, etc.)

There will be tools for each component.

* **Stream Processing:** Compared to Batch Processing, here there is no storage. It is just like processing the live data.

Process data on the fly(live), as it comes in.

Below are the Streaming Methods available to be used in Stream processing.

* At least once
* At most once
* Exactly once

Big data works on map-reduce (key-value pairing)

**Steps involved in the Map Reduce Process:**

1. Organize the data into keys and values
2. sort by the key
3. combine the data with matching keys
4. Repeat until you have the final key-value outcome.

Tools used for Big Data: Hadoop, Spark, samza, beam

**Data Warehouse**

* A subject-oriented, integrated, time-variant, non-volatile(not deleted) 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 in time.
* bulk number of machines used to store data

**DataWarehouse:** database used for data reporting and analysis.

-> Business Data -> snapshot of company's situation(from operational databases, external sources)

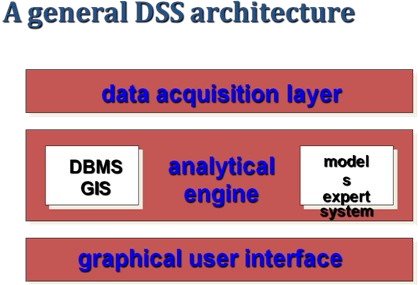
-> Business data model

**Properties of DW:**

* **Subject Oriented** (organized according to subject instead of an application, mainly focuses on modeling and analysis of data for decision makers not on daily operations or transaction processing)
* **Integrated** (integrating multiple heterogenous data sources like databases, flat files, records, etc., this property ensures consistency in naming conventions, encoding structures, attribute measures, etc. among different data sources)
* **Time variant** (time is significantly longer than that of os i.e., provide info from a historical perspective (eg., past 5-10 years))
* **Non-Volatile** (no updates are allowed, once data is entered into DW, they are never removed, only added)

**DSS(Decision Support Systems)** -> helps to assess and resolve everyday business questions.

-> It works by compiling useful information from a combination of raw data, documents, personal knowledge, or business models.



**Structured (relational) and Unstructured (nonrelational) Components**

1. A Structured component is one that directly helps us to proceed toward a decision.
2. An Unstructured component is that which is to be still processed and requires human interaction with the DSS.

**DSS Architectural Styles:**

* OLTP: Online Transaction processing -> Used by traditional OS -> RDMS
* OLAP: Online Analytical Processing -> Used by Data Warehouse.

**Operational DB:** use an OLTP architecture, accessed by the Operational system to carry out regular operations of an organization.

**OLTP**

**Definition:** methodology to provide end users with access to large amounts of data

Transaction-oriented applications.

Eg: An ATM for a bank is an example of a commercial transaction processing application.

**OLTP Architecture:** Raw Audit data table -> Warehouse -> Auditors for reporting, analysis, and mining.

**Benefits of OLTP:**

* Simplicity (Simple to use and implement)
* Efficiency (faster and more accurate operations)
* Data integrity (more integrated data)
* Fast query processing (useful in case of environments having multiple access).

**Pitfalls of OLTP:**

* requires instant update
* not suitable for data analysis
* need to query multiple tables

***DSS Data vs Operational data***

* time span (Shorter term in Operational, Longer term in DSS)
* granularity (detailed and transaction-level in Operational, Less granular in DSS)
* Dimensionality (lower dimensionality in Operational, higher dimensionality in DSS)

**ETL uses the below steps to house its key functions.**

* Staging (the area where data is at rest)
* Integration (Integrating from/to multiple data sources)
* Access layers (Data Marts)

**Procedure Involved in Operational DB:**

1. The data that arrive at the data warehouse are first passed to the Operational Data Store (ODS).
2. Data is integrated from multiple sources for additional operations on the data.
3. This integrated data is passed back to operational systems for decision-making.