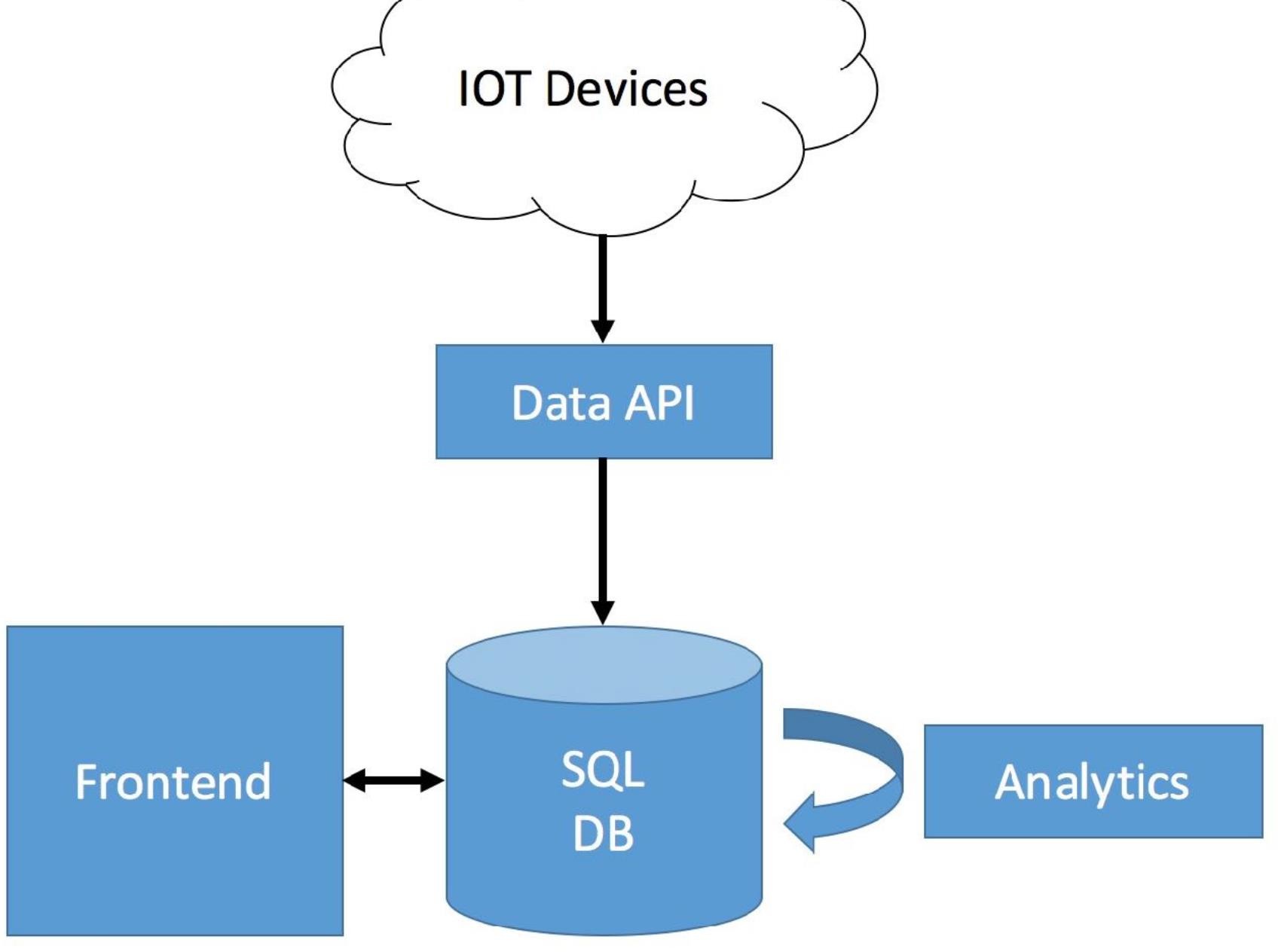
Data Engineering -

* Design, building & scaling the organize data for analysis.

ETI: -

* **Extract**- data is collected and extracted from various sources
* **Transform**- Once the data is extracted, it undergoes transformation processes
* **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**

* unprocessed data in the format of Json
* without any schema

**PROCESSED DATA**

* raw data applied with schema is processed data
* stored on events tables etc.

**COOKED DATA**

* processed data when summarized used for analytics

**BIG DATA:**

To store all the Kind of Datas we need BIG DATA.

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

**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

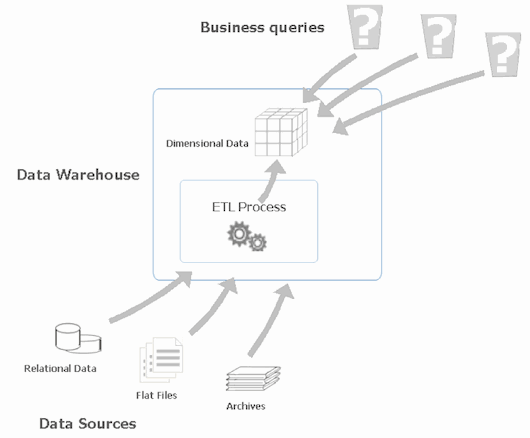
**An introduction to Data Warehousing:**

* 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.
  + It mainly focuses on modeling and analysis of data for decision makers.
  + Example- retail warehouse, there might be a subject area focused on sales, containing data related to products sold, customer transactions.
* **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. Once the data entered the data warehouse, they are never removed.
  + Example- In an educational institution's data warehouse, student enrollment data for past semesters remains unchanged. Even if there are updates to the source systems, the data in the warehouse preserves the state it was in when initially loaded.

**Data Warehouse Architecture:**

  
 **Operational Data Store**

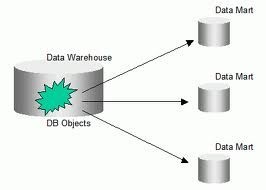
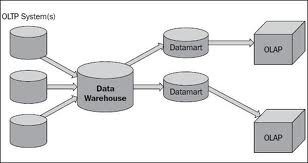
* The data that arrived at data warehouse are first passed to operational data store
* Data is integrated from multiple sources for more operations on the data.  
   OLTP Vs Warehouse Applications

**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 Warehouseis a database used for data reporting and analysis.
* The data stored in the warehouse is uploaded from the operational systems.

**Data Warehouse Life cycle:**

