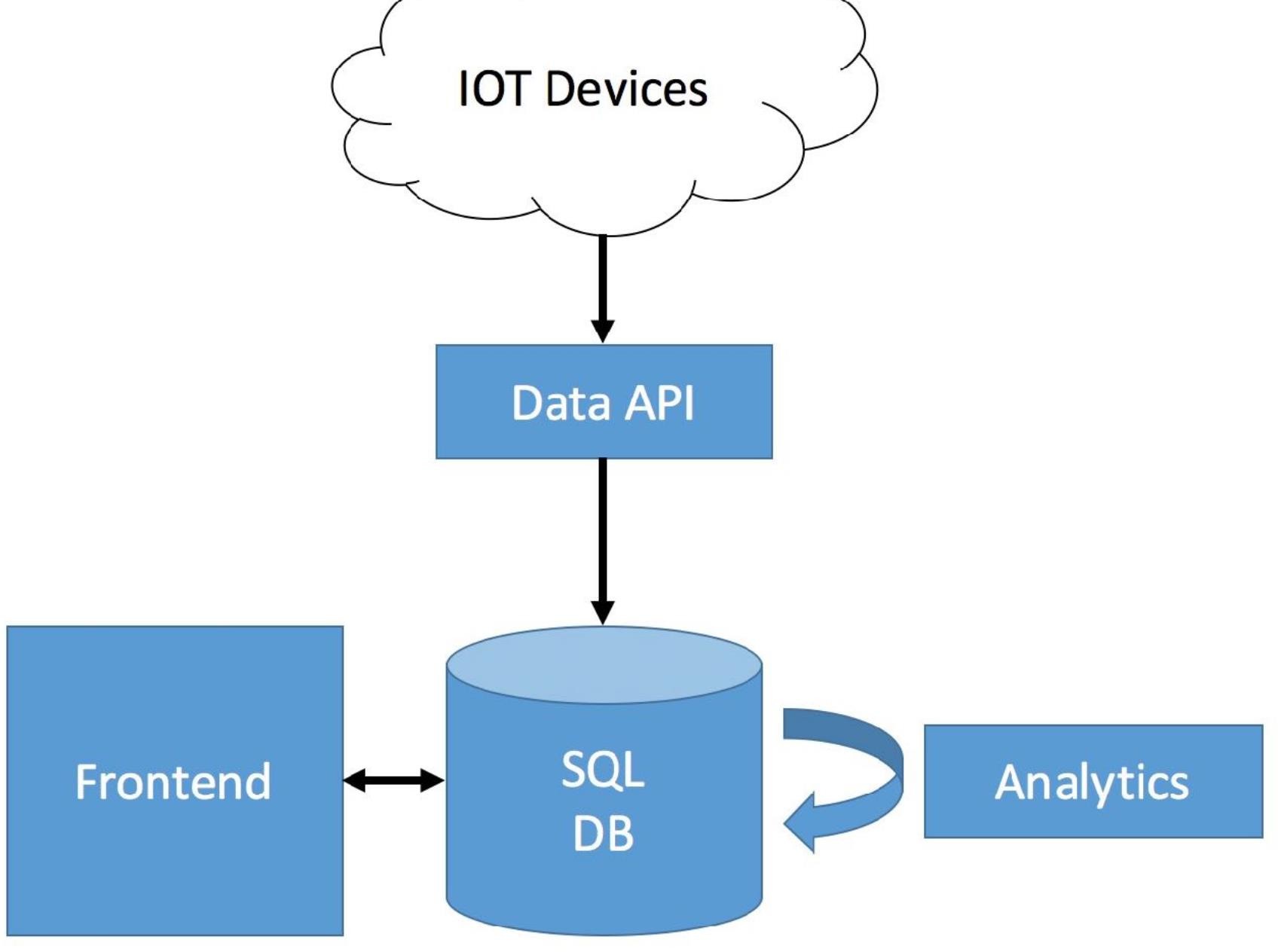
**DATA ENGINEERING:**

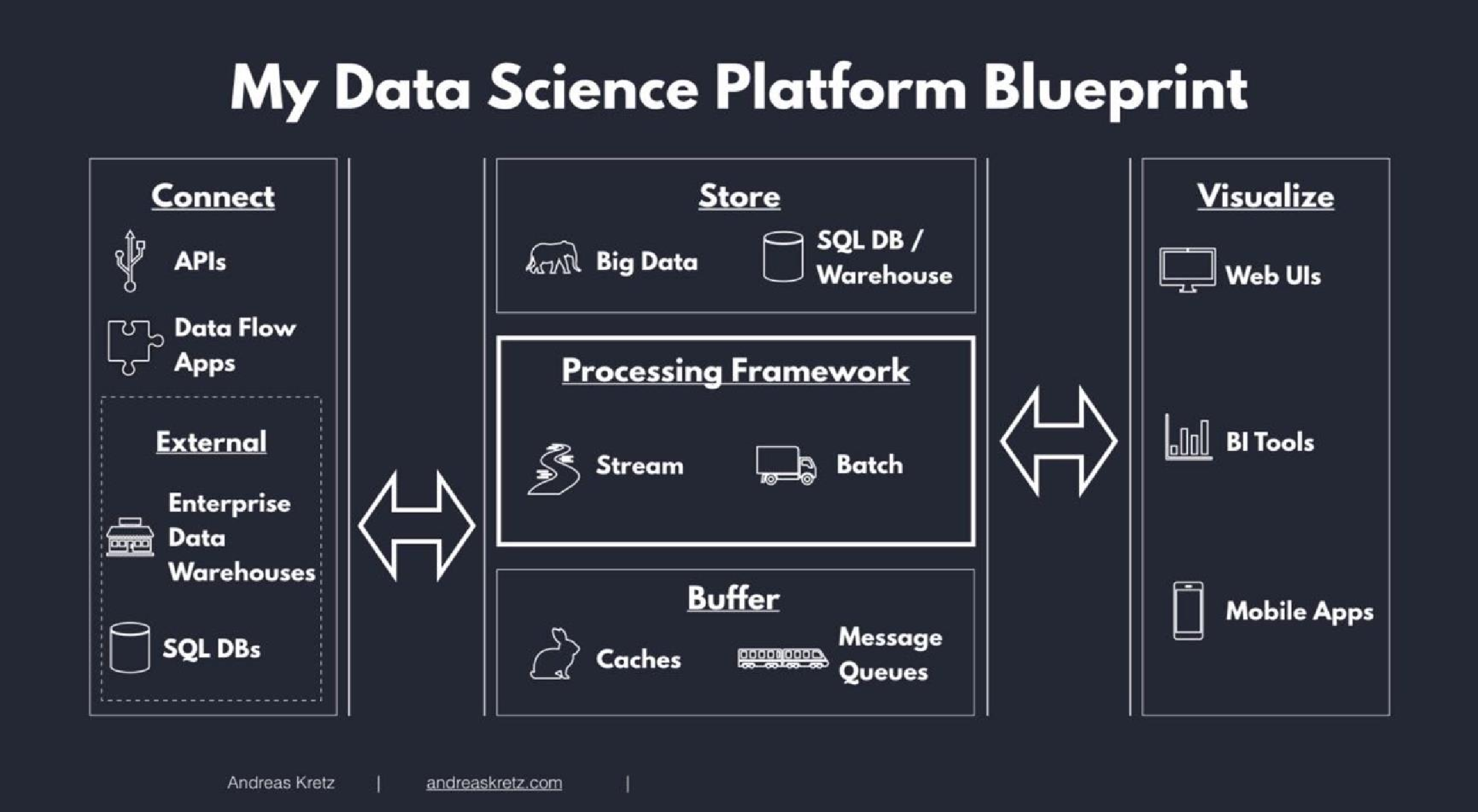
Designing, building and scaling systems that organize data for analytics.

**ETL** 🡺 Extract, Transform, Load

**Architecture of ETL:**



**Data Science Platform Blueprint:**



**Data Classification:**

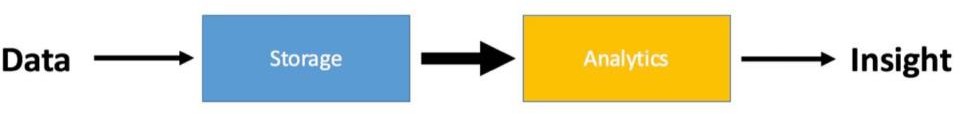
1. Raw data 🡺unprocessed data
2. Processed data 🡺 raw data + schema
3. Cooked data 🡺 processed data +summarize

**Big Data Properties: (V’s of BIG DATA)**

1. Volume 🡺 How much data you have
2. Velocity 🡺 speed of data
3. Variety 🡺 Types of data
4. Veracity 🡺 reliable of data

**Data processing methods:**

1. Batch processing



1. Stream processing 🡺 process data on the fly, as it comes in

Streaming methods🡺 i. At least one

ii. At most one

iii. Exactly one

**Map Reduce:**

It is main part of big data.

It has key – value pairing.

Sort by key.

Ex: Hadoop, Apache Spark, Beam, Samza.

**Data storages:**

1. Relational Database(SQL) 🡺 Table format
2. Non Relational Database(NoSQL) 🡺 Not in a table format

**Data warehousing: (DW)**

This place helpful for storing large amount of data in one place and easily accessible for performing tasks.

This is the place keep all kind of data like raw data, relational data, etc.

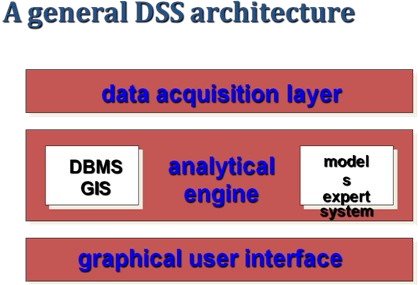
**Features of Data warehousing:**

* Subject oriented (Data organized according to subject)
* Integrated (Integrated with heterogeneous data)
* Time variant (It provides long life (5 to 10 years))
* Non – Volatile (No updates are allowed, once the data entered into warehouse)
  + - It useful for know company history.

**Decision support system: (DSS)**

It helpful for taking decisions on warehousing, databases or an applications.

**DSS Architecture:**



**DSS architectural styles:**

1. OLTP (Online Transaction Processing) 🡺 used in RDBMS
2. OLAP(Online Analytical Processing) 🡺 used in data warehouse

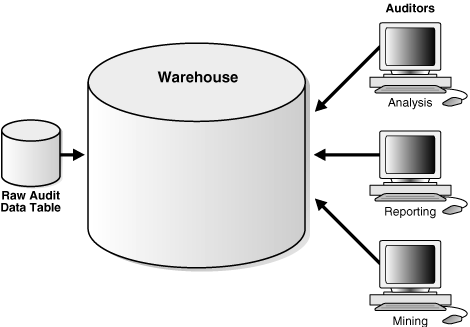
**Operational Data:**

The Operational Database is one which is accessed by an Operational System to carry out regular operations of an organization.

It uses OLTP for faster transaction processing.

It performs the Inserting, Deleting, Updating.

**OLTP Architecture:**



**Example :** ATM Machines uses this type.

**Benefits of OLTP:**

1. Simplicity
2. Efficientcy
3. Data integrity
4. Fast query processing

**Pitfalls of OLTP:**

1. OLTP requires instant updates.
2. Not suitable for data analysis.

Data store contain 2 types of data:

1. Business data 🡺 operational data + external data
2. Business data model

**Data Mart:**

The data mart is a subset of the data warehouse that is usually oriented to a specific business line or team.