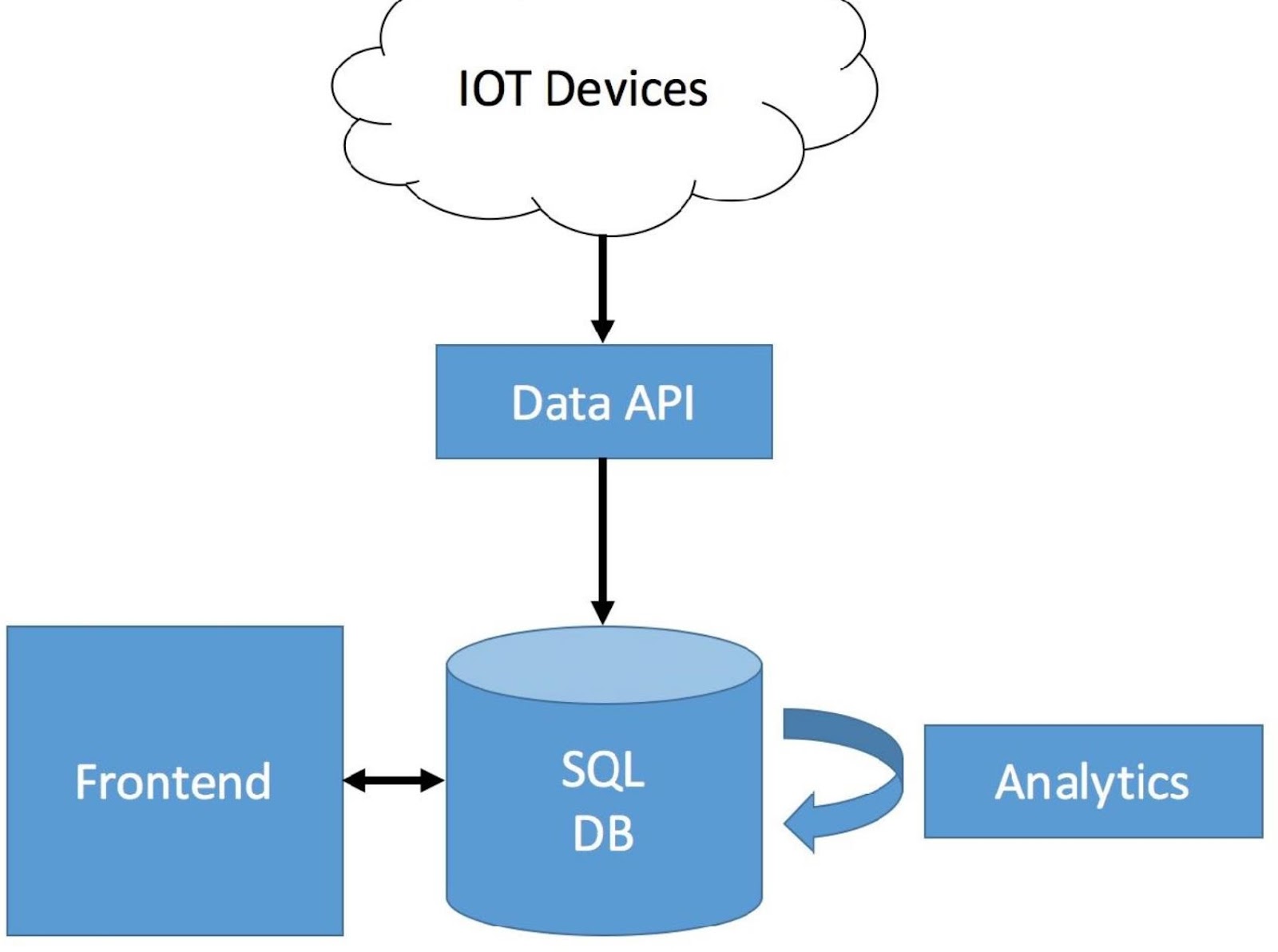
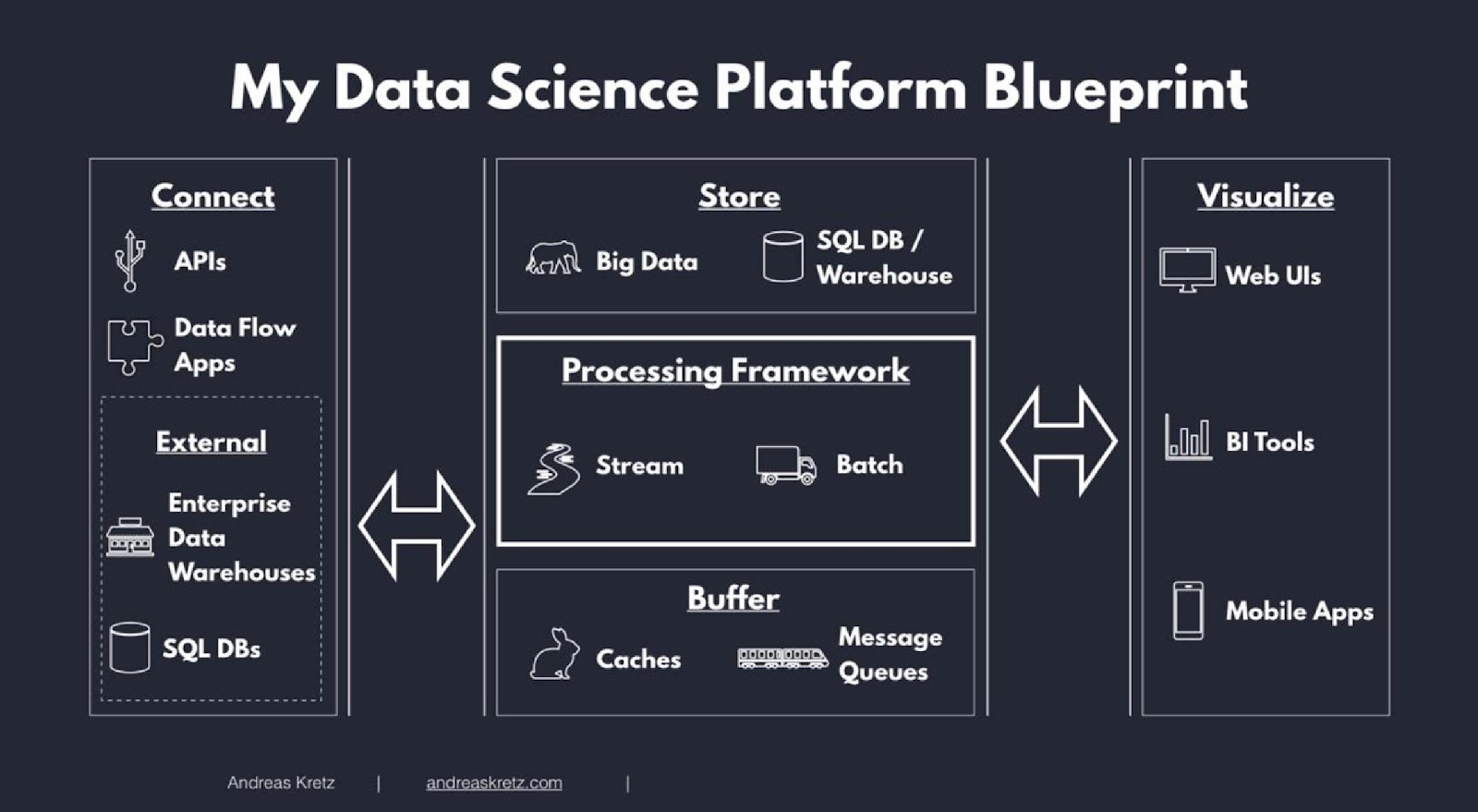
**DAY-1**

**Data Engineering-** The process of designing, building and scaling the systems that let people collect and analyse raw data from multiple sources and formats in order to organize the data for analytics.

**Basic Architecture Of ETL:**



**My Data Science Platform Blueprint:**



**Data Classification:**

**Raw Data:**

* Here unprocessed data is present used on source and no schema is applied over here.

**Processed Data:**

* Here we will apply schema to the raw data.
* We store in event tables or destinations in pipelines.

**Cooked Data:**

* The Summarization of data is present here which is processed.

Whenever we want to store a vast amount of data we go for BIG DATA.

We can describe the bigdata in **4V’s.**

**Volume**: describes how much data you have.

**Velocity**: describes that how fast the data is getting to you.

**Variety**: describes that how different that your data is.

**Veracity:** describes that how reliable your data is.

**Batch Processing:**

Data, Storage, Analytics and insights. For each of these we will use individual tools.

**Stream Processing:** Process the data on the fly, as it comes in. It involves atleast once, atmost once and exactly once.

**Map-Reduce:**

It involves Key-Value Pairing. Here we will

* We organize our data into keys and values,
* We Sort our data by make use of the key,
* Combine the data with matching Keys.
* Repeat until you have the final key-value outcome.

**Data Warehouse Architecture:**

**Data Warehouse:** It is a central repository of information that can be stored and analysed to make more informative decisions. Generally Data flows into a data warehouse from transactional systems, relational databases, and other sources also.

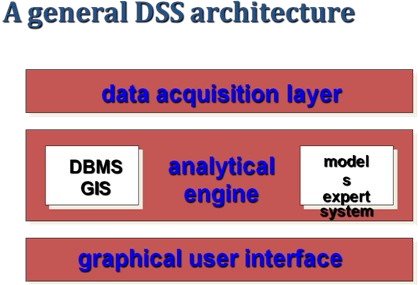
**Features of Data Warehouse:**

* **Subject-oriented**-Here we organize the data by make use of subject or content-wise instead of application. Here we do modelling and analysis of data for decision makers, not on daily operations or transaction processing.
* **Integrated**- Integrated means combining the data from all resources. It may include cloud platforms, relational databases, flat-files etc. we have to make sure Make sure that names, ways of organizing, and how we measure things are the same across all the different places we get data from.
* **Time-variant-** Data warehouses are time-variant, which means they maintain historical data and provide a record of changes over time. This is essential for tracking trends, making comparisons, and conducting historical analyses.
* The ability to store and manage historical data allows decision-makers to identify patterns, monitor performance, and make informed decisions based on a comprehensive understanding of the data's evolution over time.
* **Non-volatile:** Data warehouses are non-volatile, meaning once data is stored in the warehouse, it is not frequently updated or deleted. Instead, data is loaded in periodic batches through processes like Extract, Transform, Load (ETL). This non-volatility ensures the stability and consistency of the data for analytical purposes. Users can rely on the data warehouse for consistent and stable information without the risk of constant changes that may occur in operational databases.

**Need for Decision Support System (DSS) in business:**

In this competitive world, in order to succeed our business we have to make correct decisions to improve it. For that we need Decision Support System.

To do well in business nowadays, a company must have computer systems that can help with all kinds of information and decision-making.



**DSS involves Structured and Unstructured Components.**

**Structured Components** are nothing but relation databases which are concisely useful for taking correct decisions.

*Example*: If we want to predict diabetes in a person ,we have to check different attributes like age, insulin level, haemoglobin level in blood in a table format.

**Unstructured Components** are the components where human interaction needs to make it taking decisions.

*Example:*

Patient needs correct Body Mass Index and glucose level in body.

By using the above statement we cannot able to find the person has diabetes or not.

In order to implement DSS ,we have

**DSS Architectural Styles,**

**OLTP:** which stands for Online Transaction Processing System used to managing and processing day to day transactional operations in a business.It may used for real time for processing data, data integrity etc.

Traditionally it is used in RDBMS.

*Example:* Banking System.

***Advantages :***

**Real time Transaction Processing:** we can use these in our daily life in order to make transactions.

Data Integrity and Fast Query Processing.

***Disadvantages:***

**Limited Analysis *.***It cannot perform transactions with large data sets fastly. It needs instant updates.

**Redundancy in data**: It need normalisation still has redundancies which results inefficient data accessing.

**Slow in executing Queries:** If we are performing with complex queries it gives slow response.

In order to overcome all these, OLAP came into picture.

**OLAP:** Which stands for Online Analytical Processing used by Data Warehouse.

**Data Warehouse:** It a like a database where data comes from multiple sources to analyse and generate reports.

The data store contains two main types of data.

**-Business Data**-which comes from operating systems and different business domains.

-**Business Data Model-**It’s a visual representation that defines how the is organised.

**DSS data vs Operational data:**

This can be differ in three terms:

* Time span
* Granularity
* Dimensionality

An ETL based Data Warehouse uses as its key functions.

1. **Staging**- data comes from multiple places and used to analyse.
2. **integration** -combining the data from different sources.
3. **access layers (Data Marts)-**These are like interfaces where the user can interact and access data from data marts.