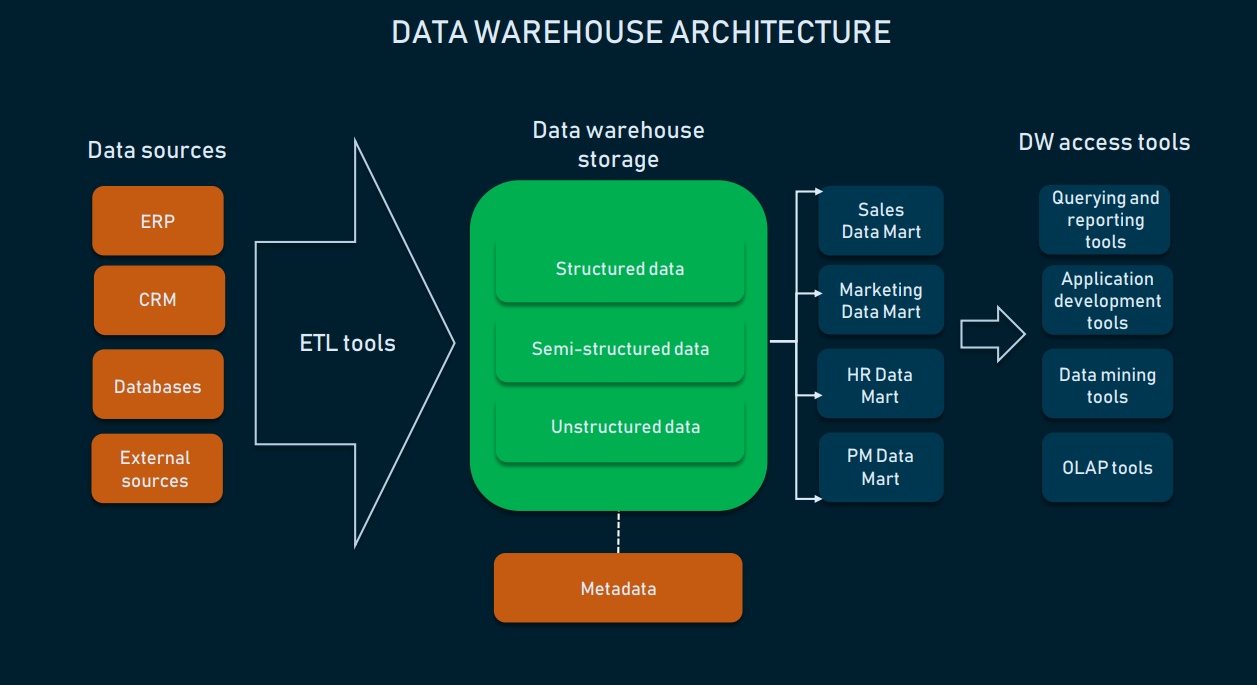
Week 1 learnings

Q - What is Big Data?

* In simple words, huge amount of data which is generated in very less time is considered as Big Data.
  + For e.g.: Petabytes of Videos getting uploaded on YouTube, Getting millions of tweets on Twitter in a minute, Millions of email sent in a minute etc.
* How do we classify any data as big data?
  + If the data has below characteristics, then we can consider it as big data.
    - **Volume**: This simply means, large amount of data. With big data, we’ll have to process high volumes of unstructured data. Initially the value of data is unknown, such as twitter data feeds, Clickstream on a webpage or a mobile app or senor enabled equipment.
    - **Velocity**: The data which is been received in high speed and processed as close to real-time as possible. This determines the potential of data that how fast the data is generated and processed to meet the demands.
      * For e.g.: There are more than 3.5 billion searches per day are made on Google. Also, Facebook users are increasing by 22%(Approx.) Year by year.
    - **Variety**: It refers to nature of data that is structured, semi-structured and unstructured data.
      * Structured data: This data is basically an organized data. This type of data has defined length and format. For eg : xlsx, csv file or data in relational database.
      * Semi- Structured data: This data is basically a semi-organised data. This type of data do not have a concrete structure. Log and json files are the examples of this type of data.
      * Unstructured data: This data basically refers to unorganized data. It generally refers to data that doesn’t fit neatly into the traditional row and column structure of the relational database. Texts, pictures, videos etc. are the examples of unstructured data which can’t be stored in the form of rows and columns.
    - **Veracity**: It refers to inconsistency in data, like biases, noise and abnormality. In general it checks how much accurate and truthful of data set is and how much we can rely on it.
    - **Value**: It refers, how much knowledge we can discover from the data in hand, to achieve/solve the business problem or maybe to make a good profit.

Q – What is the role of Data Engineer?

1. Majorly they **Extract, Transform and Load** the data.
   1. **Extracting data**: The information is located somewhere, so first we have to extract it. In terms of corporate data, the source can be some database, a website’s user interactions, an internal ERP/CRM system, etc. Or the source can be a sensors, for e.g. IoT devices. Or the data may come from public sources available online.
   2. **Transformation**: Raw data may not make much sense to the end users, because it’s hard to analyze in such form. Transformations aim at cleaning, structuring, and formatting the data sets to make data consumable for processing or analysis. In this form, it can finally be taken for further processing or queried from the [reporting layer](https://www.altexsoft.com/blog/data-analytics-tools/).
   3. **Data storing/transition**: The main architectural point in any data pipeline is storages. We need to store extracted data somewhere. In data engineering, the concept of a [data warehouse](https://www.altexsoft.com/blog/enterprise-data-warehouse-concepts/) embodies an ultimate storage for all data gathered for analytical purposes.

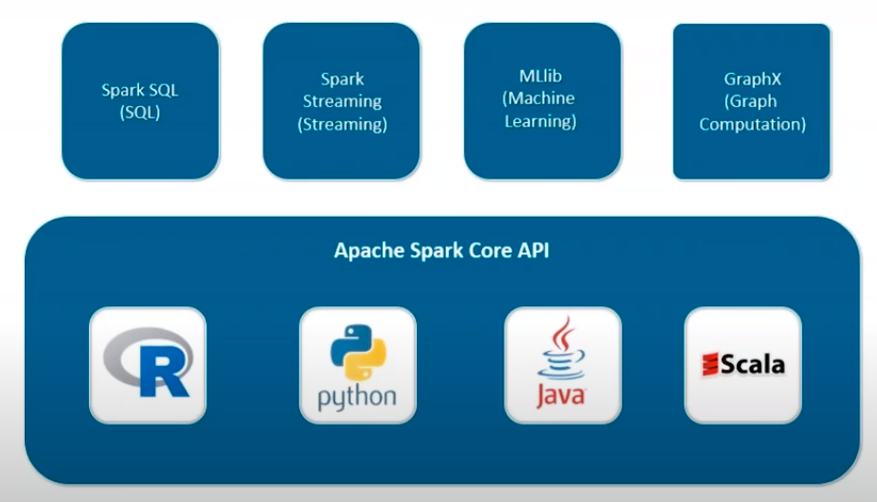


**ETL in Data warehousing**

1. DE are also responsible for **designing the architecture of data platform**.
2. They automate tasks by **creating Data Pipeline**.
3. They also maintain the **Data Pipeline** by testing the reliability and performance at each steps.
4. They also **monitored the automated jobs** by checking the performance and stability of the data pipeline.

Q - What is Spark and Why Spark is used?

1. Apache Spark is an Open source framework used for processing data quickly. Data processing can be done in 2 ways.
   1. Real-time data processing: This is achieved via in memory computation.
   2. Batch data processing: This is achieved via scheduling a program.
2. **Spark Eco System**



* 1. **Spark SQL** – It is used to leverage the power of SQL queries on the spark data which can be accessed in the form of RDDs and other external sources.
  2. **Spark Streaming** – This component provides real time processing of data and batch processing.
  3. **Mlib(Machine Learning)** – This component is used to create and deploy scalable machine learning pipelines.
  4. **GraphX(Graph Computation)** – Spark also comes with a library to manipulate graph databases and perform computations called GraphX.

1. **Features of Spark**
   1. Fast Processing – This is achieved via RDDs
   2. Flexibility – As it supports multiple languages like Java, Scala, Python and R.
   3. In-memory Computation - Spark stores the data in the RAM of servers which allows quick access and in turn accelerates the speed of analytics
   4. Real-Time processing - Spark is able to process real-time streaming data. Unlike MapReduce which processes only stored data, Spark is able to process real-time data and is, therefore, able to produce instant outcomes.
   5. Better analystics – This is achieved using the components like like Spark SQL, Machine Learning libraries, GraphX analystics etc.