**Apache Spark**

* Apache Spark is introduced to work with large amounts of data without interruption and for fast execution we use apache spark.
* Apache Spark first use for GFS(google file system) to store the data on google and

Retrieve the information back for web pages etc. for this kind of problems or solve these problems we use apache spark.

* We also use HDFS(Hadoop file System) to store the data. It allows us to use multiple systems for storing the data and retrive data on requirements.
* After HDFS work with Data flow a new term is introduced that is **Data lake.**
* **Data lake follow 4 steps :**

1. **Ingest** (Data Collection and integration)

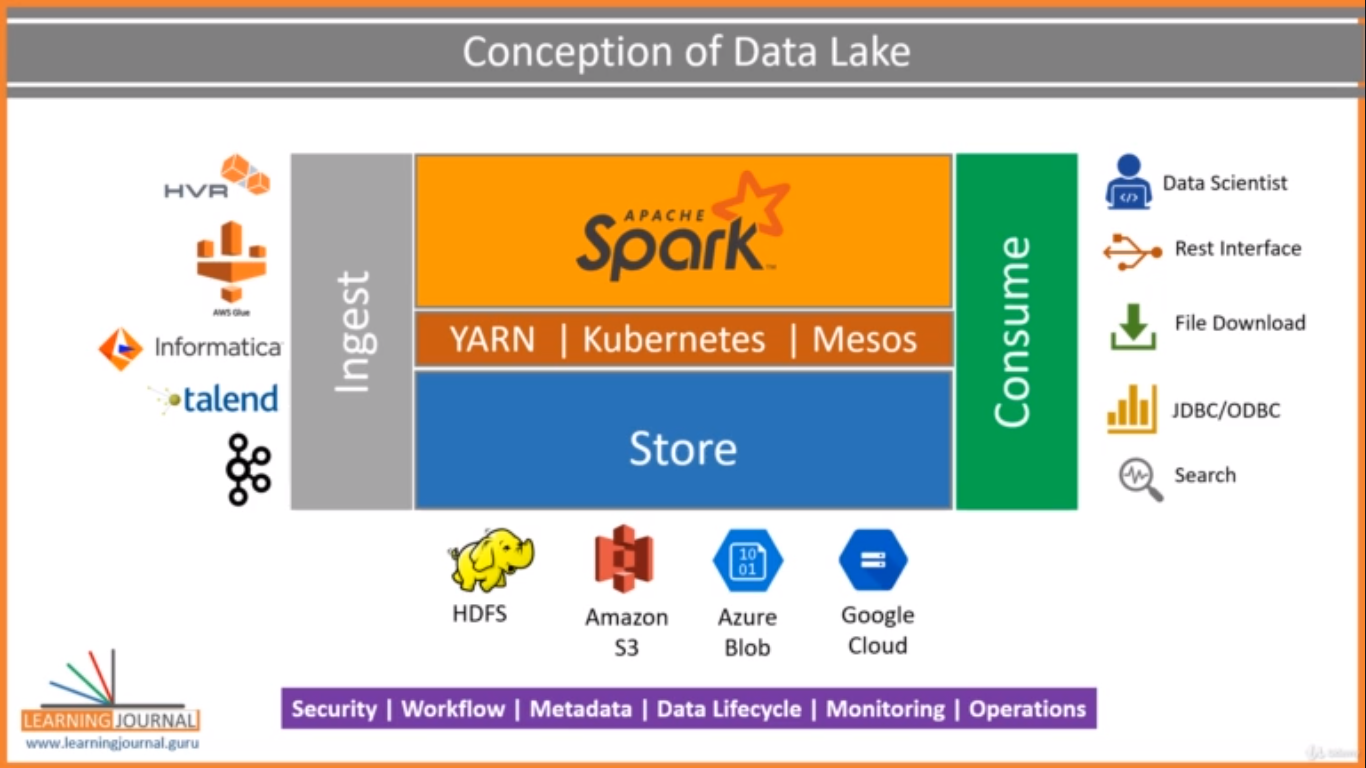
Ingest means data comes in raw format and various tools are used to store the data from various sources to the data ingest layer.

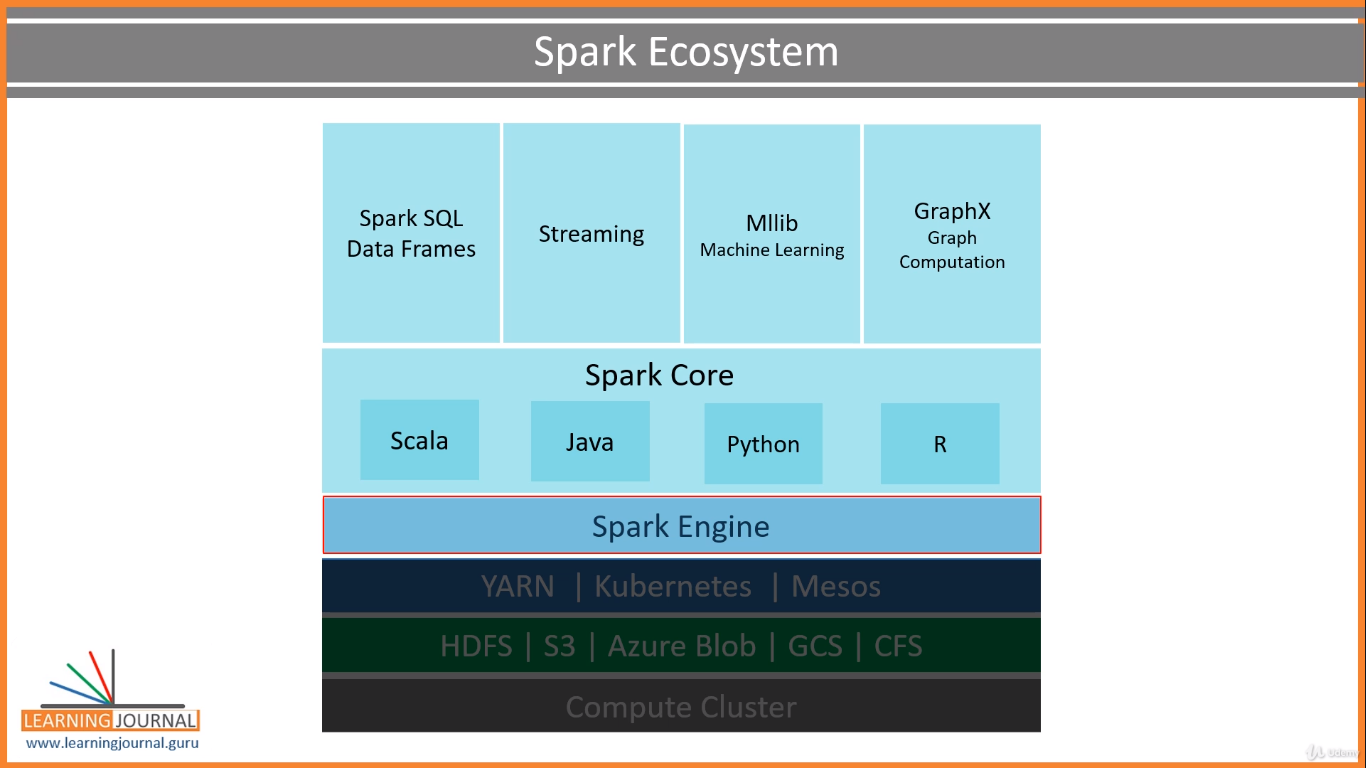
1. Store (Data Storage and management)
2. Process (Data Processing and transformation)

Process means - Initial data Quality Check transforming and preparing data and

Applying ML Model

1. Consume(Data Access and retrieval)



* These are the 4 problems that google faceing with data and now solve this problem with the help of Data lake and for storage the data currently user using various kinds of services like HDFS on Premises, Amazon S3, GC etc. and among all of these cloud is using too much because cloud is extremely available on low cost and high scalability.
* **Working with Apache Spark first we understand what is spark ?**
* Apache Spark is used for running the data process work load but for storage the data HDFC, AWS S3, GCP etc services is used for storing the data and spark engine is responsible for the data processing by dividing the work load into sub parts and managing and monitoring this task.
* **Spark EcoSystem-**
* ****
* **Spark SQL DataFrame**
* Spark SQL is use to run sql query in spark for processing our data.
* **Spark Streaming**
* Spark streaming is use to stream the data without interruption.
* **Mllib**
* Here Mllib is use for ML and AI LIb.
* **GraphX**

This lib is use for graph algorithm

We use spark because it is use dataframe or database like sql so for that it is easy to work like sql so spark is so popular and very useful and we can use it with our fav programing language along with sql queries.

**So Finally we are going to work with Apache Spark System**

1. **Apache Installation**