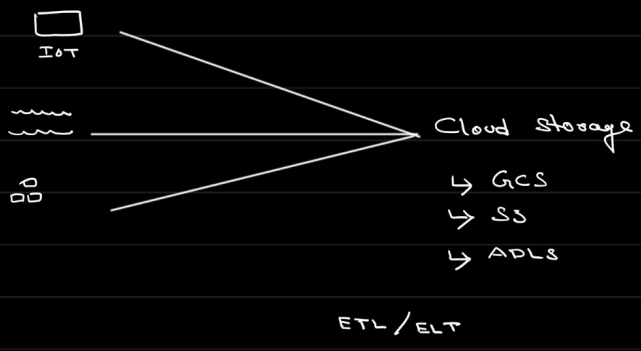




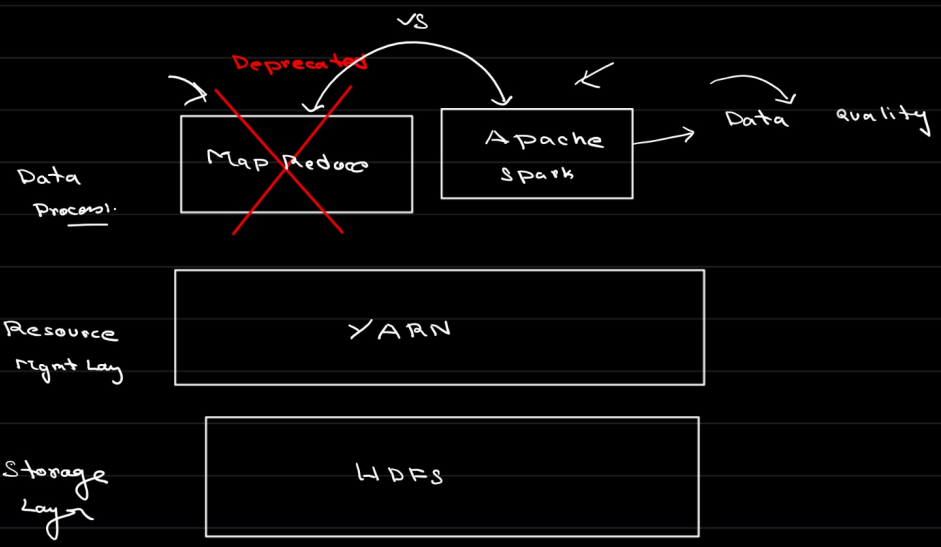



## Challenges of Big Data

1. Storage : Distribute storage system
2. Processing : MPP (Massive Parallel processing)

## Hadoop 2.x

1. Storage Layer : HDFS (Hadoop Distributed FS)
2. Resource Mgmt Layer : YARN
3. Processing Layer : Map Reduce

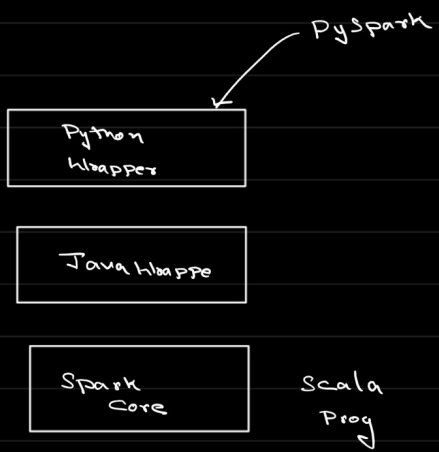



Big Data Workloads

- Data Integration ETL
- Batch Computation
- ML
- Real-time streaming

- Apache Spark is natively written using Scala

- NumPy
- SciPy
- Pandas
- Seaborn
-



# Spark Ecosystem