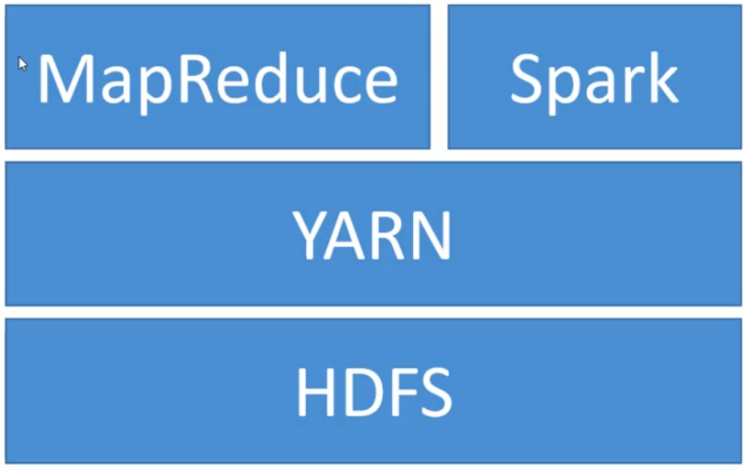
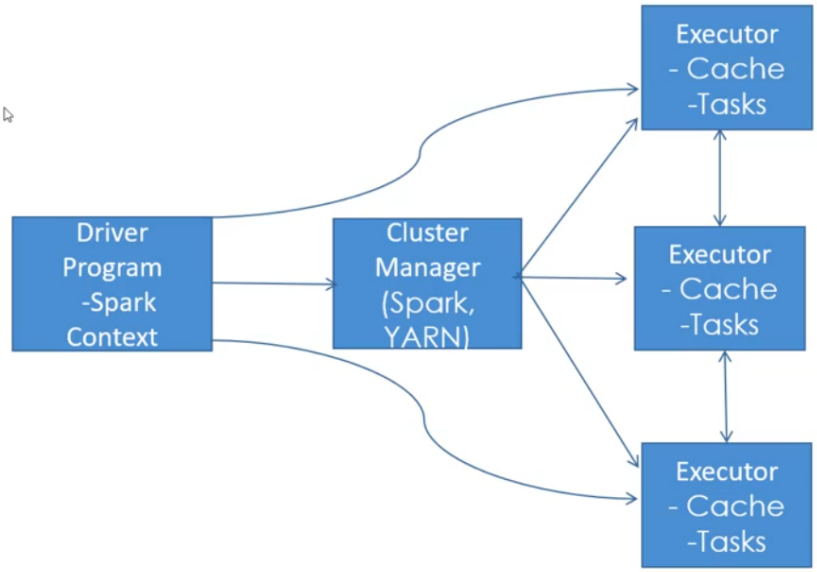
spark

Apache Spark



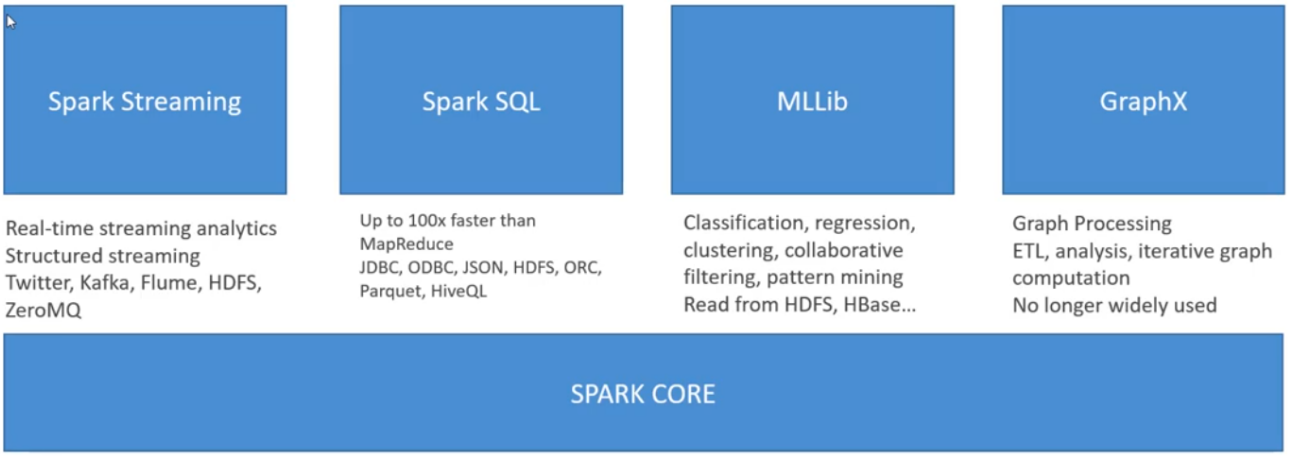
- Distributed processing framework for big data
- In-memory caching, optimized query execution
- Supports Java, Scala, Python, and R
- Supports code reuse across
  - Batch processing
  - Interactive Queries
    - Spark SQL
  - Real-time Analytics
  - Machine Learning
    - MLLib
  - Graph Processing
- Spark Streaming
  - Integrated with Kinesis, Kafka, on EMR
- Spark is NOT meant for OLTP

How Spark Works



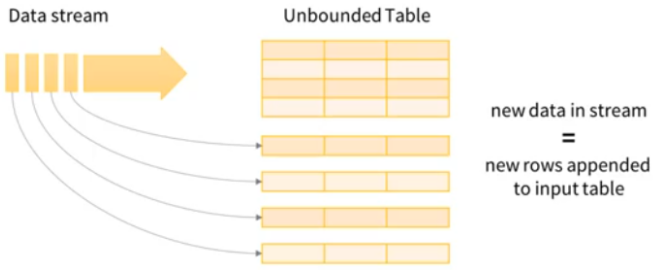
- Spark apps are run as independent processes on a cluster
- The SparkContext (driver program) coordinates them
- SparkContext works through a Cluster Manager
- Executors run computations and store data
- SparkContext sends application code and tasks to executors

Spark Components



Memory management, fault recovery, scheduling, distribute & monitor jobs, interact with storage
Scala, Python, Java, R

Spark Structured Streaming
A constantly growing DataSet



Data stream as an unbounded Input Table

```
val inputDF = spark.readStream.json("s3://logs")
inputDF.groupBy($"action", window($"time", "1 hour")).count()
       .writeStream.format("jdbc").start("jdbc:mysql//...")
```

Spark + Redshift

- spark-redshift package allows Spark datasets from Redshift
  - It's a Spark SQL data source
- Useful for ETL using Spark