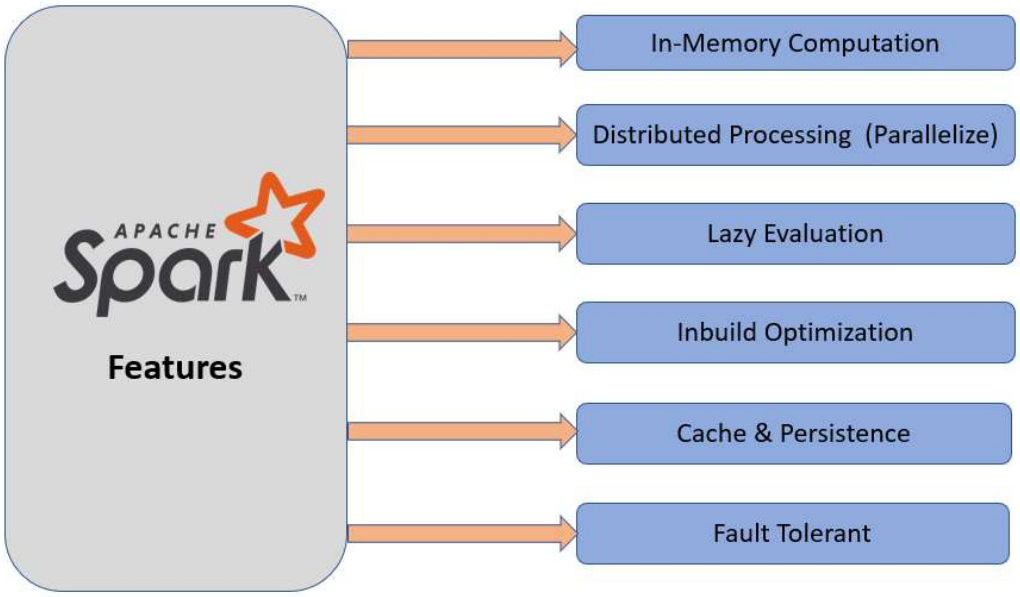
Apache Spark (Pyspark) Introduction

- ✓ Apache Spark is bigdata analytical processing engine for large scale distributed data processing applications.
- ✓ Spark is easier & faster than Hadoop Map-Reduce
- ✓ Spark is Open Source Software developed by UC Berkeley
- ✓ Pyspark is python API for Apache Spark.
- ✓ Java, Python, Apache-spark are required for working with Pyspark.

Features:



Languages Supported:

Pyspark Modules & Packages:

- PySpark RDD (pyspark.RDD)
- PySpark DataFrame and SQL (pyspark.sql)
- PySpark Streaming (pyspark.streaming)
- PySpark MLib (pyspark.ml, pyspark.mllib)
- PySpark GraphFrames (GraphFrames)
- PySpark Resource (pyspark.resource) It's new in PySpark 3.0

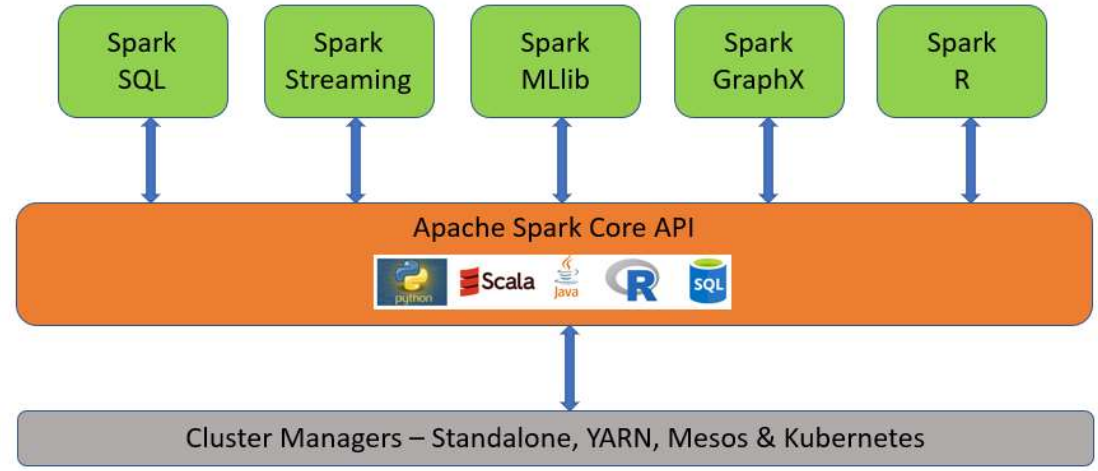
Data Sources:

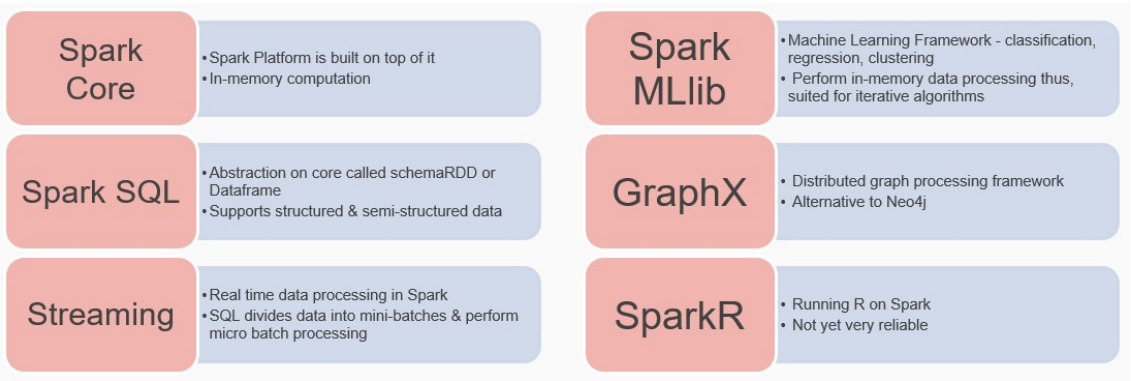
Spark can process data from HDFS, AWS-S3 & other files systems

Data Formats:

Can Read & Write in variety of data formats like Hive tables, JSON, CSV, Parquet, Avro, ORC

Apache Spark Ecosystem:





Standalone – A simple cluster manager included with Spark that makes it easy to set up a cluster.

Apache Mesos – Mesons is a Cluster manager that can also run Hadoop MapReduce and Pyspark applications.

Hadoop YARN – The resource manager in Hadoop 2. This is mostly used, cluster manager.

Kubernetes – An open-source system for automating deployment, scaling, and management of container apps

Advantages:

1. Pyspark is an in-memory computation, distributed processing engine.
2. Applications running on Pyspark are 100x faster than traditional systems.
3. Pyspark also is used to process real-time data using Streaming and Kafka.
4. Pyspark natively has machine learning and graph libraries.