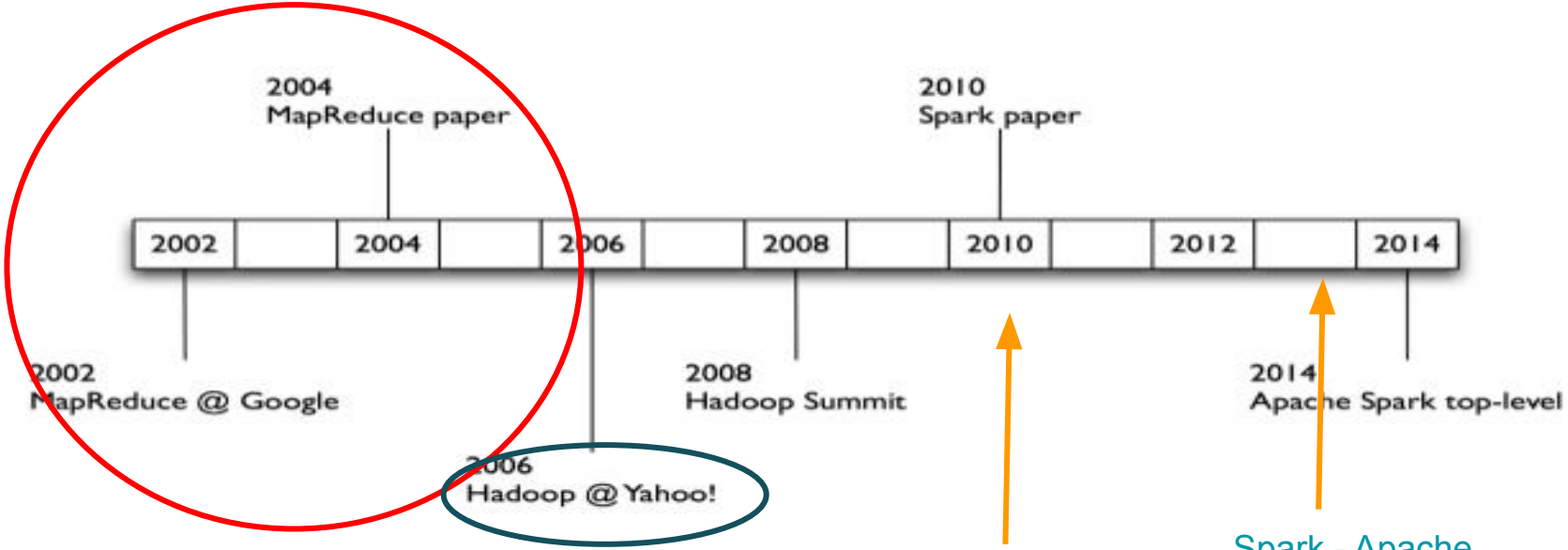
Apache Spark

in-memory, parallel distributed processing



MapReduce - GFS

Doug Cutting
&
Mike Cafarella
MapReduce - HDFS

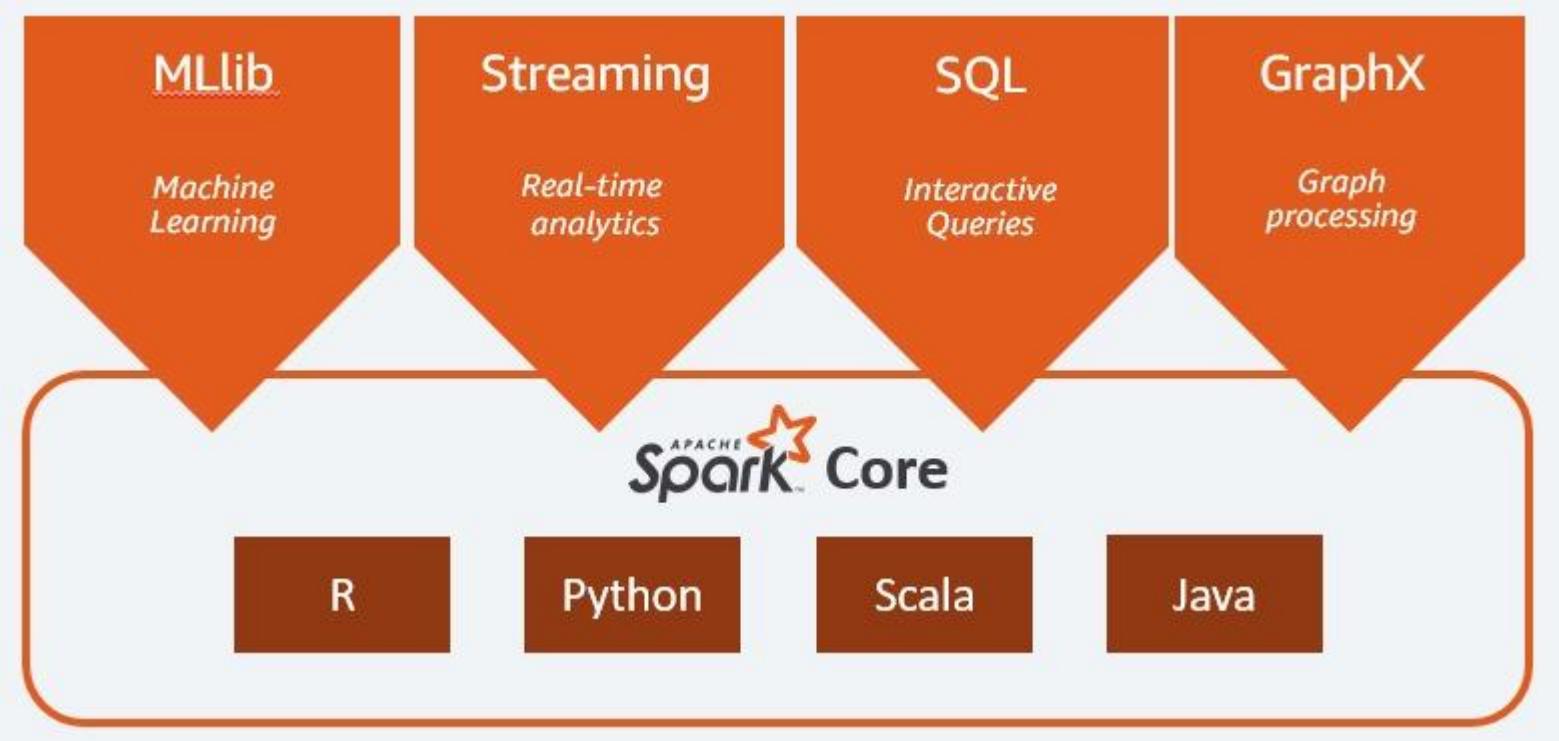
2009 Research project at UC Berkeley AMPLab

Spark - Apache Software Foundation

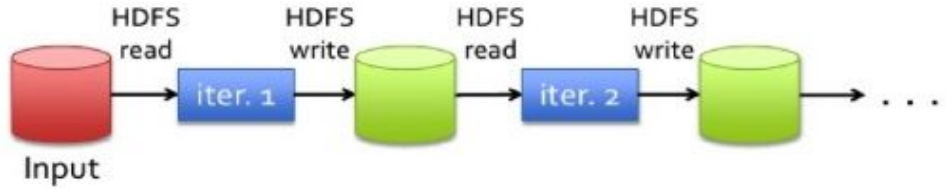
Lucene
Apache-Hadoop

GFS - Google File System
HDFS - Hadoop Distributed File System

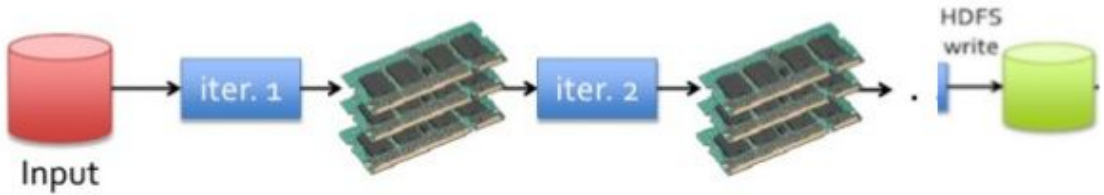

APACHE
Spark
Stream Processing
& Batch Processing
Speed
Lazy Evaluation
Concise
In-memory Processing


MLlib
Machine Learning
Streaming
Real-time analytics
SQL
Interactive Queries
GraphX
Graph processing
APACHE Spark Core
R
Python
Scala
Java

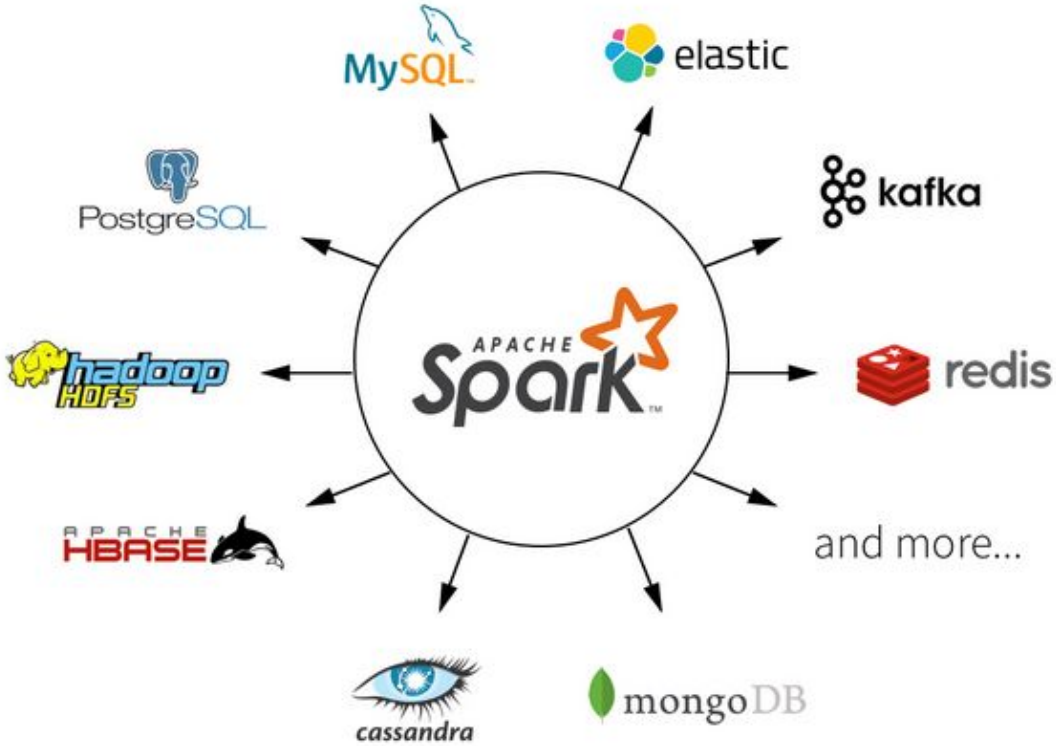
Hadoop



Spark



Fuente


MySQL
elastic
PostgreSQL
kafka
hadoop
HDFS
APACHE
Spark
redis
APACHE
HBASE
and more...
cassandra
mongoDB

Resilient Distributed Database

- Colección de elementos que pueden ser divididos en múltiples nodos para ser procesados en paralelo
- Un RDD es inmutable, pero se pueden ejecutar cualquier operación y luego crear otro RDD.
  - MAP - División
  - REDUCE - Se ejecuta una transfromación y el resultado se regresa a driver

Referencias

¿por qué usar PySpark para grandes conjuntos de datos que exceden la memoria de la máquina de un solo nodo?

Comprehensive Introduction to Apache Spark, RDDs & Dataframes (using PySpark)

How to use a Machine Learning Model to Make Predictions on Streaming Data using PySpark