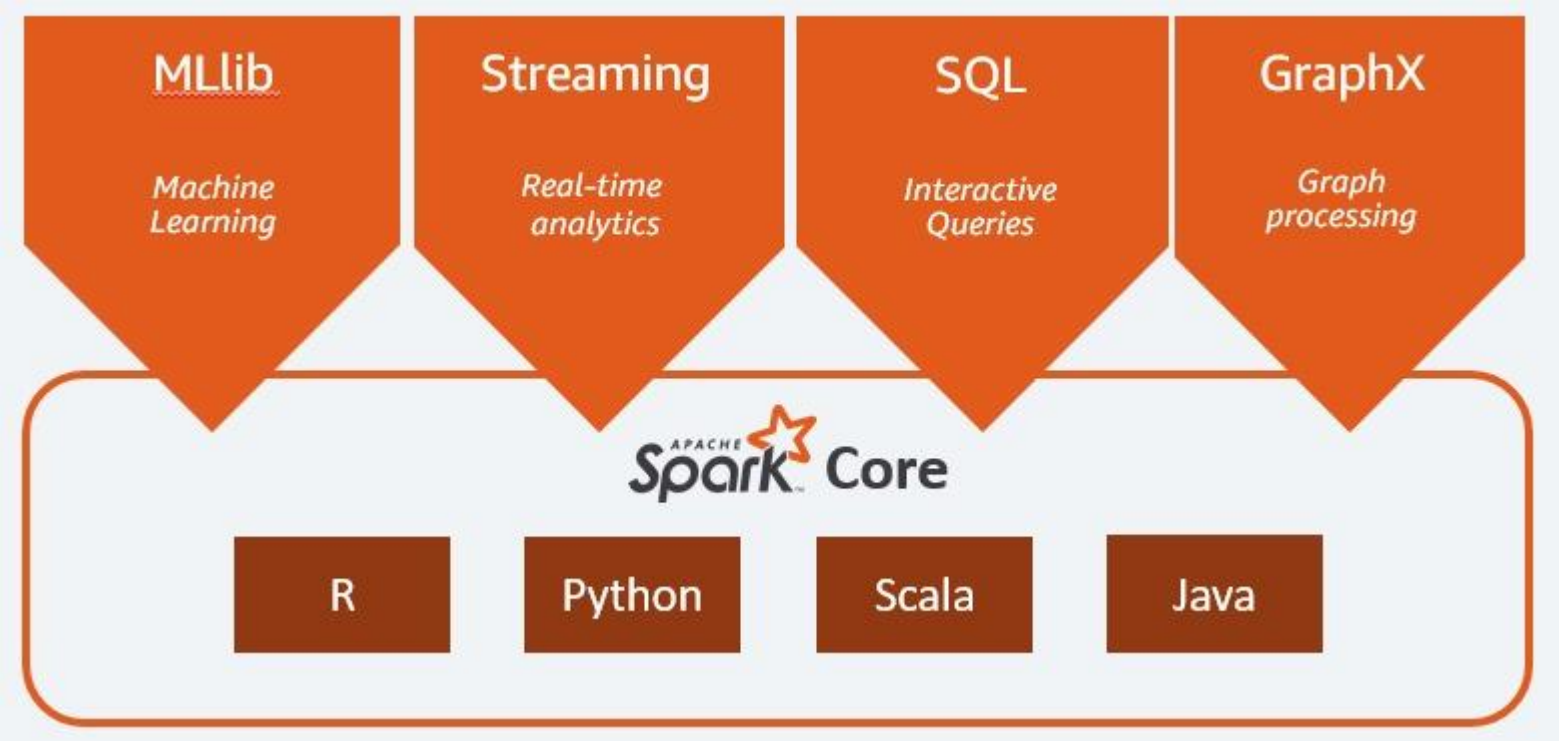
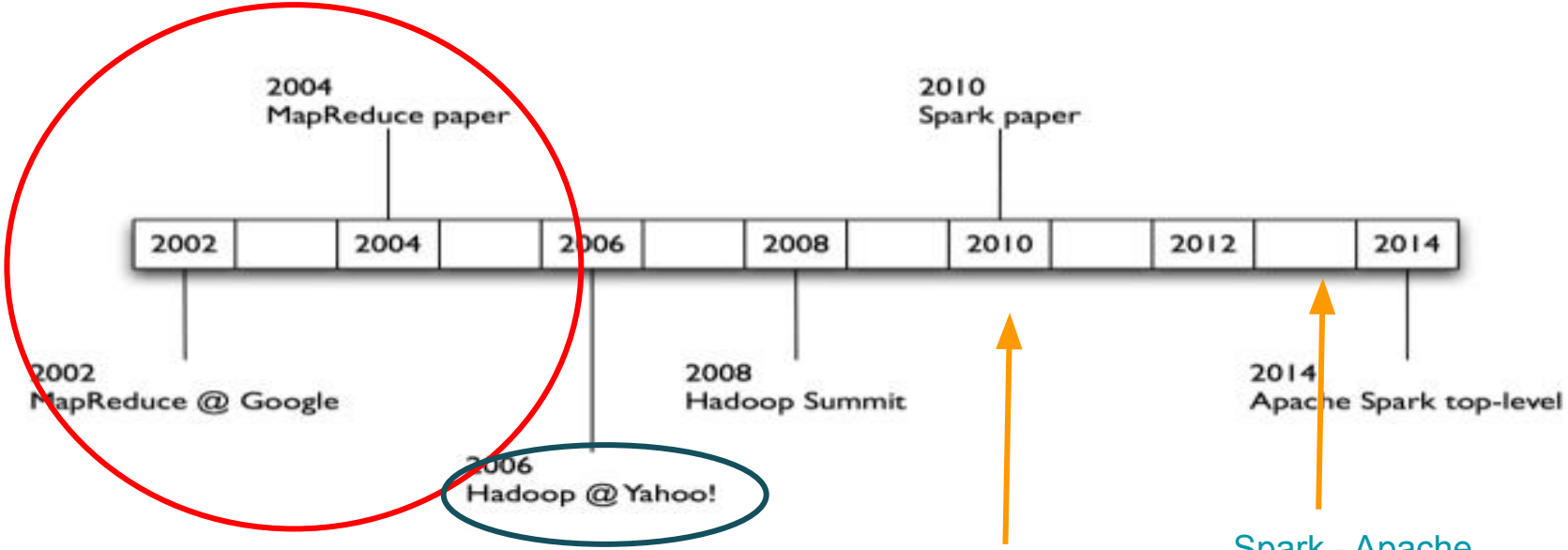
Apache Spark

Framework de computo distribuido


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



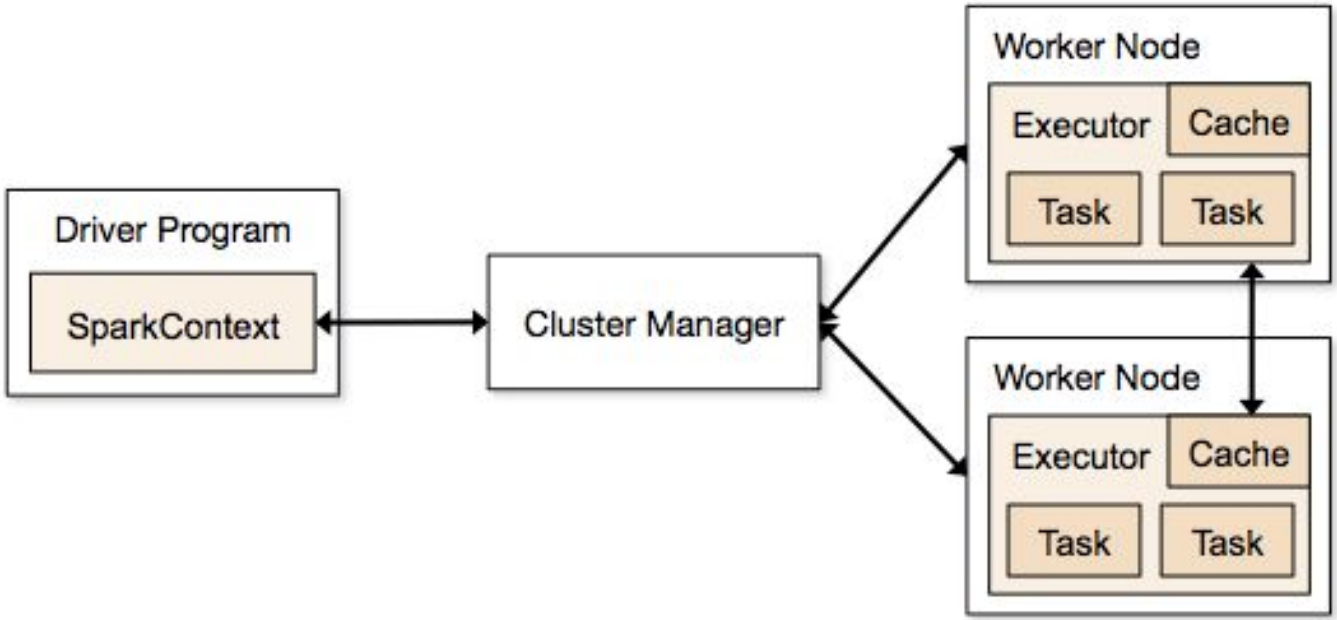
MapReduce - GFS

Doug Cutting
&
Mike Cafarella
MapReduce - HDFS

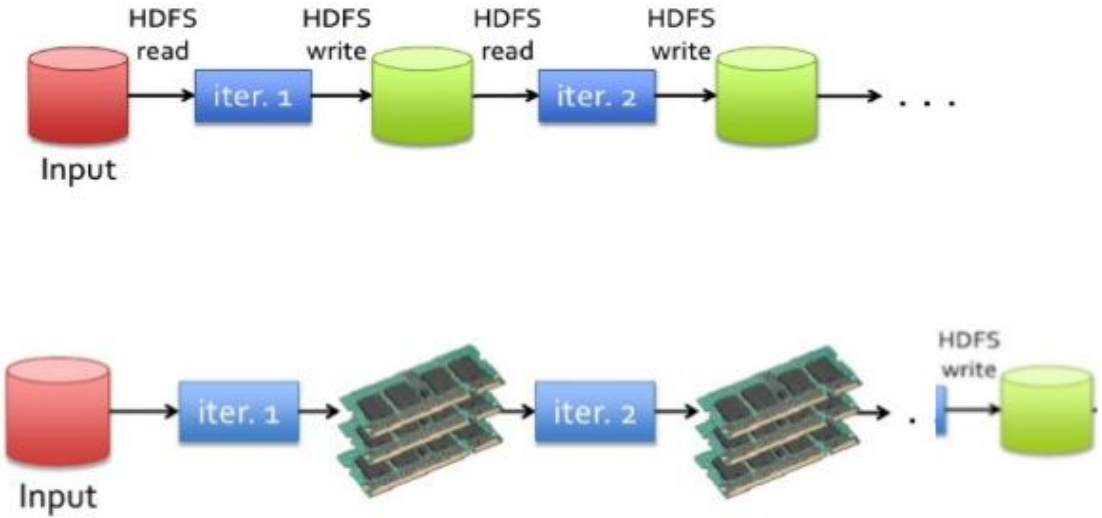
Lucene

2009 Research proyect al UC Berkley AMPLab

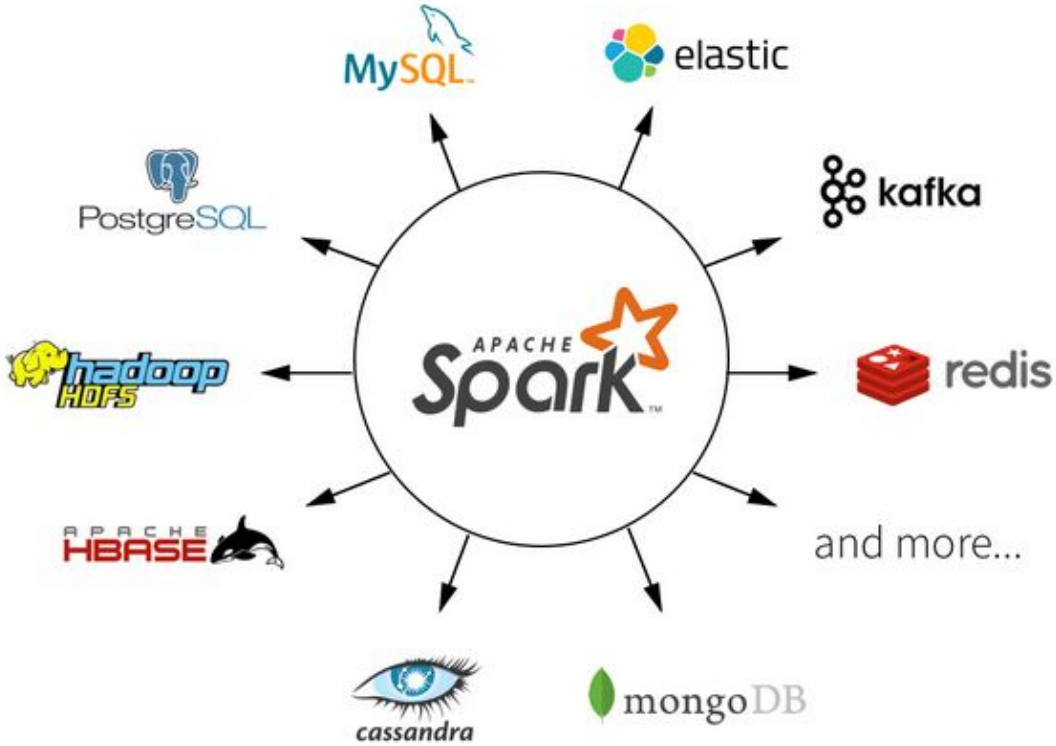
Spark - Apache Software Foundation



- Master
- RDD
- Operaciones Lazy



https://www.analyticsvidhya.com/wp-content/uploads/2016/09/big1.png


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


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

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