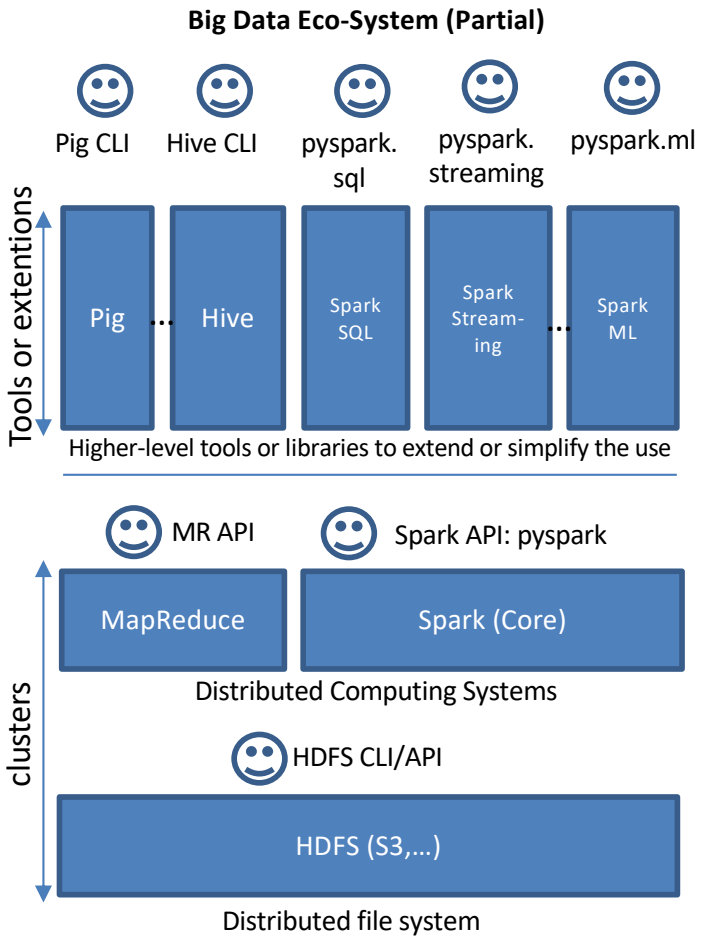
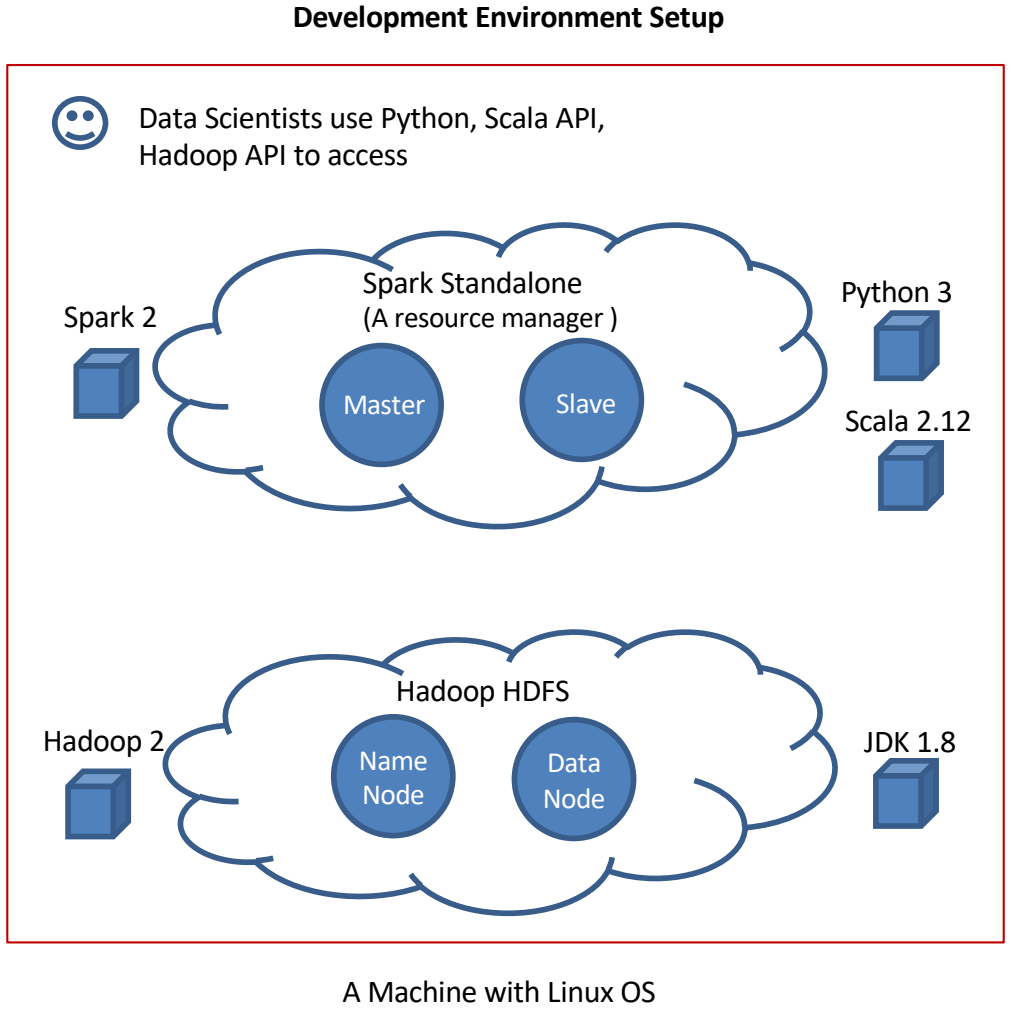

Big Data Eco-System (Partial)
Pig CLI
Hive CLI
pyspark. sql
pyspark. streaming
pyspark.ml
Tools or extentions
Pig
Hive
Spark SQL
Spark Stream-ing
Spark ML
Higher-level tools or libraries to extend or simplify the use
MR API
Spark API: pyspark
MapReduce
Spark (Core)
Distributed Computing Systems
clusters
HDFS CLI/API
HDFS (S3,...)
Distributed file system


Development Environment Setup
Data Scientists use Python, Scala API, Hadoop API to access
Spark 2
Spark Standalone (A resource manager)
Master
Slave
Python 3
Scala 2.12
Hadoop HDFS
Hadoop 2
Name Node
Data Node
JDK 1.8
A Machine with Linux OS