零基础学习Spark 1.x应用开发系列课程

Spark 1.x介绍

讲师-梦琪

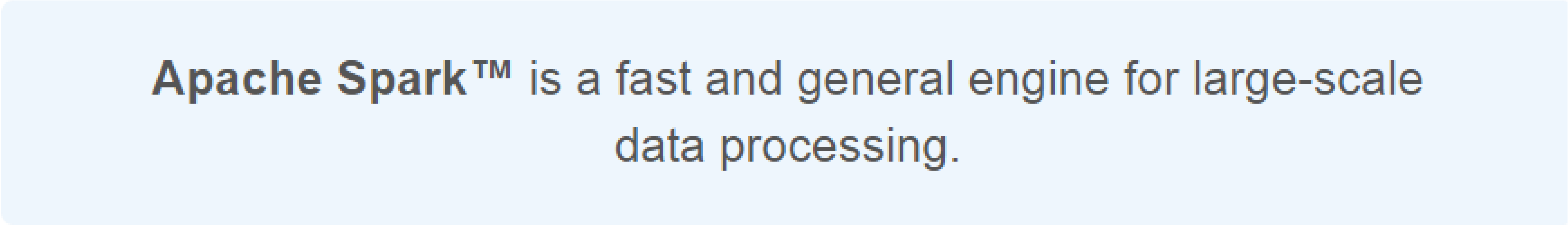
版权声明

【声明】本视频和讲义等均为云帆大数据网络课程的教学资料，所有资料只能在课程内使用，不允许在课程以外范围散播，违者将可能被追究法律和经济责任。

课程详情访问云帆大数据官方培训网站

http://www.cloudyhadoop.com

What is Spark?



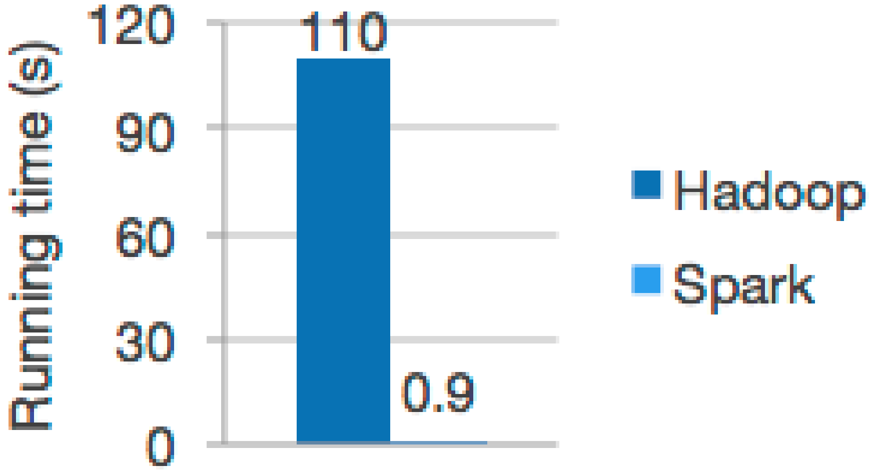
Apache Spark™ is a fast and general engine for large-scale data processing.

What is Spark?

Speed

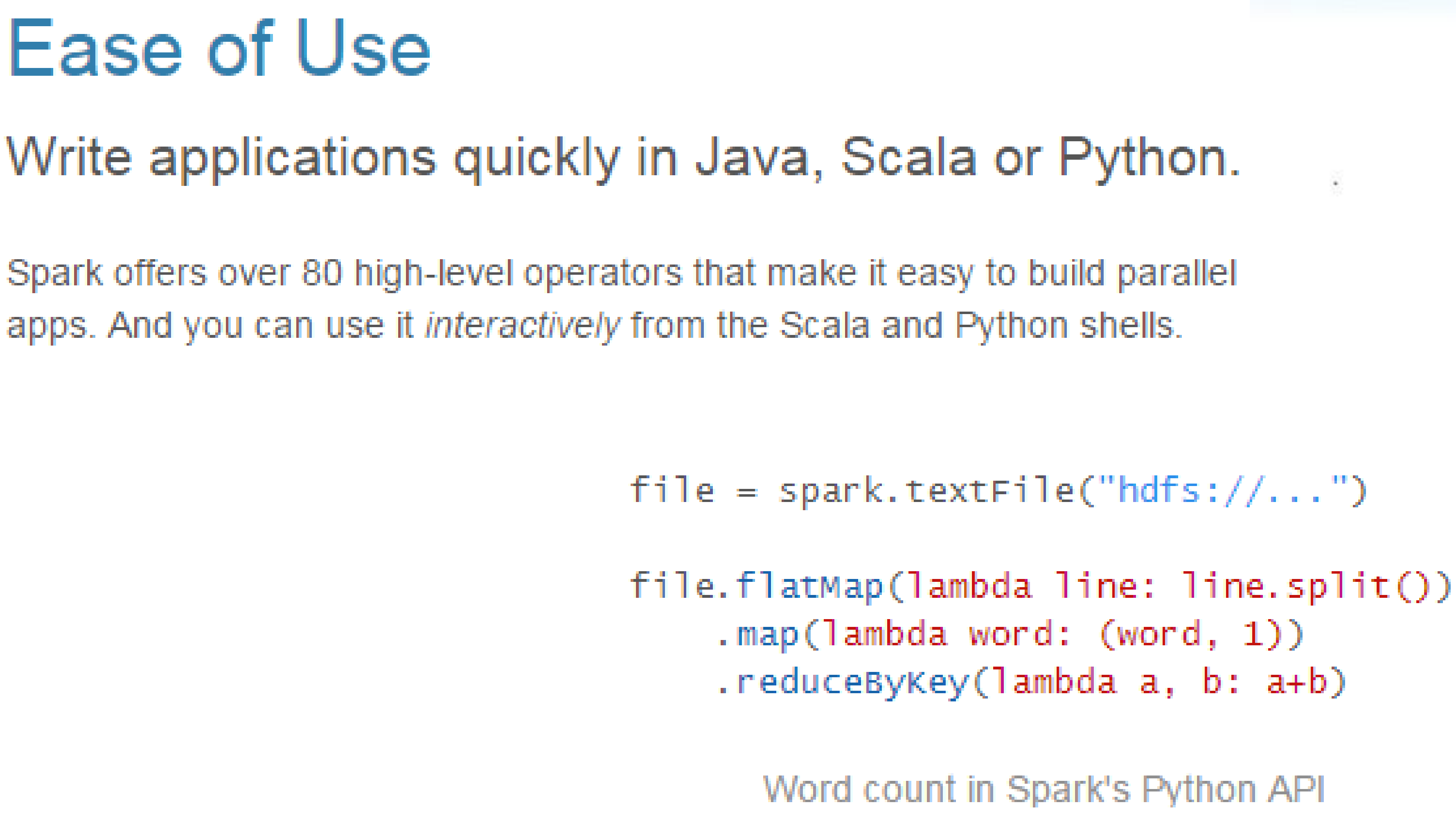
Run programs up to 100x faster than Hadoop MapReduce in memory, or 10x faster on disk.

Spark has an advanced DAG execution engine that supports cyclic data flow and in-memory computing.



Logistic regression in Hadoop and Spark

What is Spark?



Ease of Use

Write applications quickly in Java, Scala or Python.

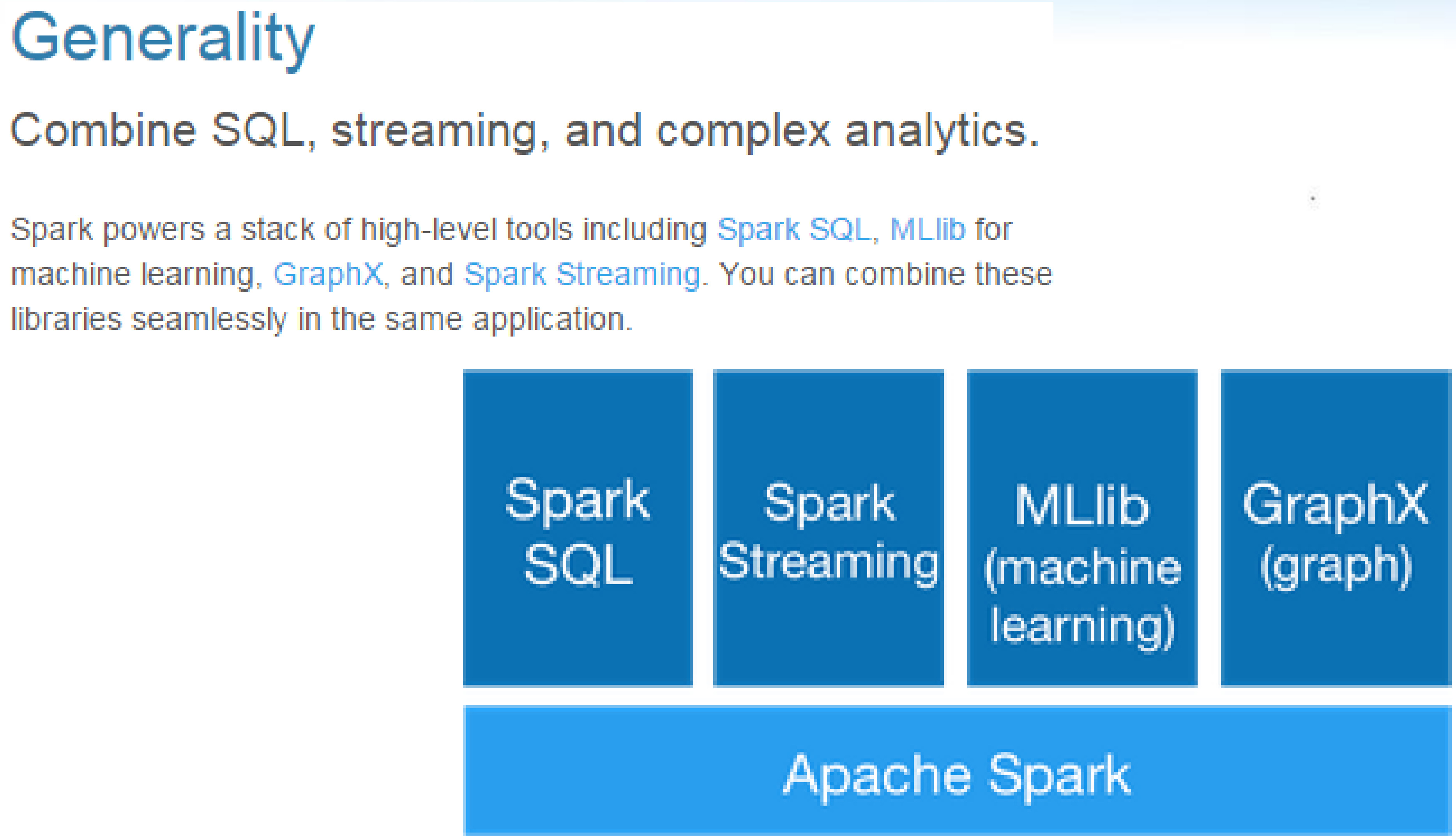
Spark offers over 80 high-level operators that make it easy to build parallel apps. And you can use it *interactively* from the Scala and Python shells.

```
file = spark.textFile("hdfs://...")

file.flatMap(lambda line: line.split())
    .map(lambda word: (word, 1))
    .reduceByKey(lambda a, b: a+b)
```

Word count in Spark's Python API

What is Spark?



Generality

Combine SQL, streaming, and complex analytics.

Spark powers a stack of high-level tools including Spark SQL, MLlib for machine learning, GraphX, and Spark Streaming. You can combine these libraries seamlessly in the same application.

Spark SQL

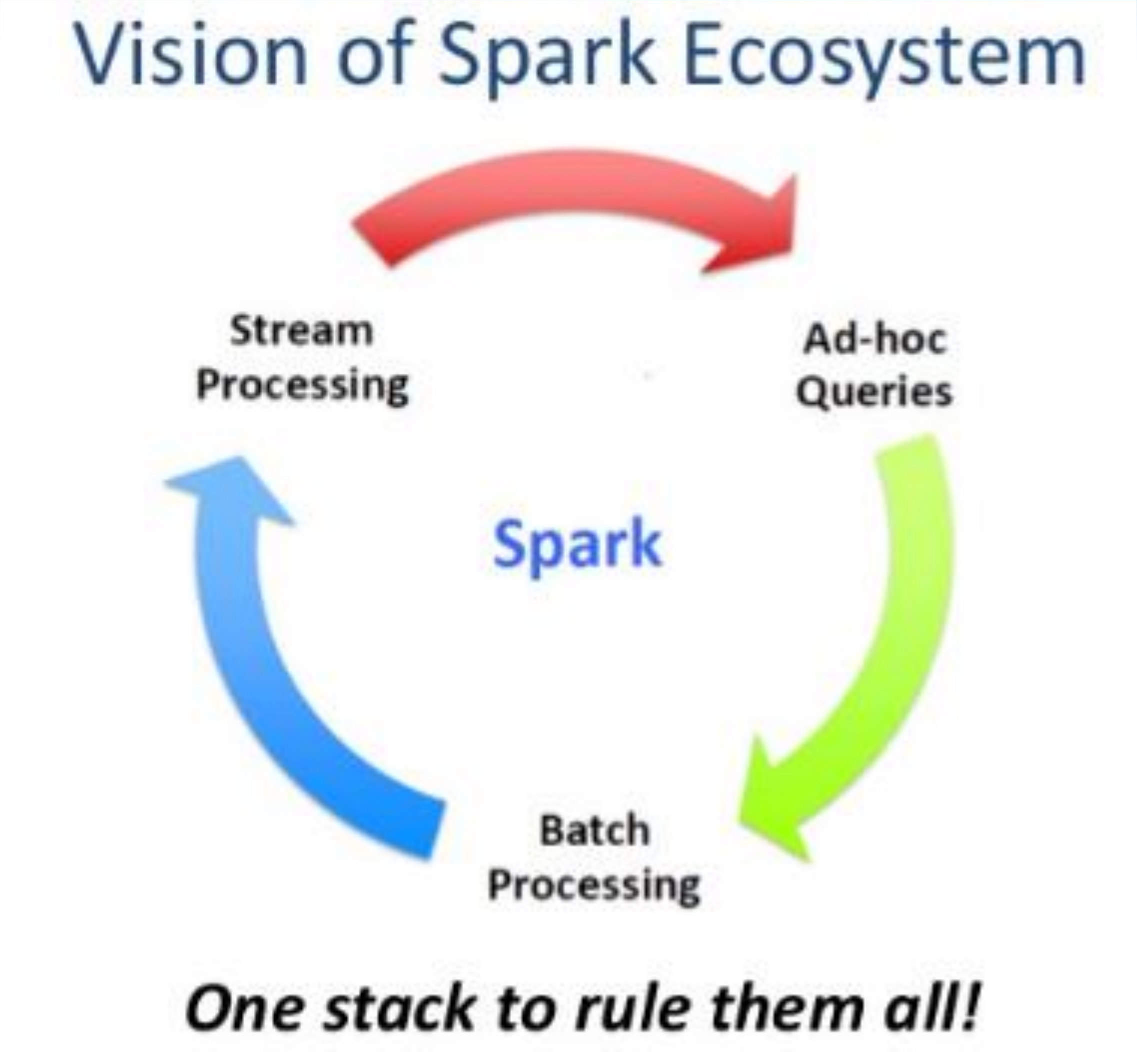
Spark Streaming

MLlib (machine learning)

GraphX (graph)

Apache Spark

One Stack to Rule them All



Vision of Spark Ecosystem

Stream Processing

Ad-hoc Queries

Spark

Batch Processing

One stack to rule them all!

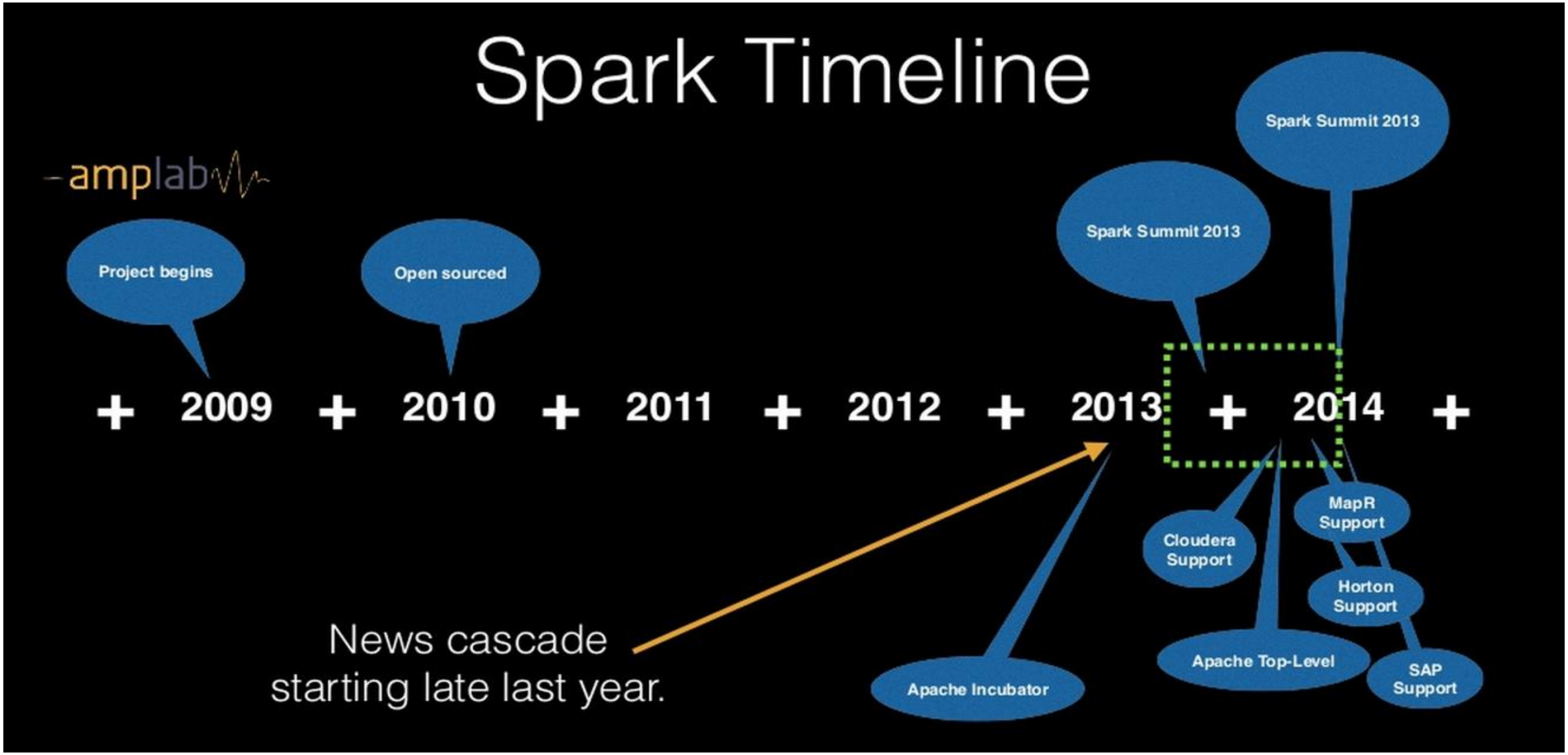
What is Spark?

Runs Everywhere

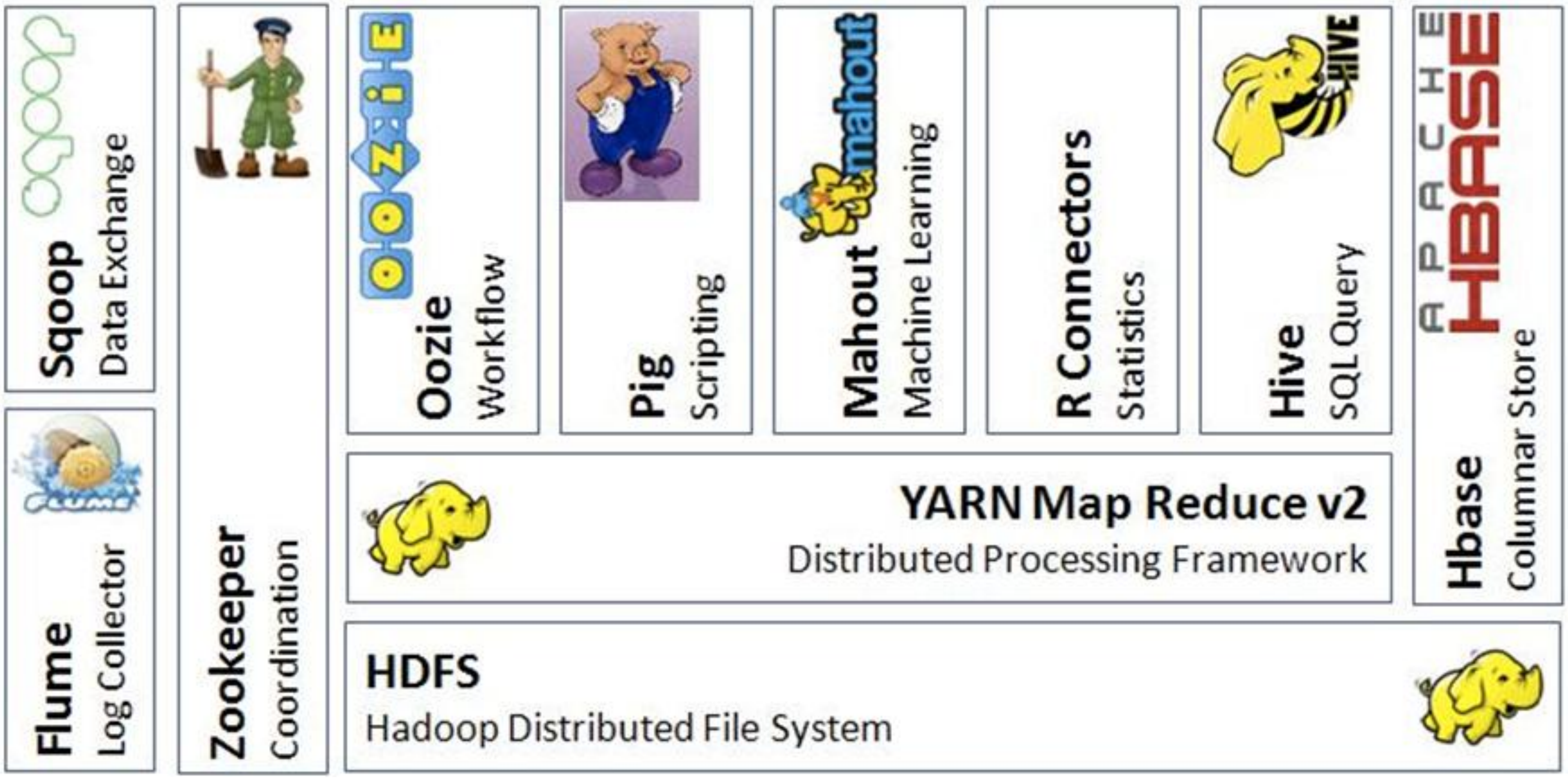
Spark runs on Hadoop, Mesos, standalone, or in the cloud. It can access diverse data sources including HDFS, Cassandra, HBase, S3.

You can run Spark readily using its standalone cluster mode, on EC2, or run it on Hadoop YARN or Apache Mesos. It can read from HDFS, HBase, Cassandra, and any Hadoop data source.

Spark Timeline



Hadoop Ecosystem



Sqoop
Data Exchange

Flume
Log Collector

Zookeeper
Coordination

Oozie
Workflow

Pig
Scripting

Mahout
Machine Learning

R Connectors
Statistics

Hive
SQL Query

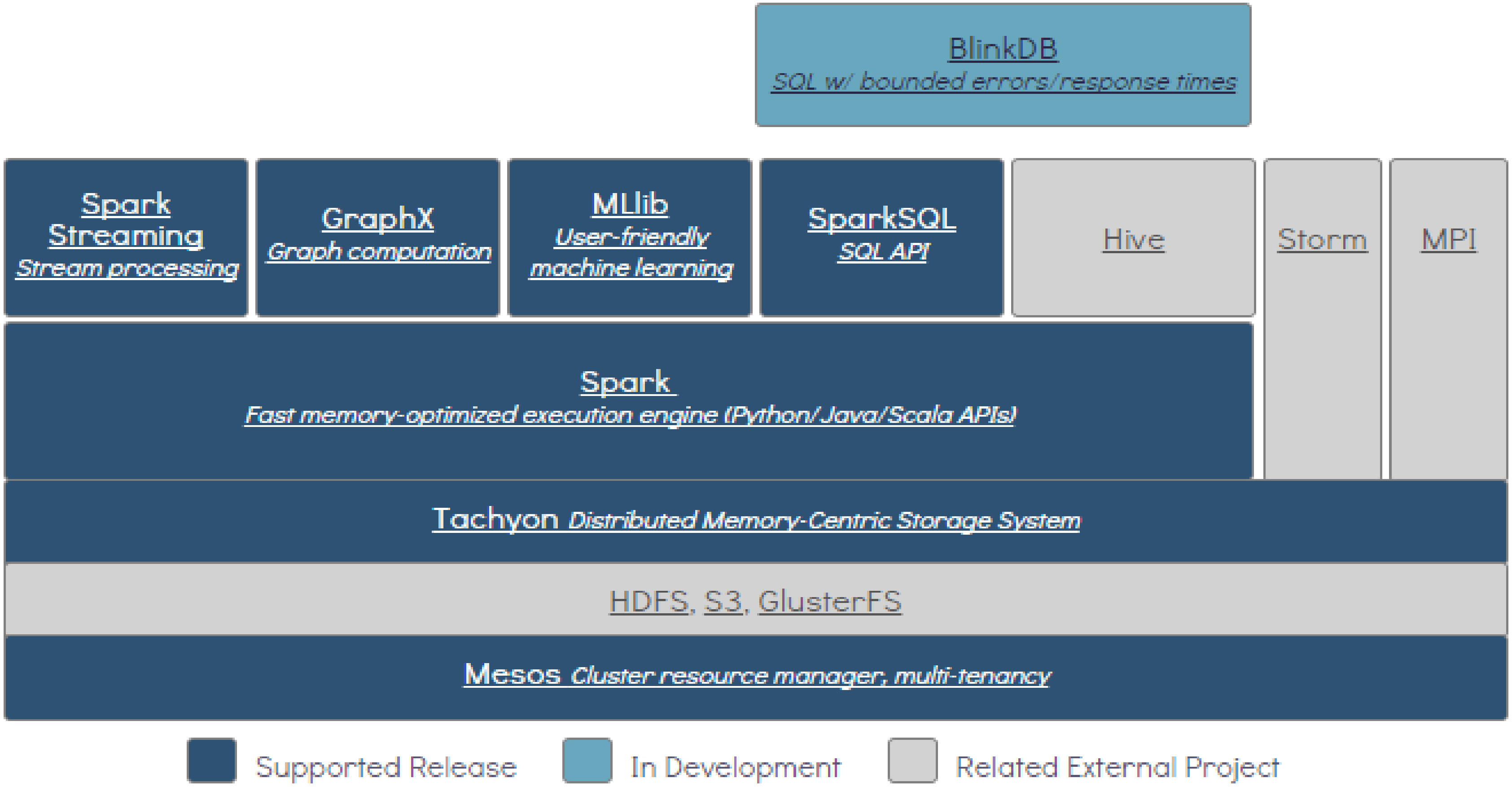
Hbase
Columnar Store

YARN Map Reduce v2
Distributed Processing Framework

HDFS
Hadoop Distributed File System

BDAS

BDAS: Berkeley Data Analytics Stack



BlinkDB
SQL w/ bounded errors/response times

Spark Streaming
Stream processing

GraphX
Graph computation

MLlib
User-friendly machine learning

SparkSQL
SQL API

Hive

Storm

MPI

Spark
Fast memory-optimized execution engine (Python/Java/Scala APIs)

Tachyon Distributed Memory-Centric Storage System

HDFS, S3, GlusterFS

Mesos Cluster resource manager, multi-tenancy

Supported Release

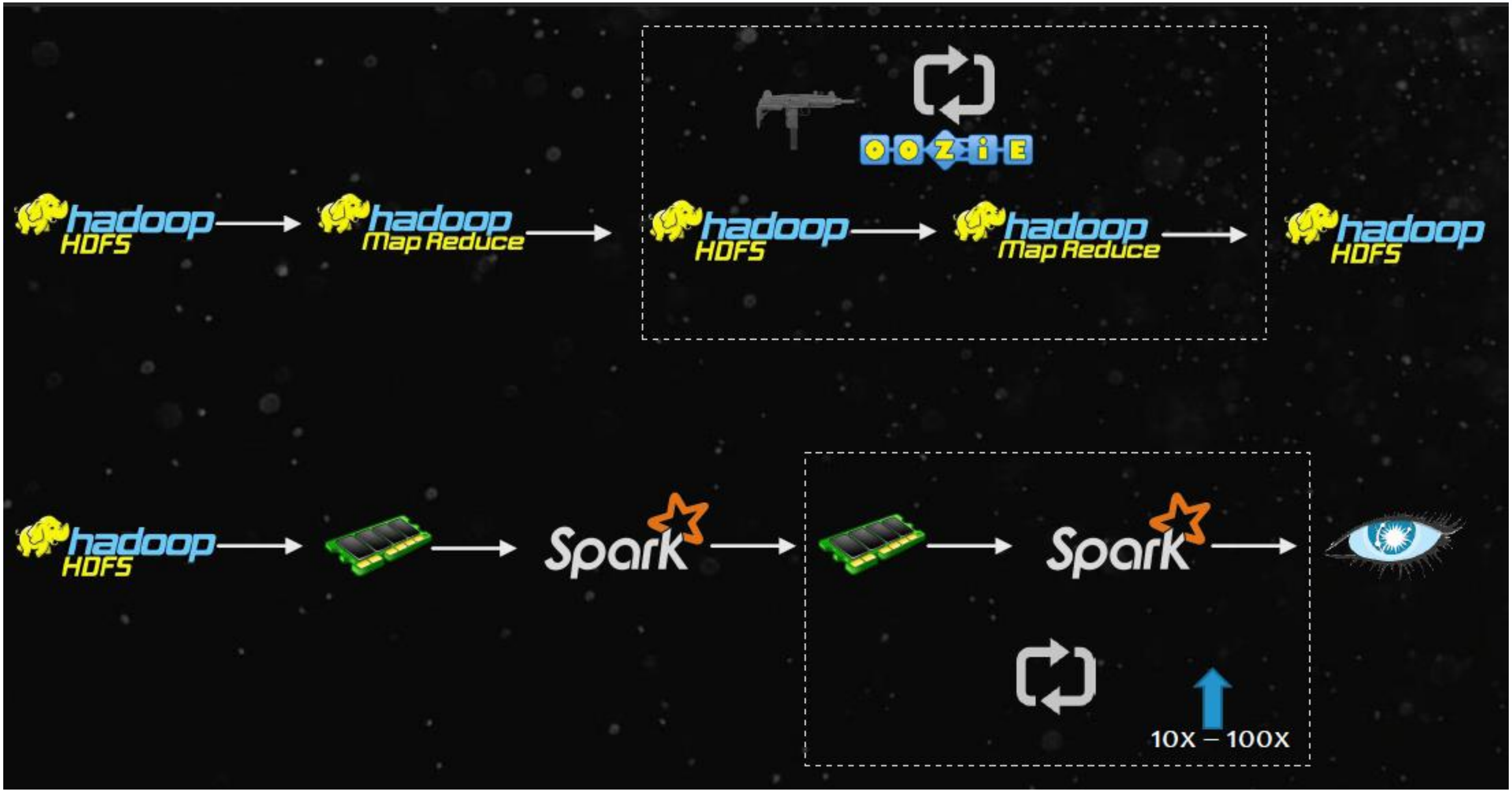
In Development

Related External Project

Spark vs MapReduce

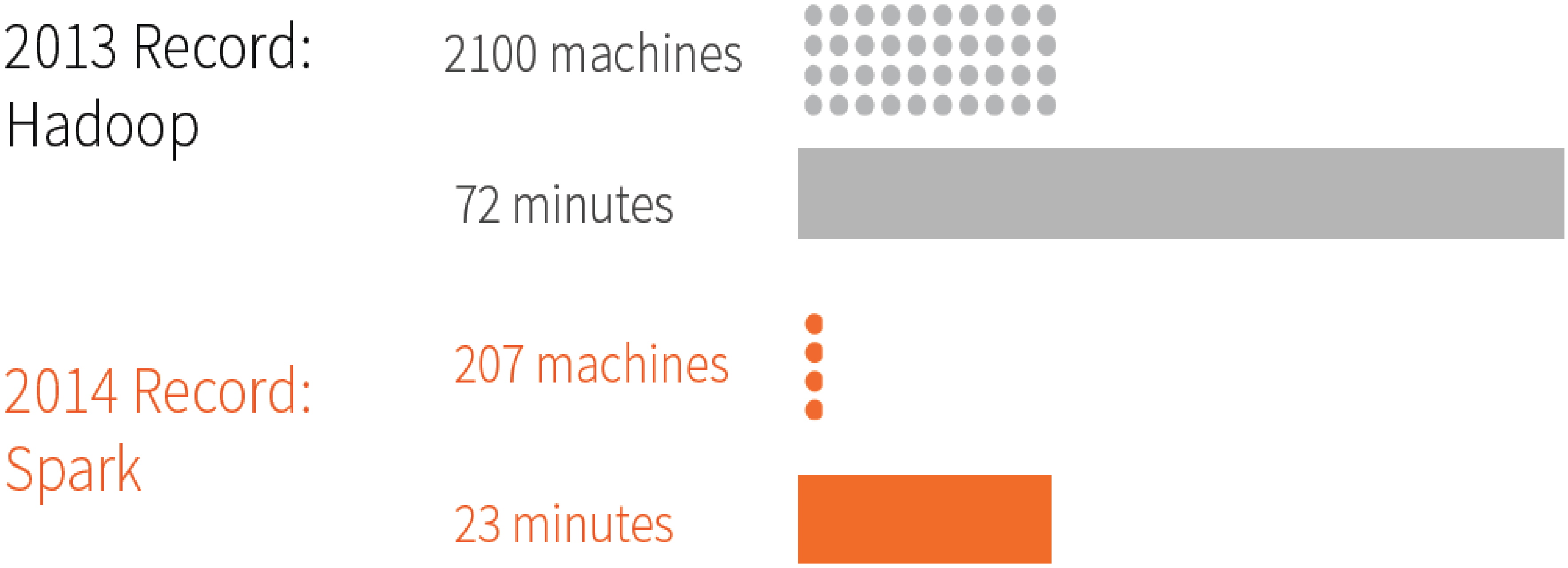
MapReduce	Spark
数据存储结构：磁盘hdfs文件系统的split	使用内存构建弹性分布式数据集RDD，对数据进行运算和cache
编程范式：Map + Reduce	DAG(有向无环图)：Transformation + action
计算中间数据落磁盘，io及序列化、反序列化代价大	计算中间数据在内存中维护，存取速度是磁盘的多个数量级
Task以进程的方式维护，任务启动就有数秒	Task以线程的方式维护，对小数据集的读取能达到亚秒级的延迟

MapReduce vs Spark



MapReduce vs Spark

On-Disk Sort Record: Time to sort 100TB



2013 Record: Hadoop

2100 machines

72 minutes

2014 Record: Spark

207 machines

23 minutes

http://www.csdn.net/article/2014-10-11/2822041-spark-breaks-previous-large-scale-sort-record

云帆大数据网络课程

- 云帆大数据是国内首家坚持实时在线授课、提供高端开发课程网络培训机构。采用新兴的互联网教育模式，坚持实时在线授课模式，既继承传统教育的学习交流特点，又发挥互联网的无处不在的时空特性，将天南地北有志向的人才组织在一起学习交流，使原先孤立的个体学习，组合成有组织的学习探讨，并且把原先的学费用降低到十分之一左右，使更多的人能学习到最新的高端课程技术。云帆大数据同时是一个平台，如果你是一个学员，可以尽情的学习和交流；如果你是一个有梦想有才华的人，可以联系我们，给你提供一片驰骋的原野。

- 云帆大数据已推出国内首家《企业级Hadoop 2.x 应用开发课程》、《企业级Hadoop 2.x 项目实战课程》和《企业级Spark 1.x 应用开发课程》，更多其他详情，请登录我们的培训网站http://www.clodyhadoop.com。

实时在线授课，专业课程辅导