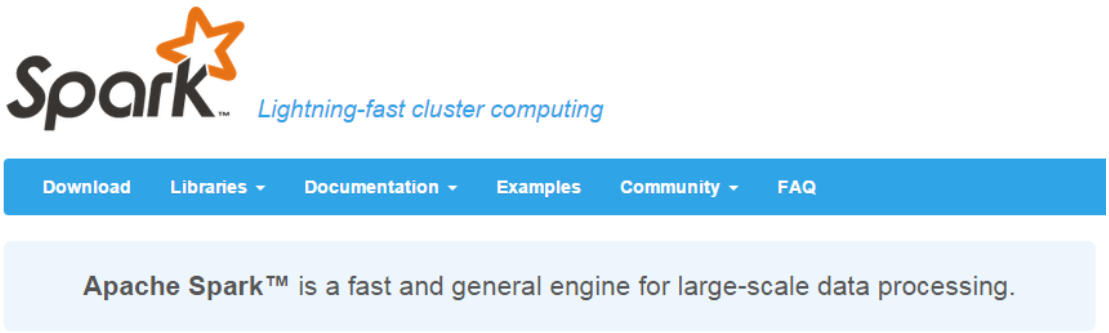
一、 Spark 概述

1. 什么是 Spark（官网：http://spark.apache.org）

Hadoop 解决海量数据的存储和计算问题；Spark 是计算引擎，只实现了计算功能。



Spark 是一种快速、通用、可扩展的大数据分析引擎。2009 年诞生于加州大学伯克利分校 AMPLab；2010 年开源；2013 年 6 月成为 Apache 孵化项目；2014 年 2 月成为 Apache 顶级项目。目前 Spark 生态系统已发展成一个包含多个子项目的集合，其中包含 SparkSQL、Spark Streaming、GraphX、MLlib 等。Spark 是基于内存的大数据并行计算框架，提高了大数据环境下数据处理的实时性，同时保证了高容错性和高可伸缩性。Spark 得到了众多大数据公司的支持，包括 Hortonworks、IBM、Intel、Cloudera、MapR、Pivotal、百度、阿里、腾讯、京东、携程、优酷土豆等。当前百度的 Spark 已应用于凤巢、大搜索、直达号、百度大数据等业务；阿里利用 GraphX 构建了大规模的图计算和图挖掘系统，实现了很多生产系统的推荐算法；腾讯 Spark 集群已达到 8000 台的规模，是当前已知的世界上最大的 Spark 集群。

2. 为什么要学 Spark

中间结果输出：基于 MapReduce 的计算引擎将中间结果写入环形缓冲区，若环形缓冲区满，则溢写到磁盘（IO 读写性能消耗较高），在磁盘上进行存储和容错。出于任务管道承接的考虑，当一些查询反映到 MapReduce 任务时会产生多个 Stage，而这些串联的 Stage 又依赖于底层文件系统（如 HDFS）来存储每一个 Stage 的中间结果输出。Spark 将中间结果写入内存，若内存不够大，则写入磁盘（非常灵活）。

Spark 是 MapReduce 的替代方案，且兼容 HDFS、Hive，可融入 Hadoop 生态系统，以弥补 MapReduce 的不足（大量磁盘 IO 消耗性能；计算较慢；）。

3. Spark 特点

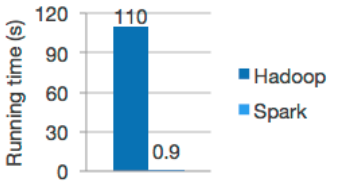
3.1. 快

与 Hadoop 的 MapReduce 相比，Spark 基于内存的运算要快 100 倍以上，基于硬盘的运算也要快 10 倍以上。Spark 实现了高效 DAG 执行引擎，可以基于内存来高效处理数据流。

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

3.2. 易用

Spark 支持 Java、Python 和 Scala 的 API，还支持超过 80 种高级算子可以快速构建并行应用。而且 Spark 支持交互式 Python 和 Scala 的 shell，可以非常方便地在这些 shell 中使用 Spark 集群来验证解决问题的方法。

Ease of Use

Write applications quickly in Java, Scala, Python, R.

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

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

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

Word count in Spark's Python API

3.3. 通用

Spark 提供了统一的解决方案用于批处理、交互式查询（Spark SQL）、实时流式处理（Spark Streaming）、机器学习（Spark MLlib）和图计算（GraphX），这些不同类型的处理都可以在同一个应用中无缝使用。Spark 统一的解决方案非常具有吸引力，毕竟任何公司都想用统一的平台去处理遇到的问题，减少开发和维护的人力成本和部署平台的物力成本。

3.4. 兼容性

Spark 可以非常方便的与其他开源产品融合。比如，Spark 可以使用 Hadoop 的 YARN 或 Apache 的 Mesos（细粒度）作为资源管理调度器；可以处理所有 Hadoop 支持的数据，包括 HDFS、HBase 和 Cassandra 等。这对于已经部署 Hadoop 集群的用户特别重要，因为不需要做任何数据迁移就可以使用 Spark 的强大处理能力。Spark 也可以不依赖第三方资源管理调度器，它实现了 Standalone 作为其内置的资源管理和调度框架，进一步降低 Spark 的使用门

槛，使得所有人都可以非常容易的部署和使用 Spark。此外 Spark 还提供了在 EC2 上部署 Standalone 集群的工具。

Runs Everywhere

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

You can run Spark using its standalone cluster mode, on EC2, on Hadoop YARN, or on Apache Mesos. Access data in HDFS, Cassandra, HBase, Hive, Tachyon, and any Hadoop data source.