Spark入门部分

整体上分为spark基本介绍、spark集群搭建、跑第一个spark demo

学习新东西的一般套路：

1. 为什么要学习（spark） ？

A. 基于内存计算

Hadoop将数据存放在datanode的磁盘上，存在一定的i/o瓶颈和内存相比比较慢，尤其是机器学习，不断地迭代数据，会消耗大量的IO

B. 支持实时计算

存在着这样的使用场景：来一批数据就分析一次数据。数据流就像水流一样 ， 新产生的数据立即能够进行分析处理。而hadoop只能用于离线计算

C. 通用计算平台

支持流式处理、SQL交互式、机器学习、图计算。能够很好的集成hadoop、kafka等

等其它大数据工具

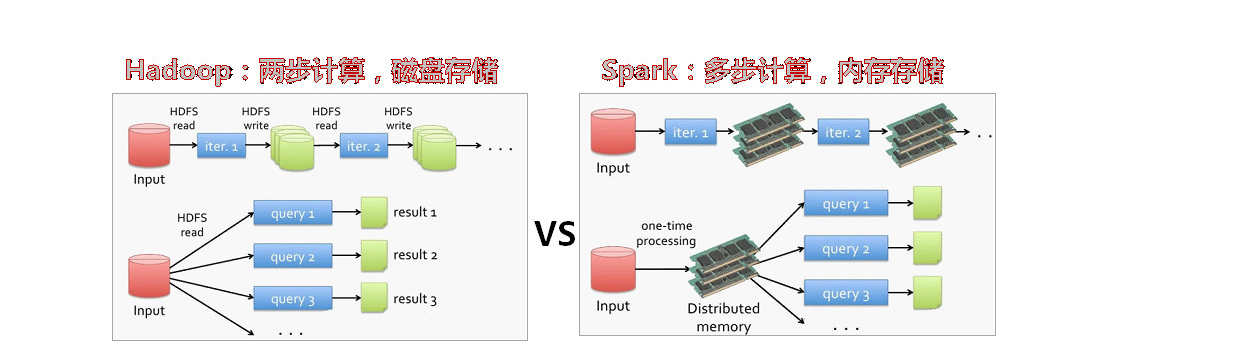
是mapreduce的替代方案，弥补maprudece的不足。总之，作为大数据工程师必须掌握的技能

1. 学新东西(spark)的什么就能满足需求 ？

Spark core / spark sql dataframe / spark streaming / Mllib / 和其它框架集成以及项目

1. 怎么学习、跟谁学习？

源头（创始人、官网）、不断总结

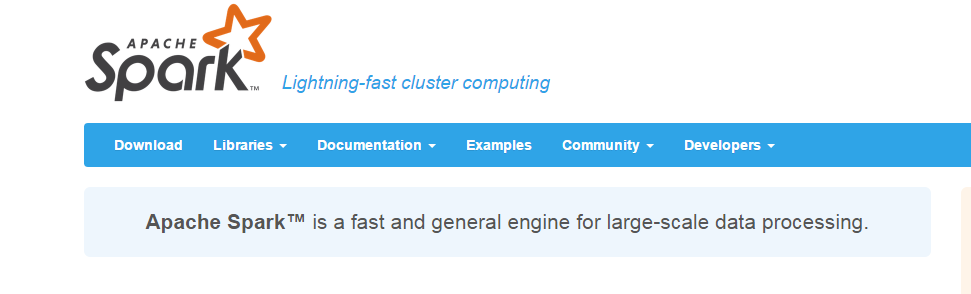


spark基本介绍

Spark 官方网站： <http://spark.apache.org/>

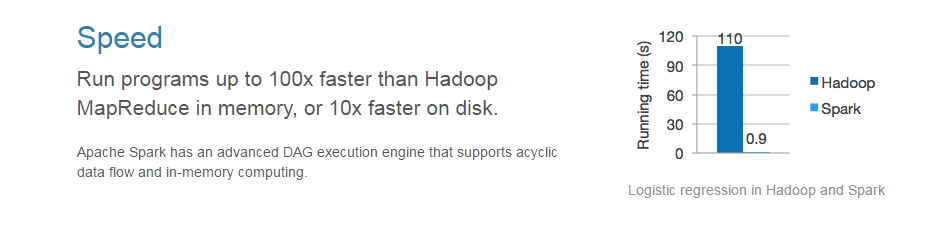
Github 地址：https://github.com/apache/spark

1. Spark是什么 ？



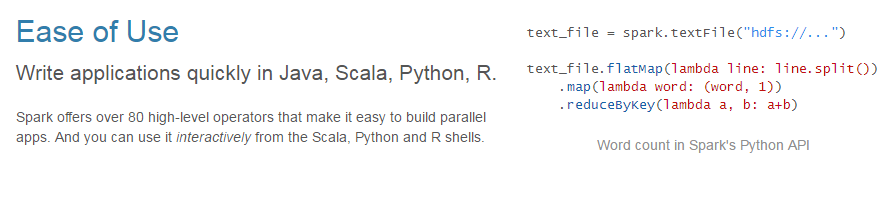
Spark是一种快速、通用、可扩展的大数据分析引擎，2009年诞生于加州大学伯克利分校AMPLab，2010年开源，2013年6月成为Apache孵化项目，2014年2月成为Apache顶级项目（hadoop在2005 年被 Apache Software Foundation 公司引入作为[Lucene](http://baike.baidu.com/view/371811.htm" \t "http://baike.baidu.com/_blank)的子项目[Nutch](http://baike.baidu.com/view/46642.htm" \t "http://baike.baidu.com/_blank)的一部分）。目前，Spark生态系统已经发展成为一个包含多个子项目的集合，其中包含SparkSQL、Spark Streaming、GraphX、MLlib等子项目，Spark是基于内存计算的大数据并行计算框架。Spark基于内存计算，提高了在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，允许用户将Spark部署在大量廉价硬件之上，形成集群。Spark得到了众多大数据公司的支持，这些公司包括Hortonworks、IBM、Intel、Cloudera、MapR、Pivotal、百度、阿里、腾讯、京东、携程、优酷土豆。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集群达到8000台的规模，是当前已知的世界上最大的Spark集群。

1. Spark 特点

快

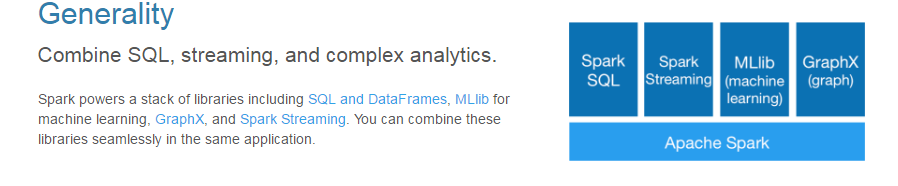
与Hadoop的MapReduce相比，Spark基于内存的运算要快100倍以上，基于硬盘的运算也要快10倍以上。Spark实现了高效的DAG执行引擎，可以通过基于内存来高效处理数据流。

易用



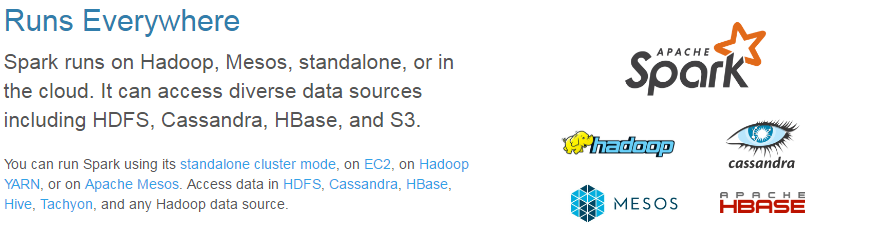
Spark支持Java、Python和Scala的API，还支持超过80种高级算法，使用户可以快速构建不同的应用。而且Spark支持交互式的Python和Scala、R的shell，可以非常方便地在这些shell中使用Spark集群来验证解决问题的方法。

通用



Spark提供了统一的解决方案。Spark可以用于批处理、交互式查询（Spark SQL）、实时流处理（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。

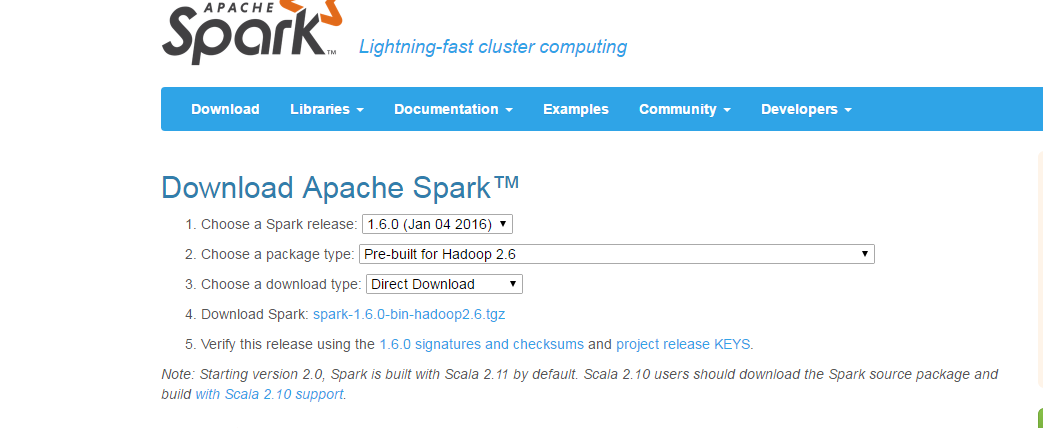
兼容性



Spark可以非常方便地与其他的开源产品进行融合。比如，Spark可以使用Hadoop的YARN和Apache Mesos作为它的资源管理和调度器，器，并且可以处理所有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也可以不依赖于第三方的资源管理和调度器，它实现了Standalone作为其内置的资源管理和调度框架，这样进一步降低了Spark的使用门槛，使得所有人都可以非常容易地部署和使用Spark。此外，Spark还提供了在EC2上部署Standalone的Spark集群的工具。

Spark 集群安装

1. 下载spark安装包



安装步骤：

1. 将下载好的spark安装包放到linux上面(bigdata01上面)

2. 解压安装包到指定位置

tar -zxvf spark-1.6.0-bin-hadoop2.6.tgz -C /usr/local/apps

1. 进入到spark安装目录，进行相关配置 : cd /usr/local/apps/spark-1.6.0-bin-hadoop2.6

A:cd bin , cd sbin 分别进行观察，总结出两个目录的作用

B: cd .. , cd conf/ , mv spark-env.sh.template spark-env.sh , vi spark-env.sh

在该配置文件下配置如下配置：

export JAVA\_HOME=/usr/local/apps/jdk1.7

export SPARK\_MASTER\_IP=bigdata01

export SPARK\_MASTER\_PORT=7077

保存退出

1. 重命名并修改slaves.template文件

mv slaves.template slaves , vi slaves

在该文件中添加子节点所在的位置（Worker节点）

bigdata02

bigdata03

保存退出

5. 将配置好的Spark拷贝到其他节点上 cd /usr/local/apps

scp -r spark-1.6.0-bin-hadoop2.6/ bigdata02:/usr/local/apps

scp -r spark-1.6.0-bin-hadoop2.6/ bigdata03:/usr/local/apps

1. Spark集群配置完毕，目前是1个Master，2个Work，在bigdata01上启动Spark集群

date -s "2017-02-07 19:45:00" // 集群时间同步

/usr/local/apps/spark-1.6.0-bin-hadoop2.6/sbin/start-all.sh

/usr/local/apps/spark-1.6.0-bin-hadoop2.6/sbin/start-stop.sh(关闭集群)

1. 启动后执行jps命令，主节点上有Master进程，其他子节点上有Work进行，登录Spark管理界面查看集群状态（主节点）：[http://bigdata01:8080/](http://node1.itcast.cn:8080/)
2. 运行sparkshell：

/usr/local/apps/spark-1.6.0-bin-hadoop2.6/bin/spark-shell --master spark://bigdata01:7077 \

--total-executor-cores 2 --executor-memory 512m

第一个spark demo

1. 执行第一个spark程序：

/usr/local/apps/spark-1.6.0-bin-hadoop2.6/bin/spark-submit \

--class org.apache.spark.examples.SparkPi \

--master spark://bigdata01:7077 \

--executor-memory 512m \

--total-executor-cores 2 \

/usr/local/apps/spark-1.6.0-bin-hadoop2.6/lib/spark-examples-1.6.0-hadoop2.6.0.jar \

100

spark 安装包文件夹说明

1. bin : 里面存放的是操作spark集群的命令
2. conf: 存放的是配置文件，文件的配置包括slaves , spark运行环境配置等等
3. data: 存放的是机器学习demo会用到的数据
4. ec2: 文件夹下面包含了spark-ec2脚本，用来在Amazon EC2环境下启动spark集群
5. examples: spark自带的小demo，建议大家学完spark以后一个一个去研究
6. lib: 存放的都是jar包
7. Licenses ：存放的都是许可文件（这个文件夹不用关心）
8. logs: 运行时生成的文件夹，存放的都是运行期的日志
9. python: 对python的支持
10. R: 对R语言的支持
11. sbin: 关闭、打开集群的shell
12. work: 工作目录，运行后生成。存放的就是提交的应用记录

spark 第一天作业

1. 讲集群搭建起来，运行，并在web界面观察
2. 自行选择5个org\apache\spark\examples包下面的demo 在集群里面跑，并观察效果
3. Spark安装目录下面的文件夹一个一个cd 进去观察，理解每一个文件夹的作用