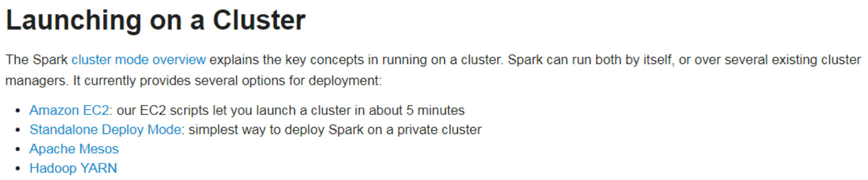
环境搭建

1. 预备环境
    1. JDK
    2. Scala
    3. Hadoop
2. 部署模式



Launching on a Cluster

The Spark cluster mode overview explains the key concepts in running on a cluster. Spark can run both by itself, or over several existing cluster managers. It currently provides several options for deployment:

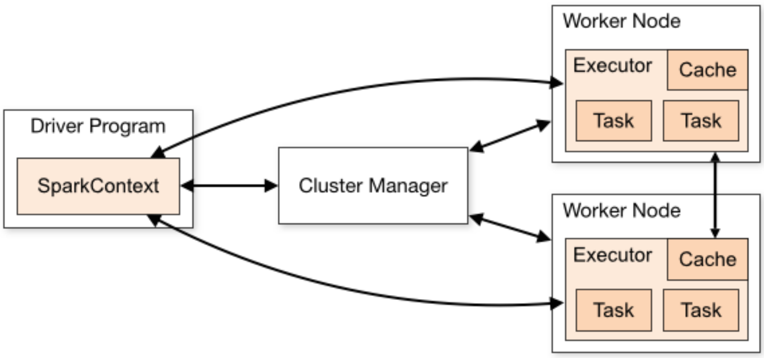
- Amazon EC2: our EC2 scripts let you launch a cluster in about 5 minutes
- Standalone Deploy Mode: simplest way to deploy Spark on a private cluster
- Apache Mesos
- Hadoop YARN

3. 环境搭建
    1. tar zxf spark-1.6.3-bin-hadoop2.6.tgz -C /opt
    2. vi ~/.bash_profile
        a) export SPARK_HOME
        b) export PATH=$SPARK_HOME/bin:$PATH
    3. spark 环境变量配置
        a) vi spark-env.sh

```
JAVA_HOME=/opt/jdk1.8.0_191
SCALA_HOME=/opt/scala-2.12.7
SPARK_MASTER_IP=master
SPARK_WORKER_MEMORY=1g
HADOOP_CONF_DIR=/opt/hadoop-2.6.5/etc/hadoop
```

        b) 修改 slaves
        c) scp -r /opt/spark-1.6.3-bin-hadoop2.6 root@slave1:/opt/
        d) sbin/start-all.sh
    4. 访问<IP>:8080

Spark 架构原理



1.Driver

进程

负责执行编写的 spark 程序

2.Master

进程，负责整个集群资源的调度、分配、监控等职责

3.Worker

进程

1.负责存储 RDD 的某个或某些 partition

2.启动其他进程或线程，对 RDD 的 partition 处理和计算

4.Executor

进程

启动多个 task 线程

4. Task

对 RDD 的 partition 进行并行计算