零基础学习Spark 1.x应用开发系列课程

Spark 1.x环境搭建

讲师-梦琪

版权声明

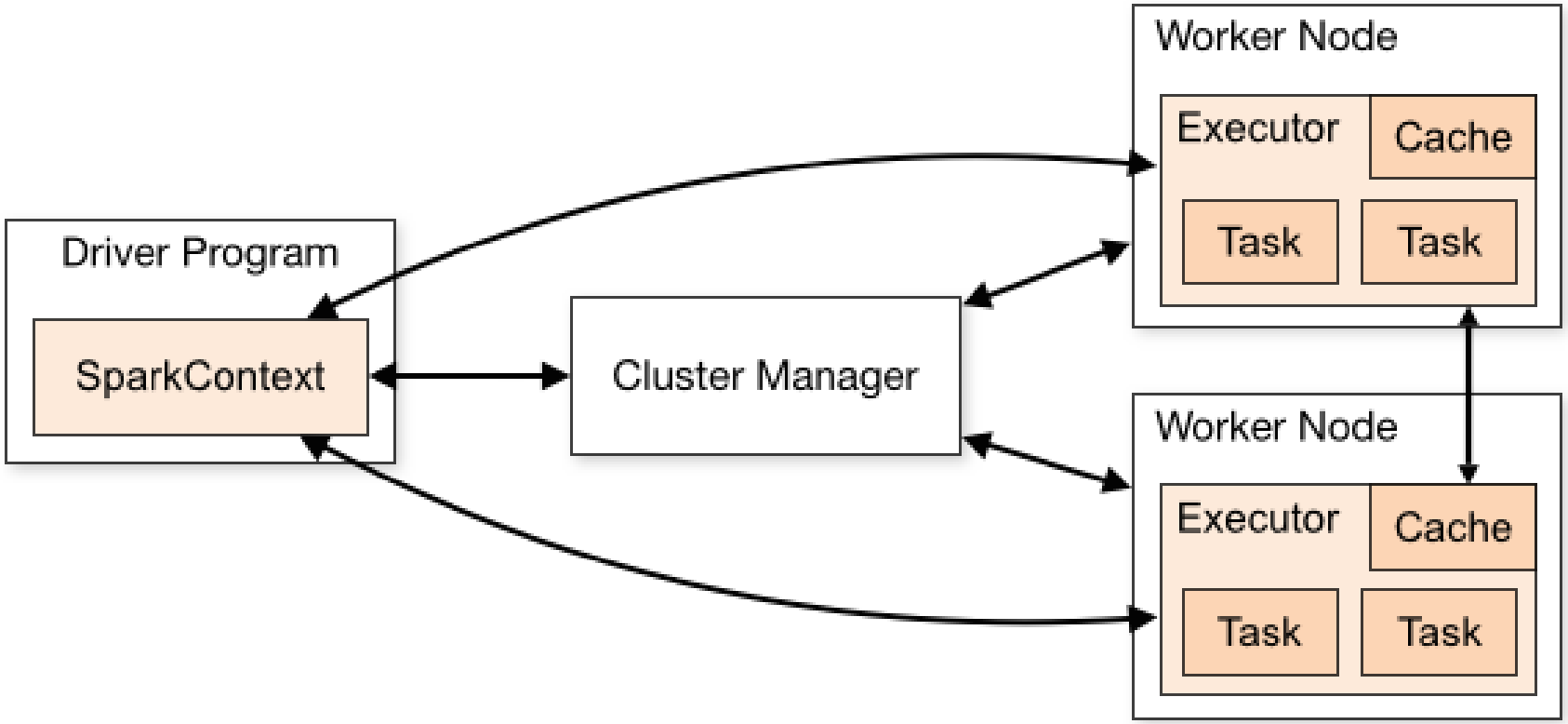
课程详情访问云帆大数据官方培训网站

http://www.cloudyhadoop.com

WAYS TO RUN SPARK

- ✓Local
- ✓Standalone
- ✓YARN
- ✓Mesos

Cluster Mode



Spark 1.x环境搭建步骤

- 安装JDK（建议JDK 7以上）
- 安装Scala（2.10.4）
- 安装Hadoop 2.x（至少HDFS）
- 安装Spark Standalone

Hadoop 2.x安装部署

◆下载

http://apache.dataguru.cn/hadoop/common/

◆解压

$ tar -zxvf hadoop-2.5.0.tar.gz

◆替换本地库

$ rm -rf ./$HADOOP_HOME/lib/native/

$ cp -r $HADOOP_SRC_HOME/hadoop-dist/target/hadoop-2.5.0/lib/native/* $HADOOP_HOME/lib/native/

◆修改配置文件（$HADOOP_HOME/etc/hadoop/目录下）

hadoop-env.sh、core-site.xml、hdfs-site.xml、yarn-site.xml、mapred-site.xml

◆ 注意点，**native下面的链接文件**

编辑配置文件

◆ 配置文件 **hadoop-env.sh**

```
export JAVA_HOME=/opt/modules/jdk1.7.0_67
```

◆ 配置文件 **core-site.xml**

```
<configuration>
        <property>                          ——指定NameNode主机名与端口号
                <name>fs.default.name</name>
                <value>hdfs://hadoop-yarn.dragon.org:8020</value>
        </property>
        <property>
                <name>hadoop.tmp.dir</name>
                <value>/opt/modules/hadoop-2.5.0/data/tmp</value>
        </property>
</configuration>
```

编辑配置文件

◆ 配置文件 **hdfs-site.xml**

```
<property>                    —— 设置HDFS的副本数
        <name>dfs.replication</name>
        <value>1</value>
</property>
```

◆ 配置文件 **slaves**

hadoop-yarn.dragon.org

Spark 1.x搭建部署

- 解压

  tar -zvxf spark-1.3.0-bin-2.6.0

- 配置环境变量

  export SPARK_HOME=/opt/modules/spark-1.3.0-bin-2.6.0

- 配置文件

  spark-env.sh spark-default.conf

- 启动

  start-all.sh

- 验证
  - jps
  - Web UI

Spark Core

WordCount

```
sc.textFile("data/README.md")
        .map(line => line.split("\t"))
        .map(_,1)
        .reduceByKey(_+_, 3)
        .collect()
```

Spark Core

A Resilient Distributed Dataset (RDD), the **basic abstraction** in Spark.

Represents an **immutable**, **partitioned** collection of elements

that can be operated on **in parallel**.

```
Internally, each RDD is characterized by five main properties:

 - A list of partitions
 - A function for computing each split
 - A list of dependencies on other RDDs
 - Optionally, a Partitioner for key-value RDDs (e.g. to say that the RDD is hash-partitioned)
 - Optionally, a list of preferred locations to compute each split on (e.g. block locations for
   an HDFS file)
```

Spark Core

Transformations

•Create a new dataset from and existing one.

•**Lazy** in nature. They are executed only when some action is performed.

•Example :
- map(func)
- filter(func)
- distinct() …

Actions

•Returns to the driver program a value or exports data to a storage system after performing a computation.

•Example:
- count()
- reduce(func)
- collect
- take()…

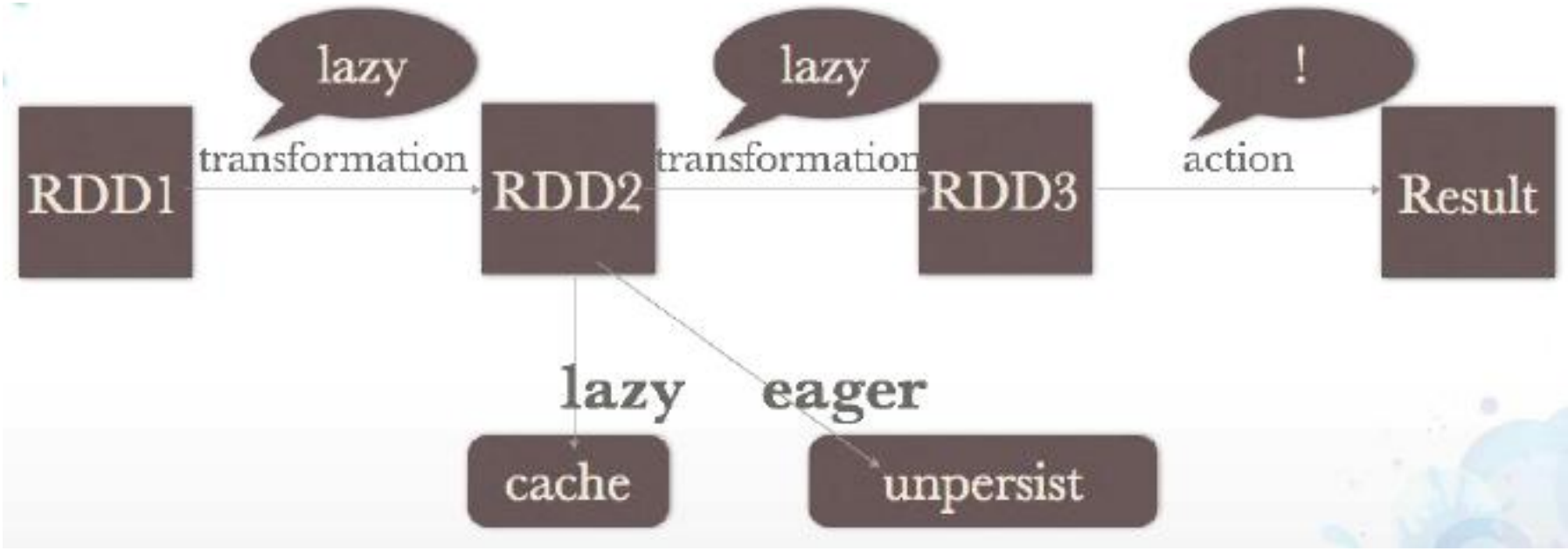
Persistence

•For caching datasets in-memory for future operations.

•Option to store on disk or RAM or mixed (Storage Level).

•Example:
- persist()
- cache()

Spark Core


lazy
lazy
!
RDD1
transformation
RDD2
transformation
RDD3
action
Result
lazy
eager
cache
unpersist

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