零基础学习Spark 1.x应用开发系列课程

Spark RDD详解

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

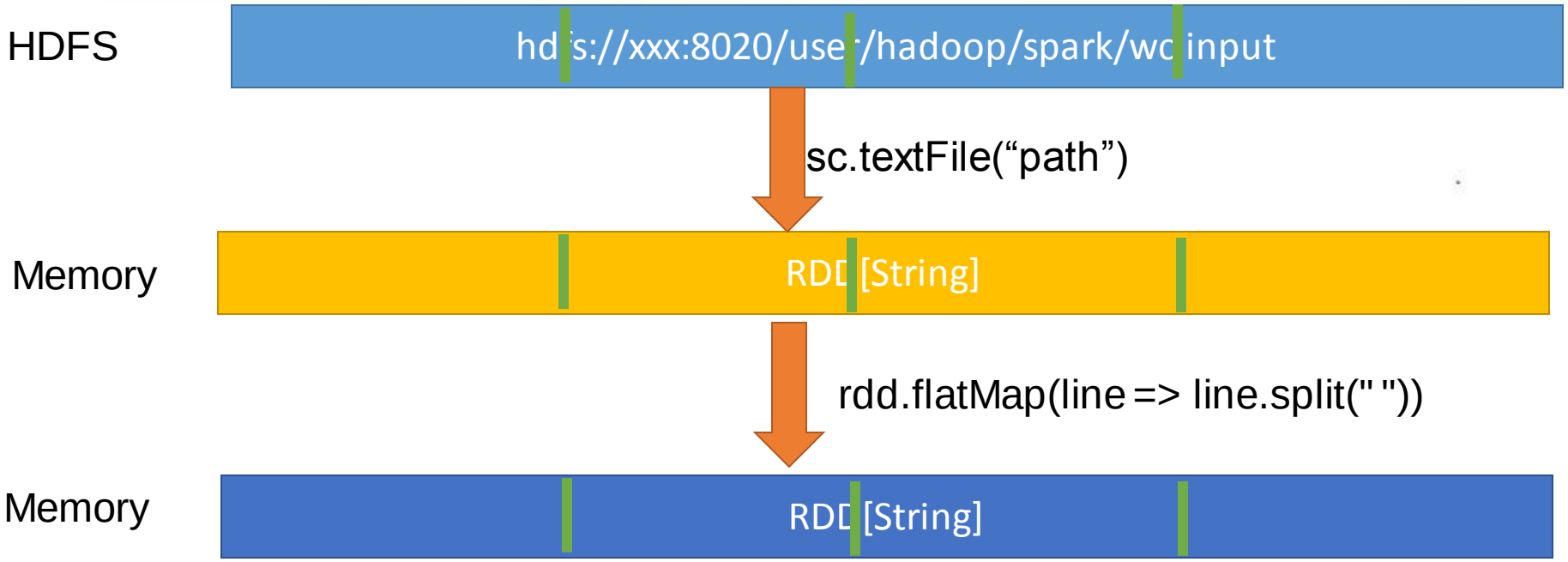
http://www.cloudyhadoop.com

Spark WordCount

```
val rdd=sc.textFile("hdfs://hadoop-spark.dragon.org:8020/user/hadoop/spark/wc.input")
```

```
val wordcount=rdd.flatMap(_.split(" "))
                 .map((_,1))
                 .reduceByKey(_ + _)
wordcount.collect()
```

```
val wordsort=wordcount.map(x=>(x._2,x._1))
                      .sortByKey(false)
                      .map(x=>(x._2,x._1))
wordsort.collect()
```


HDFS
hdfs://xxx:8020/user/hadoop/spark/wc input
sc.textFile("path")
Memory
RDD [String]
rdd.flatMap(line => line.split(" "))
Memory
RDD [String]

Spark Core—RDD

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—RDD

RDD: Resilient Distributed Dataset

RDD的特点：

1、A list of **partitions**

一系列的分片：比如说64M一片；类似于Hadoop中的split；

2、A **function** for computing each split

在每个分片上都有一个函数去迭代/执行/计算它

3、A list of **dependencies** on other RDDs

一系列的依赖：RDDa转换为RDDb，RDDb转换为RDDc，那么RDDc就依赖于RDDb，RDDb就依赖于RDDa

4、Optionally, a **Partitioner** for key-value RDDs (e.g. to say that the RDD is hash-partitioned)

对于key-value的RDD可指定一个partitioner，告诉它如何分片；常用的有hash，range

5、Optionally, a list of **preferred location(s)** to compute each split on (e.g. block locations for an HDFS file)

要运行的计算/执行最好在哪(几)个机器上运行。数据本地性。

为什么会有哪几个呢？

比如：hadoop默认有三个位置，或者spark cache到内存是可能通过StorageLevel设置了多个副本，所以一个partition可能返回多个最佳位置。

Spark Core—RDD

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()…

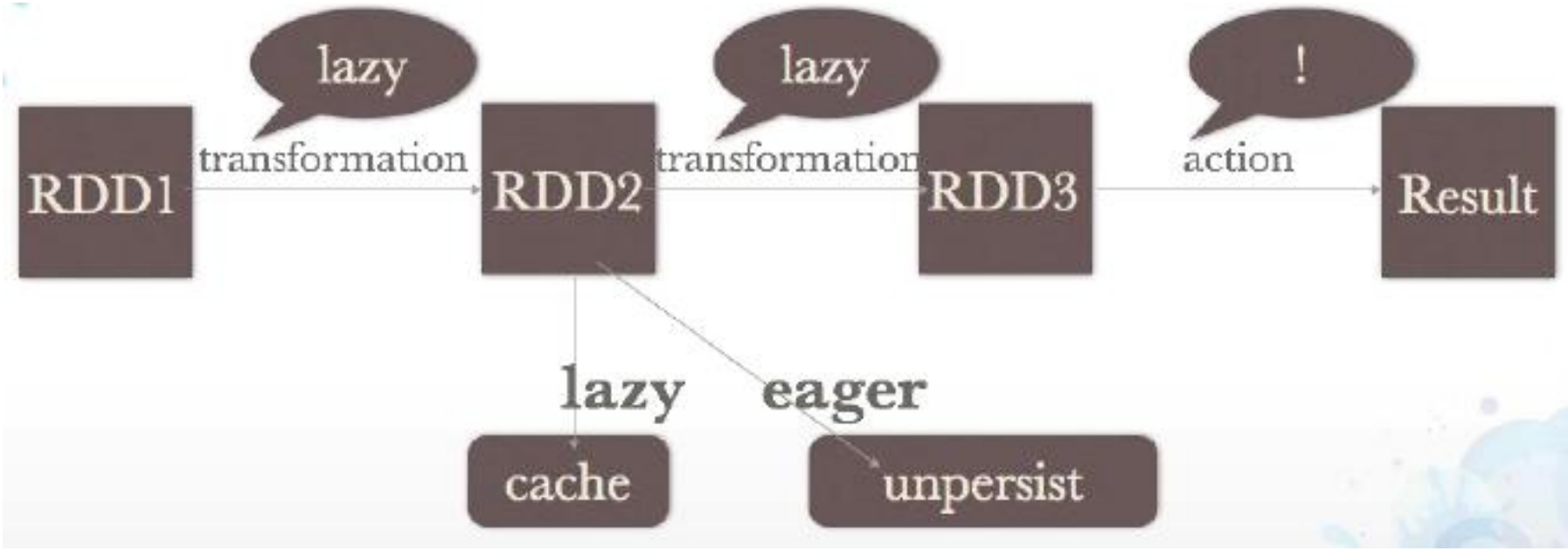
Persistence

•For caching datasets in-memory for future operations.

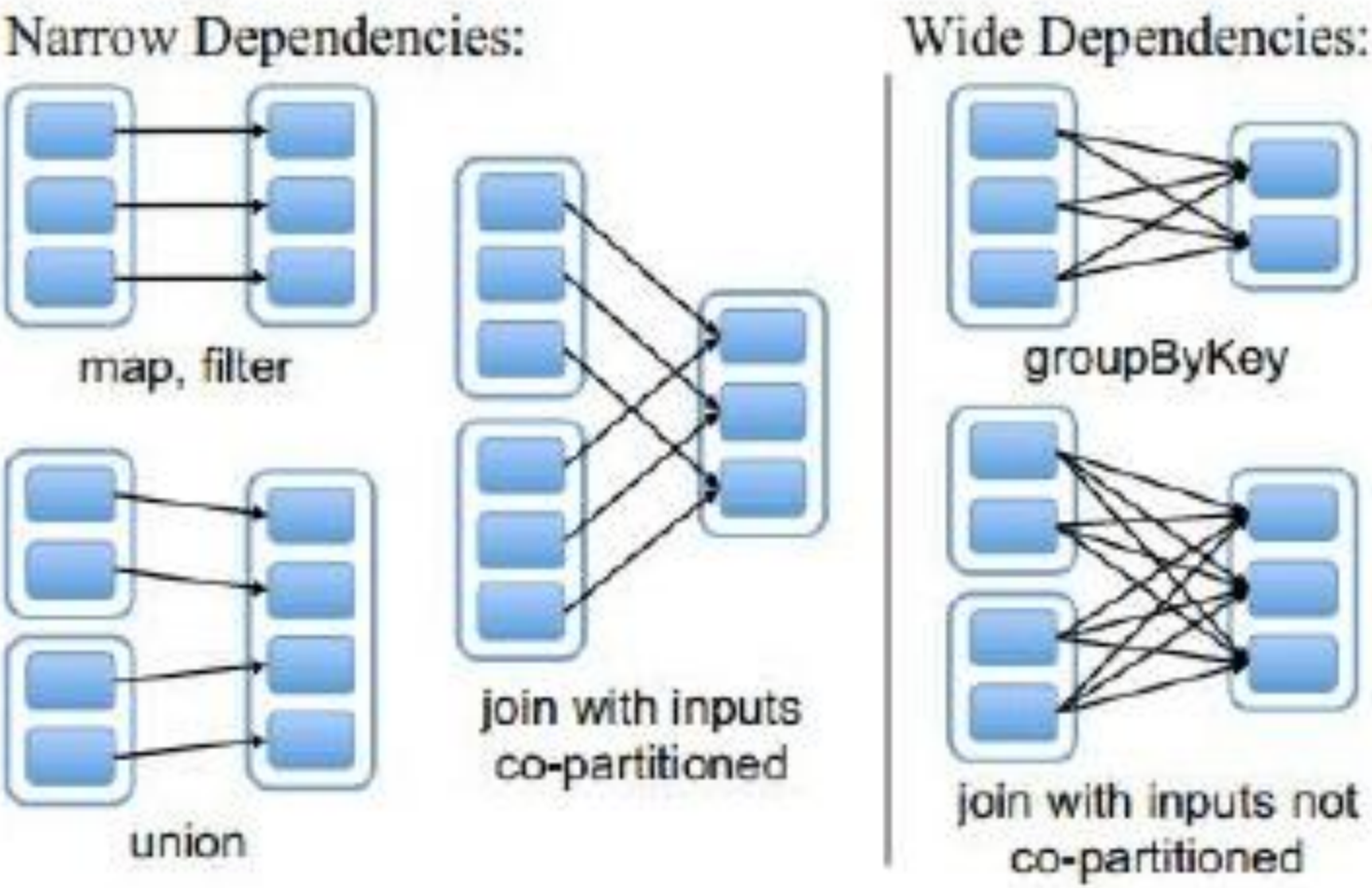
•Option to store on disk or RAM or mixed (Storage Level).

•Example:
- persist()
- cache()

Spark Core—RDD



RDD Dependencies



RDD Dependencies

- 窄依赖（ narrow dependencies ）
  - 子 RDD 的每个分区依赖于常数个父分区（即与数据规模无关）
  - 输入输出一对一的算子，且结果 RDD 的分区结构不变，主要是 map 、 flatMap
  - 输入输出一对一，但结果 RDD 的分区结构发生了变化，如 union 、 coalesce
  - 从输入中选择部分元素的算子，如 filter 、 distinct 、 subtract 、 sample
- 宽依赖（ wide dependencies ）
  - 子 RDD 的每个分区依赖于所有父 RDD 分区
  - 对单个 RDD 基于 key 进行重组和 reduce ，如 groupByKey 、 reduceByKey ；
  - 对两个 RDD 基于 key 进行 join 和重组，如 join

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.cloudyhadoop.com。

实时在线授课，专业课程辅导