零基础学习Spark 1.x应用开发系列课程

RDD操作详解

讲师-梦琪

版权所有云帆大数据 www.cloudyhadoop.com

版权声明

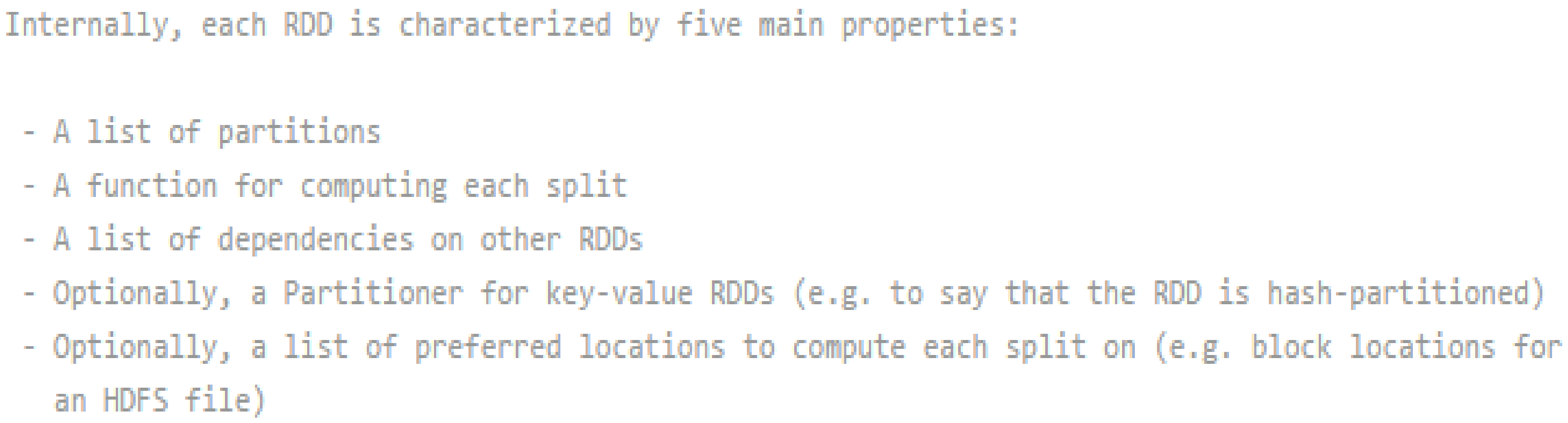
【声明】本视频和讲义等均为云帆大数据网络课程的教学资料，所有资料只能在课程内使用，不允许在课程以外范围散播，违者将可能被追究法律和经济责任。

课程详情访问云帆大数据官方培训网站

http://www.cloudyhadoop.com

Spark Core—RDD

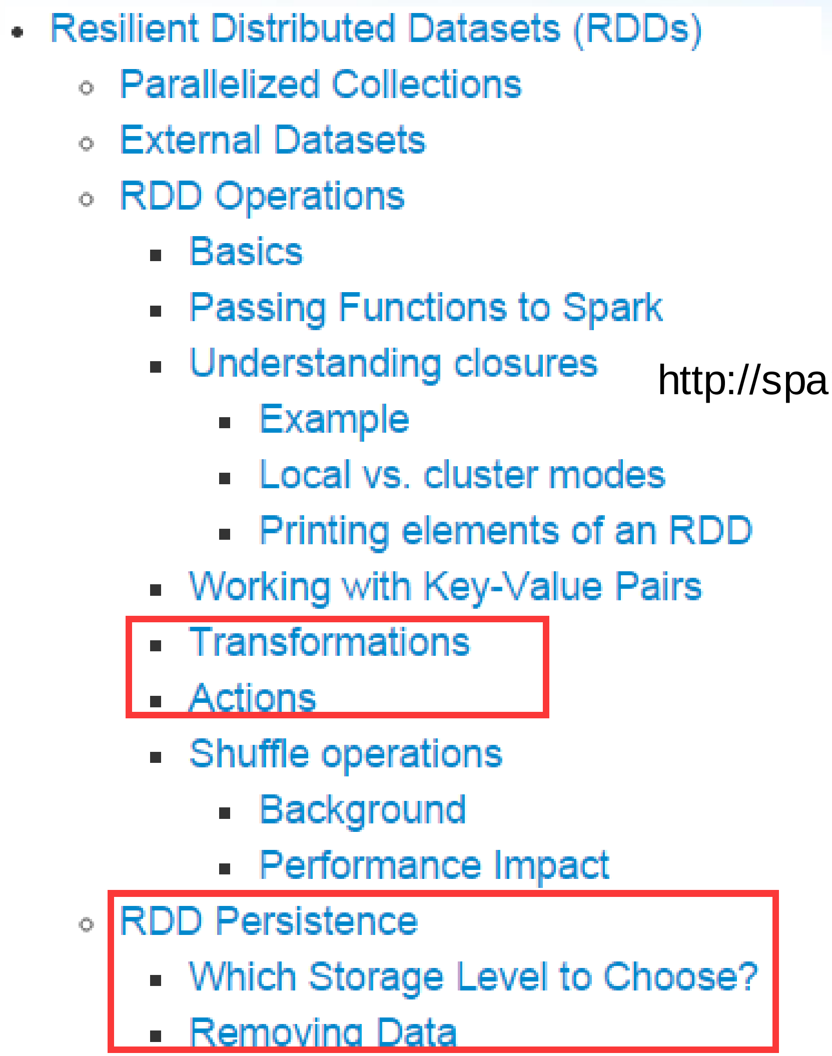
A Resilient Distributed Dataset (RDD), the **basic abstraction** in Spark. Represents an **immutable**, **partitioned** collection of elements that can be operated on **in parallel**.



```
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)
```

Resilient Distributed Datasets



- Resilient Distributed Datasets (RDDs)
  - Parallelized Collections
  - External Datasets
  - RDD Operations
    - Basics
    - Passing Functions to Spark
    - Understanding closures
      - Example
      - Local vs. cluster modes
      - Printing elements of an RDD
    - Working with Key-Value Pairs
    - Transformations
    - Actions
    - Shuffle operations
      - Background
      - Performance Impact
  - RDD Persistence
    - Which Storage Level to Choose?
    - Removing Data

http://spark.apache.org/docs/1.3.1/programming-guide.html

create RDDs

- ◆ **Parallelized Collections**
- ◆ **External Datasets**

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()…

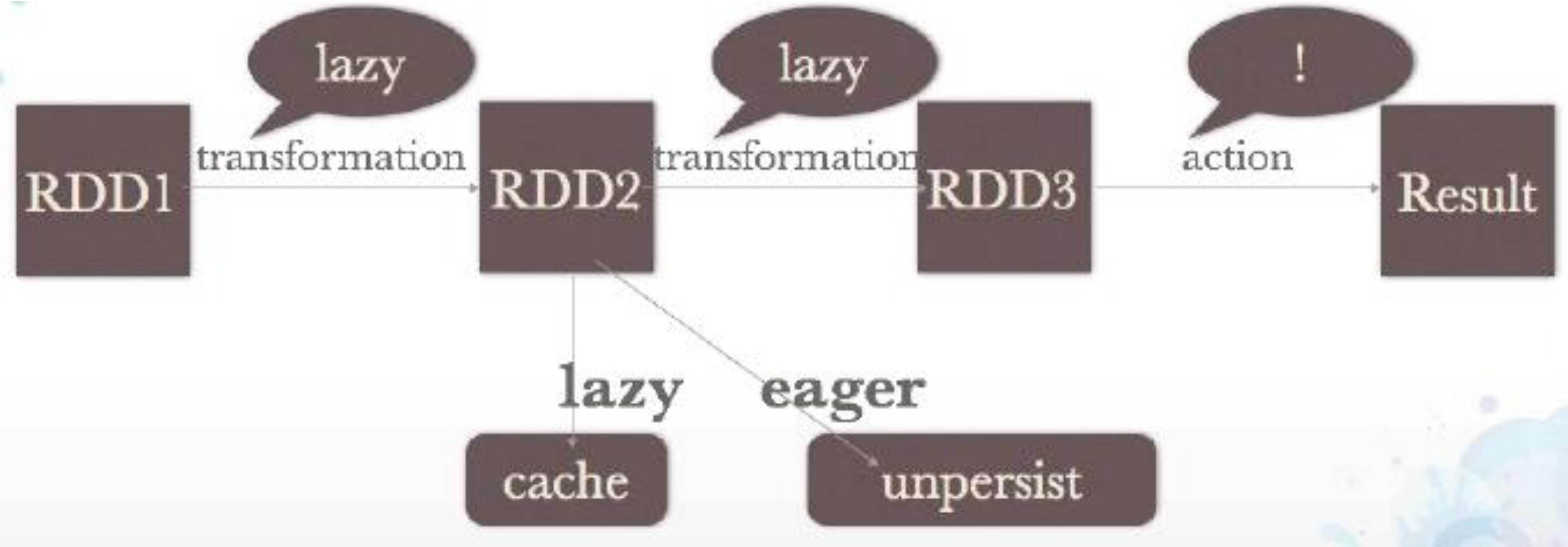
Persistence

- For caching datasets in-memory for future operations.

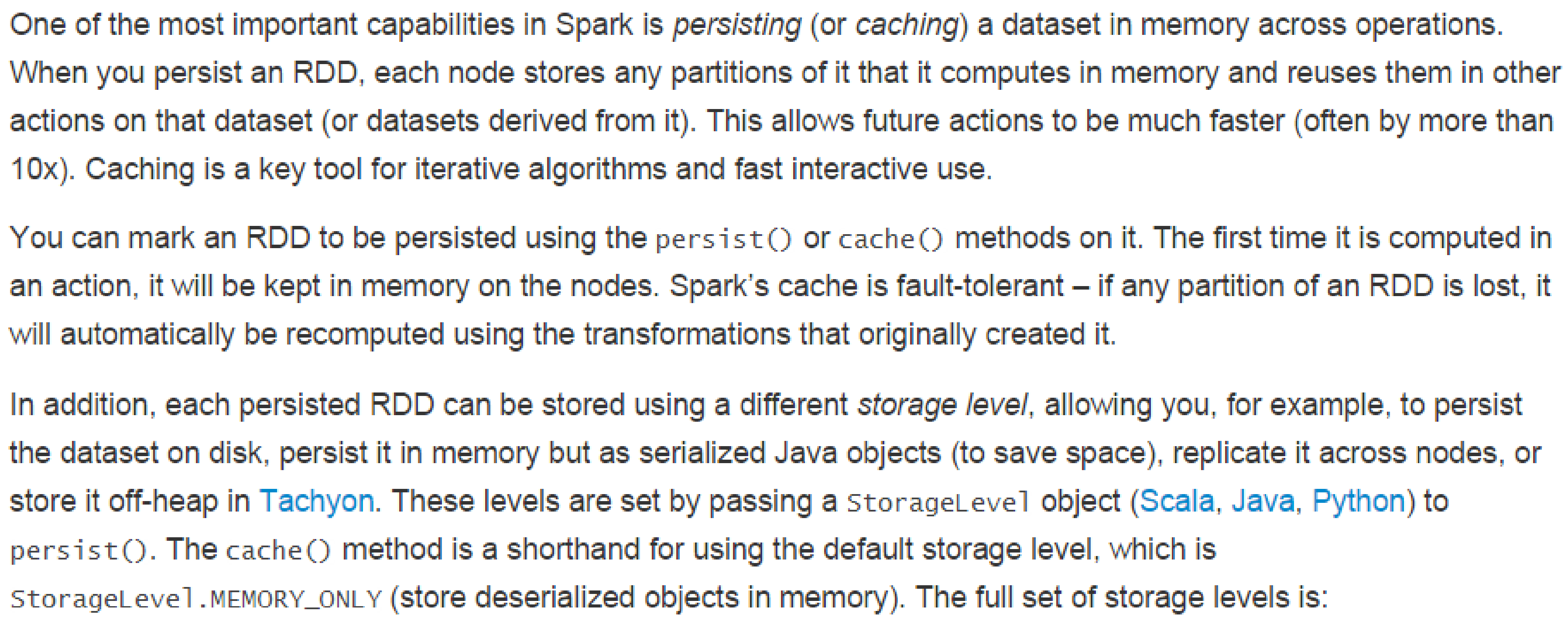
- Option to store on disk or RAM or mixed (Storage Level).

- Example:
  - persist()
  - cache()

Spark Core—RDD



RDD Persistence



One of the most important capabilities in Spark is *persisting* (or *caching*) a dataset in memory across operations. When you persist an RDD, each node stores any partitions of it that it computes in memory and reuses them in other actions on that dataset (or datasets derived from it). This allows future actions to be much faster (often by more than 10x). Caching is a key tool for iterative algorithms and fast interactive use.

You can mark an RDD to be persisted using the `persist()` or `cache()` methods on it. The first time it is computed in an action, it will be kept in memory on the nodes. Spark's cache is fault-tolerant – if any partition of an RDD is lost, it will automatically be recomputed using the transformations that originally created it.

In addition, each persisted RDD can be stored using a different *storage level*, allowing you, for example, to persist the dataset on disk, persist it in memory but as serialized Java objects (to save space), replicate it across nodes, or store it off-heap in Tachyon. These levels are set by passing a `StorageLevel` object (Scala, Java, Python) to `persist()`. The `cache()` method is a shorthand for using the default storage level, which is `StorageLevel.MEMORY_ONLY` (store deserialized objects in memory). The full set of storage levels is:

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