

MSAN 694 : Distributed Computing

Diane Woodbridge, Ph.D.
MSAN, University of San Francisco



Reviews

RDD Creation

RDD Operations - Transformations

RDD Operations - Actions

Run python scripts in Spark

Spark Interview Questions

What is Apache Spark?

Explain the key features of Spark.

What is RDD?

How to create RDD.

What is "partitions"?

Types of RDD operations?

What is "transformation"?

What is "action"?

Functions of "spark core"?

What is "spark context"?

What is an "RDD lineage"?

Which file systems does Spark support?

List the various types of "Cluster Managers" in Spark.

What is "YARN"?

What is "Mesos"?

What is a "worker node"?

What is an "accumulator"?

What is "Spark SQL" (Shark)?

What is "Spark Streaming"?

What is "GraphX"?

What is "MLlib"?

<https://www.edureka.co/blog/interview-questions/top-apache-spark-interview-questions-2016/>

Spark Interview Questions

What are the advantages of using Apache Spark over Hadoop MapReduce for big data processing?

What are the languages supported by Apache Spark for developing big data applications?

Can you use Spark to access and analyze data stored in Cassandra databases?

Is it possible to run Apache Spark on Apache Mesos?

How can you minimize data transfers when working with Spark?

Why is there a need for broadcast variables?

Name a few companies that use Apache Spark in production.

What are the various data sources available in SparkSQL?

What is the advantage of a Parquet file?

What do you understand by Pair RDD?

Is Apache Spark a good fit for Reinforcement learning?

<https://www.dezyre.com/article/top-50-spark-interview-questions-and-answers-for-2016/208>

Contents

Pair RDDs

Pair RDD Operations - Transformation

Contents

Pair RDDs

Pair RDD Operations - Transformation

Pair RDDs

Definition

- Key-value pairs – Commonly used for many operations including aggregations, ETL (extract, transform, and load) in Spark.
- Allow operations on each key in parallel or regroup data across the network such as `reduceByKey()`, `join()`, etc.



Pair RDDs

Creation

- Apply a map function with a lambda or user-defined function to have a pair of (Key, Value).
 - Key : could be a simple object (integer, string, etc.) to complex objects (tuples, etc.).
 - Value : could be a simple objects to data structures (lists, tuples, dictionaries, sets, etc.).

Example 1

From the “README.md” file,

- Extract all the words. (space separated)
- Generate key-value pairs of (Word, 1).

Contents

Pair RDDs

Pair RDD Operations - Transformation

Pair RDDs - Operations

Transformation on Pair RDDs

Function name	Purpose
keys()	Return an RDD of just the keys.
values()	Return an RDD of just the values.
sortByKey()	Return an RDD sorted by the key.
groupByKey()	Group values with the same key.
mapValues(func)	Apply a function to each value of a pair RDD without changing the key.
flatMapValues(func)	Apply a function that returns an iterator to each value of a pair RDD, and for each element returned, produce a key/value entry with the old key. Often used for tokenization.
reduceByKey(func)	Combine values with the same key.
combineByKey(createCombiner, mergeValue, mergeCombiners)	Combine values with the same key using a different result type.

<http://spark.apache.org/docs/latest/programming-guide.html>

<https://spark.apache.org/docs/0.6.2/api/core/spark/PairRDDFunctions.html>

Pair RDDs - Operations

Transformation on Pair RDDs

- `keys()` - Return an RDD of just the keys.
- `values()` - Return an RDD of just the values.
- `sortByKey()` - Return an RDD sorted by the key.

Example 2

From the “README.md” file,

- Extract all the words. (space separated)
- Generate key-value pairs of (Length of Word, Word).
- Try the following :
 - keys()
 - values()
 - sortByKey()

Example 2

Using Example 4, try the following.

- keys()
- values()
- sortByKey()

```
file = "../Data/README.md"

word = sc.textFile(file).flatMap(lambda x : x.split(" "))

len_word_pair = word.map(lambda x : (len(x),x))

len_word_pair.keys()

len_word_pair.values()

len_word_pair.sortByKey()
```

<http://spark.apache.org/docs/latest/programming-guide.html>

<https://spark.apache.org/docs/0.6.2/api/core/spark/PairRDDFunctions.html>

Pair RDDs - Operations

Transformation on Pair RDDs

- `groupByKey()`
 - Group data using the key.
 - Return an RDD of (Key, ResultIterable) pairs.

Example 3

Create a pair RDD with (length of a word, list of words) from “README.md”.

Pair RDDs - Operations

Transformation on Pair RDDs

- `mapValues(func)`
 - Pass each value in the key-value pair RDD through a map function without changing the keys.
 - Retain the original RDD's partitioning.

Example 4

From the “README.md” file,

- Extract all the words. (space separated)
- Generate key-value pairs of (Word, Occurrence).

Word Count

Pair RDDs - Operations

Transformation on Pair RDDs

- `flatMapValues(func)`
 - Pass each value in the key-value pair RDD through a flatMap function without changing the keys.
 - Retain the original RDD's partitioning.

Example 5

Create a list of (len, word) pairs from len_word_pair_group in Example 3.

```
len_word_pair_group =  
sc.textFile(file).flatMap(lambda x : x.split(" "))  
                        .map(lambda x : (len(x),x))
```

Pair RDDs - Operations

Transformation on Pair RDDs

- `reduceByKey(func)`
 - Similar to `reduce()`.
 - Run several parallel reduce operations, one for each key in the data set.
 - When called on a dataset of (Key, Val) pairs, returns a dataset of (Key, Val) pairs where the values for each key are aggregated using the given reduce function `func`.
- **Transformation** (Not an action) : Returns a new RDD consisting of each key and the reduced value for that key.

Example 6

From the “README.md” file,

- Extract all the words. (space separated)
- Generate key-value pairs of (Word, Occurrence) using reduceByKey().

Word Count

Example 4 vs Example 6

1) `word_rdd.groupByKey()
 .mapValues(lambda x : sum(x))`

2) `word_rdd.reduceByKey(lambda x,y : x+y)`

Word Count

Example 4 vs Example 6

To group values for the purpose of aggregation (such as sum or count for each key), using `reduceByKey()`, `foldByKey()` or `combineByKey()` will provide better performance.

- They combines the aggregation function before the shuffle.
 - ➔ Result in a reduced amount of data shuffled.

Pair RDDs - Operations

Transformation on Pair RDDs

- `combineByKey(createCombiner, mergeValue, mergeCombiners)`
 - Similar to `aggregate()`.
 - `createCombiner` - If it is new in a partition, create the initial value for the accumulator on the key.
 - `mergeValue` - If it is not new, apply the `mergeValue` function.
 - `mergeCombiners` - When merging the results from each partition, apply the `mergeCombiners` to merge the accumulators for the same key.

Example 7

Using `combineByKey()`, create pairs (Length of words, (Frequency, a list of words)) from “README.md”

Example 8

Which operations don't shuffle data?

- `mapValues()`
- `groupByKey()`
- `reduceByKey()`
- `combineByKey()`
- `sortByKey()`
- `flatMapValues()`
- ...

References

Distributed Computing with Spark, Reza Zadeh,
http://stanford.edu/~rezab/slides/bayacm_spark.pdf

Spark Online Documentation :
<http://spark.apache.org/docs/latest/>

Karau, Holden, et al. Learning spark: lightning-fast big data analysis. O'Reilly Media, Inc., 2015.

Zecevic, Petar, et al. Spark in Action, Manning, 2016.