8_RDD_persistance

February 21, 2022

[1]:
```
# Import SparkContext and SparkConf
from pyspark import SparkContext, SparkConf
```

[2]:
```
# Initialize spark
conf = SparkConf().setAppName("persistRDD1").setMaster("local[4]")
sc = SparkContext(conf=conf)
```

```
22/02/21 14:34:05 WARN Utils: Your hostname, ThinkCentre resolves to a loopback
address: 127.0.1.1; using 10.180.5.223 instead (on interface eno1)
22/02/21 14:34:05 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another
address
22/02/21 14:34:06 WARN NativeCodeLoader: Unable to load native-hadoop library
for your platform… using builtin-java classes where applicable
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use
setLogLevel(newLevel).
```

[3]:
```
# Create an RDD of market arrivals of various items (Karnataka)
marketRDD = sc.textFile("CommMktArrivals.csv", 2)
```

0.0.1 persist

Signature: marketRDD.persist(storageLevel=StorageLevel)

Description:

- Set this RDD's storage level to persist its values across operations after the first time it is computed.
- This can only be used to assign a new storage level if the RDD does not have a storage level set yet.
- If no storage level is specified defaults to (C{MEMORY_ONLY}).

  rdd = sc.parallelize([“b”, “a”, “c”]) rdd.persist().is_cached

[4]:
```
# Persist it to
import pyspark # Required to access StorageLevel class

# 1. one copy in MEMORY_ONLY
```

```
marketRDD.persist(pyspark.StorageLevel.MEMORY_ONLY)
marketRDD.take(2)
```

[4]:
```
['District Name,Taluk Name,Market
Name,Address,Telephone,Commodity,Year,Month,Arrival,Unit',
 'Bagalakot,Badami,BADAMI,SECRATRY A.P.M.C.BADAMI
BADAMI,220042,Bajra,2012,Jan,242,Quintal   ']
```

[5]:
```
marketRDD.unpersist()
```

[5]:
```
CommMktArrivals.csv MapPartitionsRDD[1] at textFile at
NativeMethodAccessorImpl.java:0
```

[6]:
```
# Copy on DISK_ONLY
marketRDD.unpersist()
marketRDD.persist(pyspark.StorageLevel.DISK_ONLY)
marketRDD.take(2)
```

[6]:
```
['District Name,Taluk Name,Market
Name,Address,Telephone,Commodity,Year,Month,Arrival,Unit',
 'Bagalakot,Badami,BADAMI,SECRATRY A.P.M.C.BADAMI
BADAMI,220042,Bajra,2012,Jan,242,Quintal   ']
```

[7]:
```
#marketRDD.getStorageLevel
```

[8]:
```
# 2 copies on MEMORY and DISK
marketRDD.unpersist()
marketRDD.persist(pyspark.StorageLevel.MEMORY_AND_DISK_2)
marketRDD.take(2)
```

```
22/02/21 14:34:12 WARN RandomBlockReplicationPolicy: Expecting 1 replicas with
only 0 peer/s.
22/02/21 14:34:12 WARN BlockManager: Block rdd_1_0 replicated to only 0 peer(s)
instead of 1 peers
```

[8]:
```
['District Name,Taluk Name,Market
Name,Address,Telephone,Commodity,Year,Month,Arrival,Unit',
 'Bagalakot,Badami,BADAMI,SECRATRY A.P.M.C.BADAMI
BADAMI,220042,Bajra,2012,Jan,242,Quintal   ']
```