Spark Python Application

Spark Python Application – Example

Apache Spark provides APIs for many popular programming languages. Python is on of them. One can write a python script for Apache Spark and run it using spark-submit command line interface.

In this tutorial, we shall learn to write a Spark Application in Python Programming Language and submit the application to run in Spark with local input and minimal (no) options. The step by step process of creating and running Spark Python Application is demonstrated using Word-Count Example.

Prepare Input

For Word-Count Example, we shall provide a text file as input. Input file contains multiple lines and each line has multiple words separated by white space.

Input File is located at : `/home/input.txt`

Spark Application – Python Program

Following is Python program that does word count in Apache Spark.

wordcount.py

```
import sys

from pyspark import SparkContext, SparkConf

if __name__ == "__main__":

  # create Spark context with Spark configuration
  conf = SparkConf().setAppName("Word Count - Python").set("spark.hadoop.yarn.resourcema
  sc = SparkContext(conf=conf)

  # read in text file and split each document into words
  words = sc.textFile("/home/arjun/input.txt").flatMap(lambda line: line.split(" "))

  # count the occurrence of each word
  wordCounts = words.map(lambda word: (word, 1)).reduceByKey(lambda a,b:a +b)

  wordCounts.saveAsTextFile("/home/arjun/output/")
```

Submit Python Application to Spark

To submit the above Spark Application to Spark for running, Open a Terminal or Command Prompt from the location of wordcount.py, and run the following command :

```
$ spark-submit wordcount.py
```

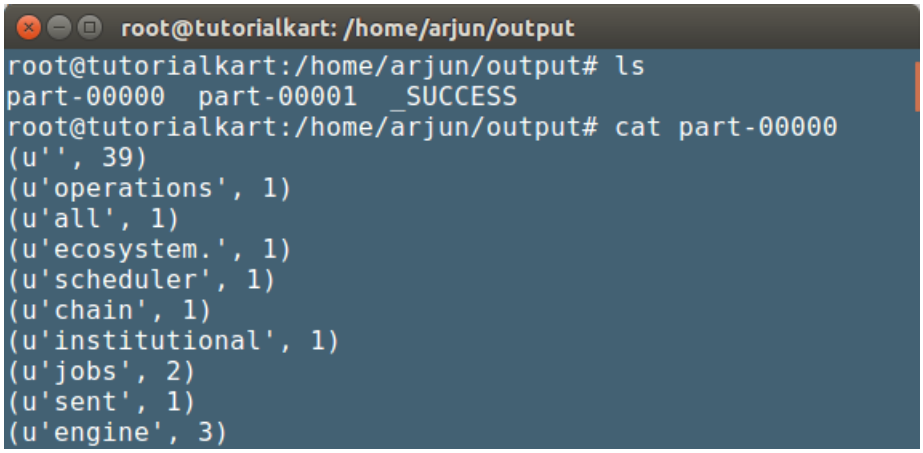
```
arjun@tutorialkart:~/workspace/spark$ spark-submit wordcount.py
17/11/14 10:54:57 INFO spark.SparkContext: Running Spark version 2.2.0
17/11/14 10:54:57 WARN util.NativeCodeLoader: Unable to load native-hadoop library for y
17/11/14 10:54:57 INFO spark.SparkContext: Submitted application: Word Count - Python
17/11/14 10:54:57 INFO spark.SecurityManager: Changing view acls to: arjun
17/11/14 10:54:57 INFO spark.SecurityManager: Changing modify acls to: arjun
17/11/14 10:54:57 INFO spark.SecurityManager: Changing view acls groups to:
17/11/14 10:54:57 INFO spark.SecurityManager: Changing modify acls groups to:
17/11/14 10:54:57 INFO spark.SecurityManager: SecurityManager: authentication disabled;
17/11/14 10:54:58 INFO util.Utils: Successfully started service 'sparkDriver' on port 38
17/11/14 10:54:58 INFO spark.SparkEnv: Registering MapOutputTracker
17/11/14 10:54:58 INFO spark.SparkEnv: Registering BlockManagerMaster
17/11/14 10:54:58 INFO storage.BlockManagerMasterEndpoint: Using org.apache.spark.storag
17/11/14 10:54:58 INFO storage.BlockManagerMasterEndpoint: BlockManagerMasterEndpoint up
17/11/14 10:54:58 INFO storage.DiskBlockManager: Created local directory at /tmp/blockmg
17/11/14 10:54:58 INFO memory.MemoryStore: MemoryStore started with capacity 366.3 MB
17/11/14 10:54:58 INFO spark.SparkEnv: Registering OutputCommitCoordinator
17/11/14 10:54:58 INFO util.log: Logging initialized @2864ms
17/11/14 10:54:58 INFO server.Server: jetty-9.3.z-SNAPSHOT
17/11/14 10:54:58 INFO server.Server: Started @2997ms
17/11/14 10:54:58 WARN util.Utils: Service 'SparkUI' could not bind on port 4040. Attemp
17/11/14 10:54:58 INFO server.AbstractConnector: Started ServerConnector@127b57de{HTTP/1
17/11/14 10:54:58 INFO util.Utils: Successfully started service 'SparkUI' on port 4041.
17/11/14 10:54:58 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@71f
17/11/14 10:54:58 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@4ee
17/11/14 10:54:58 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@1ff
17/11/14 10:54:58 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@173
17/11/14 10:54:58 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@18a
17/11/14 10:54:58 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@729
17/11/14 10:54:58 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@78a
17/11/14 10:54:58 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@4e2
17/11/14 10:54:58 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@658
17/11/14 10:54:58 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@5a7
17/11/14 10:54:58 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@7b2
17/11/14 10:54:58 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@7c7
17/11/14 10:54:58 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@58f
17/11/14 10:54:58 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@1f1
17/11/14 10:54:58 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@2cf
17/11/14 10:54:58 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@390
17/11/14 10:54:58 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@37a
17/11/14 10:54:58 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@16f
17/11/14 10:54:58 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@3ab
17/11/14 10:54:58 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@7cc
17/11/14 10:54:58 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@57f
17/11/14 10:54:58 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@5c5
17/11/14 10:54:58 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@625
17/11/14 10:54:58 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@784
17/11/14 10:54:58 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@20d
17/11/14 10:54:58 INFO ui.SparkUI: Bound SparkUI to 192.168.0.104, and started at http:/
```

```
17/11/14 10:54:58 INFO ui.SparkUI: Bound SparkUI to 192.168.0.104, and started at http://
17/11/14 10:54:59 INFO spark.SparkContext: Added file file:/home/arjun/workspace/spark/w
17/11/14 10:54:59 INFO util.Utils: Copying /home/arjun/workspace/spark/wordcount.py to /
17/11/14 10:54:59 INFO executor.Executor: Starting executor ID driver on host localhost
17/11/14 10:54:59 INFO util.Utils: Successfully started service 'org.apache.spark.networ
17/11/14 10:54:59 INFO netty.NettyBlockTransferService: Server created on 192.168.0.104:
17/11/14 10:54:59 INFO storage.BlockManager: Using org.apache.spark.storage.RandomBlockR
17/11/14 10:54:59 INFO storage.BlockManagerMaster: Registering BlockManager BlockManager
17/11/14 10:54:59 INFO storage.BlockManagerMasterEndpoint: Registering block manager 192
17/11/14 10:54:59 INFO storage.BlockManagerMaster: Registered BlockManager BlockManagerI
17/11/14 10:54:59 INFO storage.BlockManager: Initialized BlockManager: BlockManagerId(dr
17/11/14 10:54:59 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@720
17/11/14 10:55:00 INFO memory.MemoryStore: Block broadcast_0 stored as values in memory
17/11/14 10:55:00 INFO memory.MemoryStore: Block broadcast_0_piece0 stored as bytes in m
17/11/14 10:55:00 INFO storage.BlockManagerInfo: Added broadcast_0_piece0 in memory on 1
17/11/14 10:55:00 INFO spark.SparkContext: Created broadcast 0 from textFile at NativeMe
17/11/14 10:55:00 INFO mapred.FileInputFormat: Total input paths to process : 1
17/11/14 10:55:00 INFO output.FileOutputCommitter: File Output Committer Algorithm versi
17/11/14 10:55:00 INFO spark.SparkContext: Starting job: saveAsTextFile at NativeMethodA
17/11/14 10:55:00 INFO scheduler.DAGScheduler: Registering RDD 3 (reduceByKey at /home/a
17/11/14 10:55:00 INFO scheduler.DAGScheduler: Got job 0 (saveAsTextFile at NativeMethod
17/11/14 10:55:00 INFO scheduler.DAGScheduler: Final stage: ResultStage 1 (saveAsTextFil
17/11/14 10:55:00 INFO scheduler.DAGScheduler: Parents of final stage: List(ShuffleMapSt
17/11/14 10:55:00 INFO scheduler.DAGScheduler: Missing parents: List(ShuffleMapStage 0)
17/11/14 10:55:00 INFO scheduler.DAGScheduler: Submitting ShuffleMapStage 0 (PairwiseRDD
17/11/14 10:55:01 INFO memory.MemoryStore: Block broadcast_1 stored as values in memory
17/11/14 10:55:01 INFO memory.MemoryStore: Block broadcast_1_piece0 stored as bytes in m
17/11/14 10:55:01 INFO storage.BlockManagerInfo: Added broadcast_1_piece0 in memory on 1
17/11/14 10:55:01 INFO spark.SparkContext: Created broadcast 1 from broadcast at DAGSche
17/11/14 10:55:01 INFO scheduler.DAGScheduler: Submitting 2 missing tasks from ShuffleMa
17/11/14 10:55:01 INFO scheduler.TaskSchedulerImpl: Adding task set 0.0 with 2 tasks
17/11/14 10:55:01 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 0.0 (TID 0,
17/11/14 10:55:01 INFO scheduler.TaskSetManager: Starting task 1.0 in stage 0.0 (TID 1,
17/11/14 10:55:01 INFO executor.Executor: Running task 0.0 in stage 0.0 (TID 0)
17/11/14 10:55:01 INFO executor.Executor: Running task 1.0 in stage 0.0 (TID 1)
17/11/14 10:55:01 INFO executor.Executor: Fetching file:/home/arjun/workspace/spark/word
17/11/14 10:55:01 INFO util.Utils: /home/arjun/workspace/spark/wordcount.py has been pre
17/11/14 10:55:01 INFO rdd.HadoopRDD: Input split: file:/home/arjun/input.txt:0+4248
17/11/14 10:55:01 INFO rdd.HadoopRDD: Input split: file:/home/arjun/input.txt:4248+4248
17/11/14 10:55:02 INFO python.PythonRunner: Times: total = 419, boot = 347, init = 50, f
17/11/14 10:55:02 INFO python.PythonRunner: Times: total = 410, boot = 342, init = 55, f
17/11/14 10:55:02 INFO executor.Executor: Finished task 1.0 in stage 0.0 (TID 1). 1612 b
17/11/14 10:55:02 INFO executor.Executor: Finished task 0.0 in stage 0.0 (TID 0). 1612 b
17/11/14 10:55:02 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 0.0 (TID 0)
17/11/14 10:55:02 INFO scheduler.TaskSetManager: Finished task 1.0 in stage 0.0 (TID 1)
17/11/14 10:55:02 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 0.0, whose tasks hav
17/11/14 10:55:02 INFO scheduler.DAGScheduler: ShuffleMapStage 0 (reduceByKey at /home/a
17/11/14 10:55:02 INFO scheduler.DAGScheduler: looking for newly runnable stages
17/11/14 10:55:02 INFO scheduler.DAGScheduler: running: Set()
17/11/14 10:55:02 INFO scheduler.DAGScheduler: waiting: Set(ResultStage 1)
17/11/14 10:55:02 INFO scheduler.DAGScheduler: failed: Set()
17/11/14 10:55:02 INFO scheduler.DAGScheduler: Submitting ResultStage 1 (MapPartitionsRD
17/11/14 10:55:02 INFO memory.MemoryStore: Block broadcast_2 stored as values in memory
17/11/14 10:55:02 INFO memory.MemoryStore: Block broadcast_2_piece0 stored as bytes in m
17/11/14 10:55:02 INFO storage.BlockManagerInfo: Added broadcast_2_piece0 in memory on 1
17/11/14 10:55:02 INFO spark.SparkContext: Created broadcast 2 from broadcast at DAGSche
17/11/14 10:55:02 INFO scheduler.DAGScheduler: Submitting 2 missing tasks from ResultSta
17/11/14 10:55:02 INFO scheduler.TaskSchedulerImpl: Adding task set 1.0 with 2 tasks
17/11/14 10:55:02 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 1.0 (TID 2,
17/11/14 10:55:02 INFO scheduler.TaskSetManager: Starting task 1.0 in stage 1.0 (TID 3,
17/11/14 10:55:02 INFO executor.Executor: Running task 0.0 in stage 1.0 (TID 2)
17/11/14 10:55:02 INFO executor.Executor: Running task 1.0 in stage 1.0 (TID 3)
17/11/14 10:55:02 INFO storage.ShuffleBlockFetcherIterator: Getting 2 non-empty blocks o
```

```
17/11/14 10:55:02 INFO storage.ShuffleBlockFetcherIterator: Getting 2 non-empty blocks o
17/11/14 10:55:02 INFO storage.ShuffleBlockFetcherIterator: Started 0 remote fetches in
17/11/14 10:55:02 INFO storage.ShuffleBlockFetcherIterator: Started 0 remote fetches in
17/11/14 10:55:02 INFO output.FileOutputCommitter: File Output Committer Algorithm versi
17/11/14 10:55:02 INFO output.FileOutputCommitter: File Output Committer Algorithm versi
17/11/14 10:55:02 INFO python.PythonRunner: Times: total = 49, boot = -558, init = 600,
17/11/14 10:55:02 INFO python.PythonRunner: Times: total = 61, boot = -560, init = 613,
17/11/14 10:55:02 INFO output.FileOutputCommitter: Saved output of task 'attempt_2017111
17/11/14 10:55:02 INFO output.FileOutputCommitter: Saved output of task 'attempt_2017111
17/11/14 10:55:02 INFO mapred.SparkHadoopMapRedUtil: attempt_20171114105500_0001_m_00000
17/11/14 10:55:02 INFO mapred.SparkHadoopMapRedUtil: attempt_20171114105500_0001_m_00000
17/11/14 10:55:02 INFO executor.Executor: Finished task 0.0 in stage 1.0 (TID 2). 1638 b
17/11/14 10:55:02 INFO executor.Executor: Finished task 1.0 in stage 1.0 (TID 3). 1638 b
17/11/14 10:55:02 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 1.0 (TID 2)
17/11/14 10:55:02 INFO scheduler.TaskSetManager: Finished task 1.0 in stage 1.0 (TID 3)
17/11/14 10:55:02 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 1.0, whose tasks hav
17/11/14 10:55:02 INFO scheduler.DAGScheduler: ResultStage 1 (saveAsTextFile at NativeMe
17/11/14 10:55:02 INFO scheduler.DAGScheduler: Job 0 finished: saveAsTextFile at NativeM
17/11/14 10:55:02 INFO spark.SparkContext: Invoking stop() from shutdown hook
17/11/14 10:55:02 INFO server.AbstractConnector: Stopped Spark@127b57de{HTTP/1.1,[http/1
17/11/14 10:55:02 INFO ui.SparkUI: Stopped Spark web UI at http://192.168.0.104:4041
17/11/14 10:55:02 INFO spark.MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpo
17/11/14 10:55:02 INFO memory.MemoryStore: MemoryStore cleared
17/11/14 10:55:02 INFO storage.BlockManager: BlockManager stopped
17/11/14 10:55:02 INFO storage.BlockManagerMaster: BlockManagerMaster stopped
17/11/14 10:55:02 INFO scheduler.OutputCommitCoordinator$OutputCommitCoordinatorEndpoint
17/11/14 10:55:02 INFO spark.SparkContext: Successfully stopped SparkContext
17/11/14 10:55:02 INFO util.ShutdownHookManager: Shutdown hook called
17/11/14 10:55:02 INFO util.ShutdownHookManager: Deleting directory /tmp/spark-39c98eb0-
17/11/14 10:55:02 INFO util.ShutdownHookManager: Deleting directory /tmp/spark-39c98eb0-
```

Output

The word counts are written to the output folder. Verify the counts for the correctness of the program. (We have provided the output path in wordcount.py Python script).



Output has been written to two part files. Files contain tuples of word and the corresponding number of occurrences in the input file.

Conclusion

In this Apache Spark Tutorial, **Python Application for Spark**, we have learnt to run a simple Spark Application written in Python Programming language.

- Spark RDD - Distinct
- Spark RDD - Reduce

Spark Dataseet

- Spark - Read JSON file to Dataset
- Spark - Write Dataset to JSON file
- Spark - Add new Column to Dataset
- Spark - Concatenate Datasets

Spark MLlib (Machine Learning Library)

- Spark MLlib Tutorial
- KMeans Clustering & Classification
- Decision Tree Classification
- Random Forest Classification
- Naive Bayes Classification
- Logistic Regression Classification
- Topic Modelling

Spark SQL

- Spark SQL Tutorial
- Spark SQL - Load JSON file and execute SQL Query

Spark Others

- Spark Interview Questions