Big Data And Hadoop

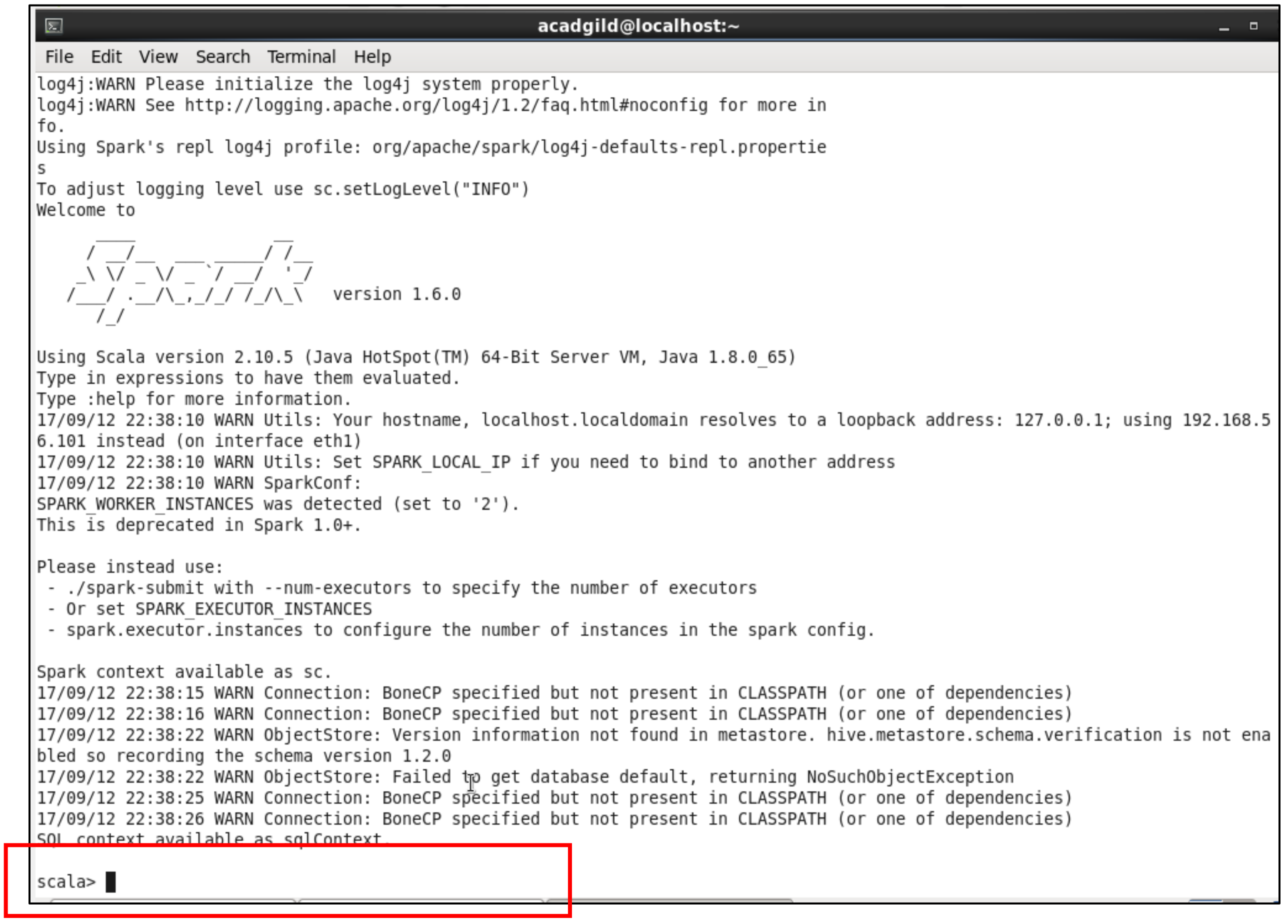
Session 13 – Assignment 3

**Problem Statement:**

Write the code to Turn a collection into a RDD and perform map operation on it to cube every number and filter the number which are divided by two and three.

**Solution:**

We first start spark context using the command **spark-shell.** The scala shell appears as follows:



```
log4j:WARN Please initialize the log4j system properly.
log4j:WARN See http://logging.apache.org/log4j/1.2/faq.html#noconfig for more in
fo.
Using Spark's repl log4j profile: org/apache/spark/log4j-defaults-repl.propertie
s
To adjust logging level use sc.setLogLevel("INFO")
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 1.6.0
      /_/

Using Scala version 2.10.5 (Java HotSpot(TM) 64-Bit Server VM, Java 1.8.0_65)
Type in expressions to have them evaluated.
Type :help for more information.
17/09/12 22:38:10 WARN Utils: Your hostname, localhost.localdomain resolves to a loopback address: 127.0.0.1; using 192.168.5
6.101 instead (on interface eth1)
17/09/12 22:38:10 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
17/09/12 22:38:10 WARN SparkConf:
SPARK_WORKER_INSTANCES was detected (set to '2').
This is deprecated in Spark 1.0+.

Please instead use:
 - ./spark-submit with --num-executors to specify the number of executors
 - Or set SPARK_EXECUTOR_INSTANCES
 - spark.executor.instances to configure the number of instances in the spark config.

Spark context available as sc.
17/09/12 22:38:15 WARN Connection: BoneCP specified but not present in CLASSPATH (or one of dependencies)
17/09/12 22:38:16 WARN Connection: BoneCP specified but not present in CLASSPATH (or one of dependencies)
17/09/12 22:38:22 WARN ObjectStore: Version information not found in metastore. hive.metastore.schema.verification is not ena
bled so recording the schema version 1.2.0
17/09/12 22:38:22 WARN ObjectStore: Failed to get database default, returning NoSuchObjectException
17/09/12 22:38:25 WARN Connection: BoneCP specified but not present in CLASSPATH (or one of dependencies)
17/09/12 22:38:26 WARN Connection: BoneCP specified but not present in CLASSPATH (or one of dependencies)
SQL context available as sqlContext.

scala> 
```

Step 1:

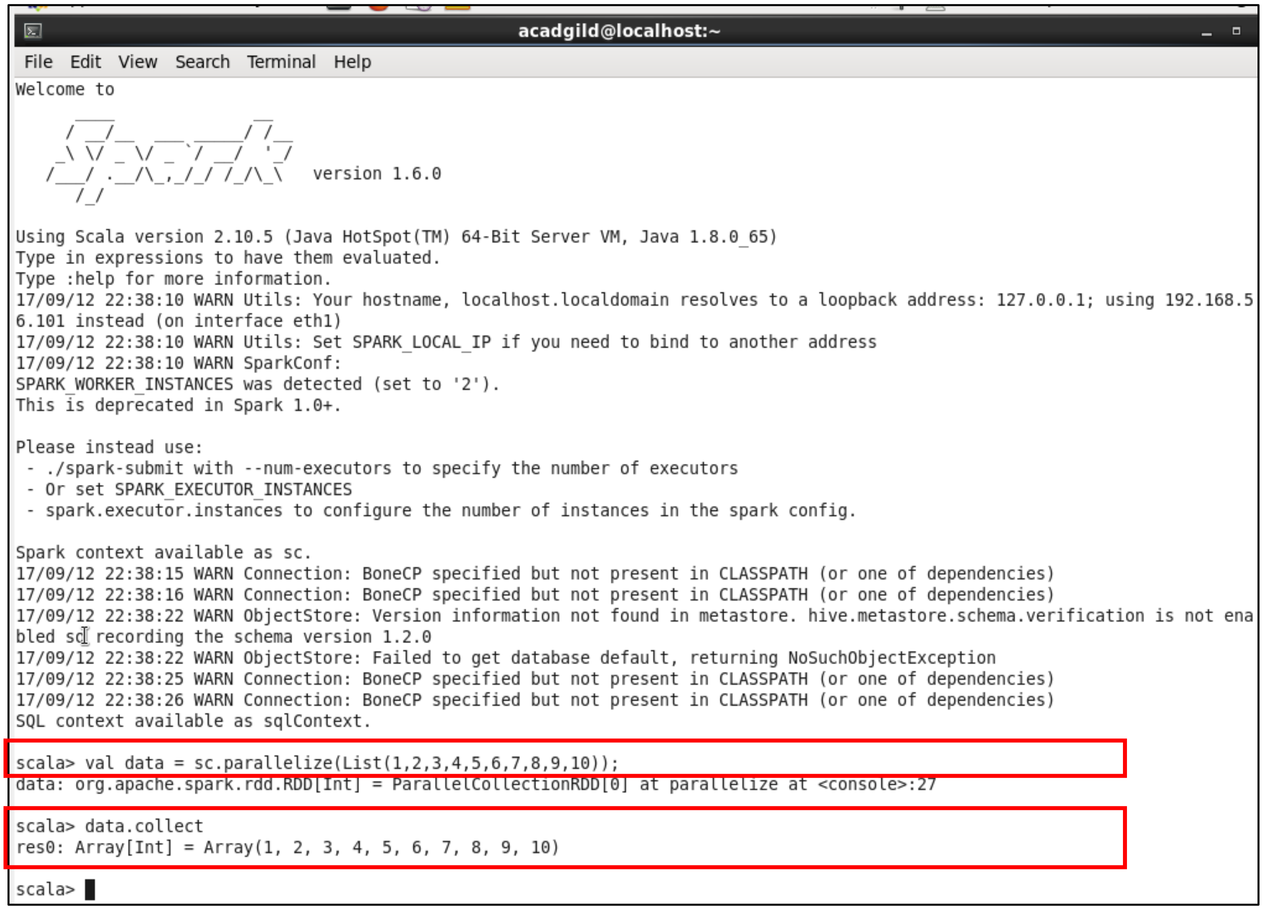
Let us consider the collection to be a List with the following elements: (1, 2, 3, 4, 5, 6, 7, 8, 9, 10).

We create an RDD representing this data using parallelize(). Parallelized collections are created by calling SparkContext's parallelize method on an existing collection in your driver program (a Scala Seq). The elements of the collection are copied to form a distributed dataset that can be operated on in parallel.

In the our example, we use parallelize as follows:

**val data = sc.parallelize( List(1, 2, 3, 4, 5, 6, 7, 8, 9, 10) ).**

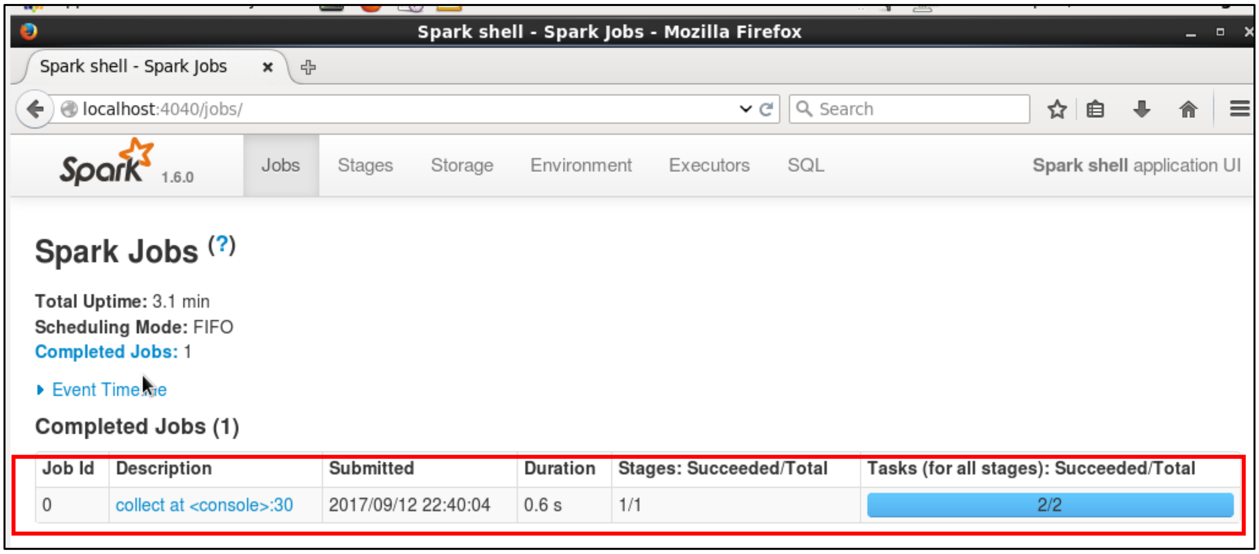
This is a transformation performed .



```
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 1.6.0
      /_/

Using Scala version 2.10.5 (Java HotSpot(TM) 64-Bit Server VM, Java 1.8.0_65)
Type in expressions to have them evaluated.
Type :help for more information.
17/09/12 22:38:10 WARN Utils: Your hostname, localhost.localdomain resolves to a loopback address: 127.0.0.1; using 192.168.5
6.101 instead (on interface eth1)
17/09/12 22:38:10 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
17/09/12 22:38:10 WARN SparkConf:
SPARK_WORKER_INSTANCES was detected (set to '2').
This is deprecated in Spark 1.0+.

Please instead use:
 - ./spark-submit with --num-executors to specify the number of executors
 - Or set SPARK_EXECUTOR_INSTANCES
 - spark.executor.instances to configure the number of instances in the spark config.

Spark context available as sc.
17/09/12 22:38:15 WARN Connection: BoneCP specified but not present in CLASSPATH (or one of dependencies)
17/09/12 22:38:16 WARN Connection: BoneCP specified but not present in CLASSPATH (or one of dependencies)
17/09/12 22:38:22 WARN ObjectStore: Version information not found in metastore. hive.metastore.schema.verification is not ena
bled so recording the schema version 1.2.0
17/09/12 22:38:22 WARN ObjectStore: Failed to get database default, returning NoSuchObjectException
17/09/12 22:38:25 WARN Connection: BoneCP specified but not present in CLASSPATH (or one of dependencies)
17/09/12 22:38:26 WARN Connection: BoneCP specified but not present in CLASSPATH (or one of dependencies)
SQL context available as sqlContext.

scala> val data = sc.parallelize(List(1,2,3,4,5,6,7,8,9,10));
data: org.apache.spark.rdd.RDD[Int] = ParallelCollectionRDD[0] at parallelize at <console>:27

scala> data.collect
res0: Array[Int] = Array(1, 2, 3, 4, 5, 6, 7, 8, 9, 10)

scala>
```

On performing **collect** action, we get the output as shown in the above snapshot. Here, the RDD got created with the specified data.

**Collect** being an action, its execution appears in the Spark UI as a job as follows:



## Spark Jobs (?)

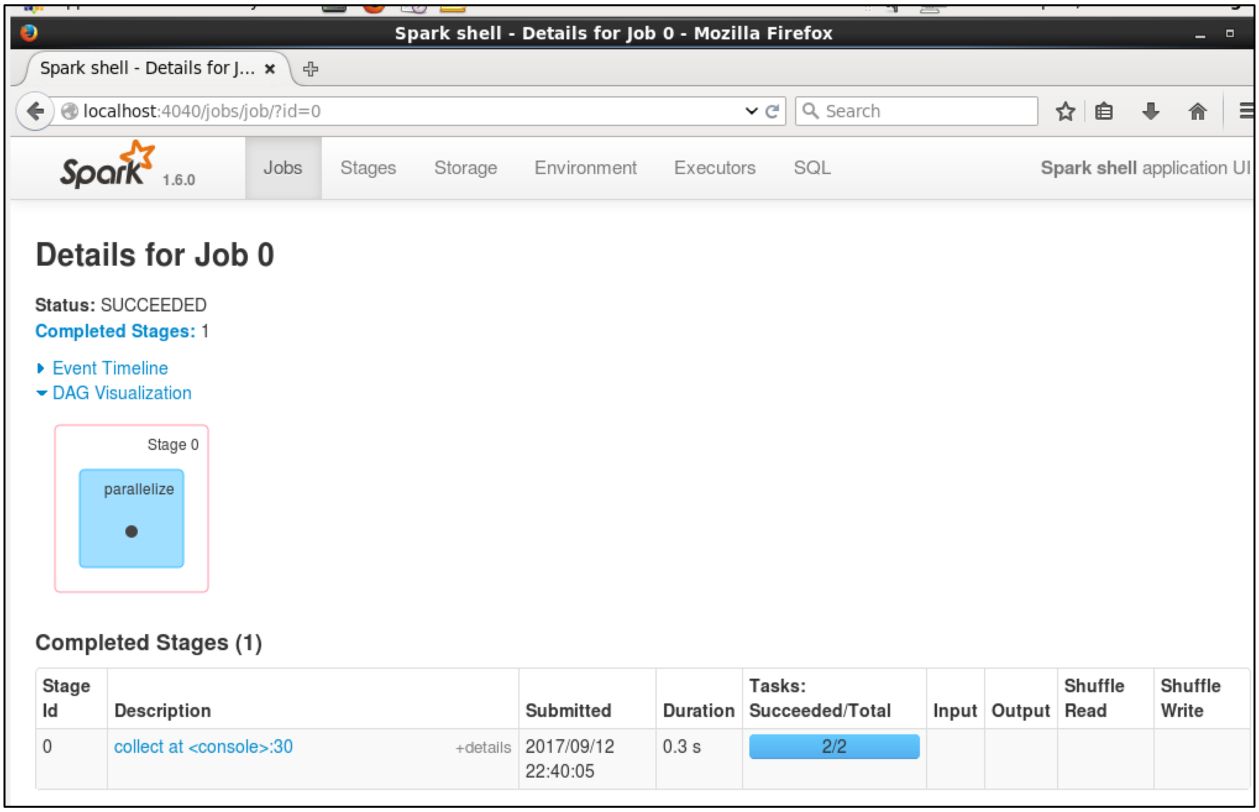
**Total Uptime:** 3.1 min
**Scheduling Mode:** FIFO
**Completed Jobs:** 1

▸ Event Timeline

### Completed Jobs (1)

| Job Id | Description | Submitted | Duration | Stages: Succeeded/Total | Tasks (for all stages): Succeeded/Total |
|---|---|---|---|---|---|
| 0 | collect at <console>:30 | 2017/09/12 22:40:04 | 0.6 s | 1/1 | 2/2 |

This action was split into two tasks as mentioned in the above snapshot. The DAG representation of this job is as follows:


Spark shell - Details for Job 0 - Mozilla Firefox
Spark shell - Details for J...
localhost:4040/jobs/job/?id=0
Search
Spark 1.6.0
Jobs
Stages
Storage
Environment
Executors
SQL
Spark shell application UI
Stage 0
parallelize

## Details for Job 0

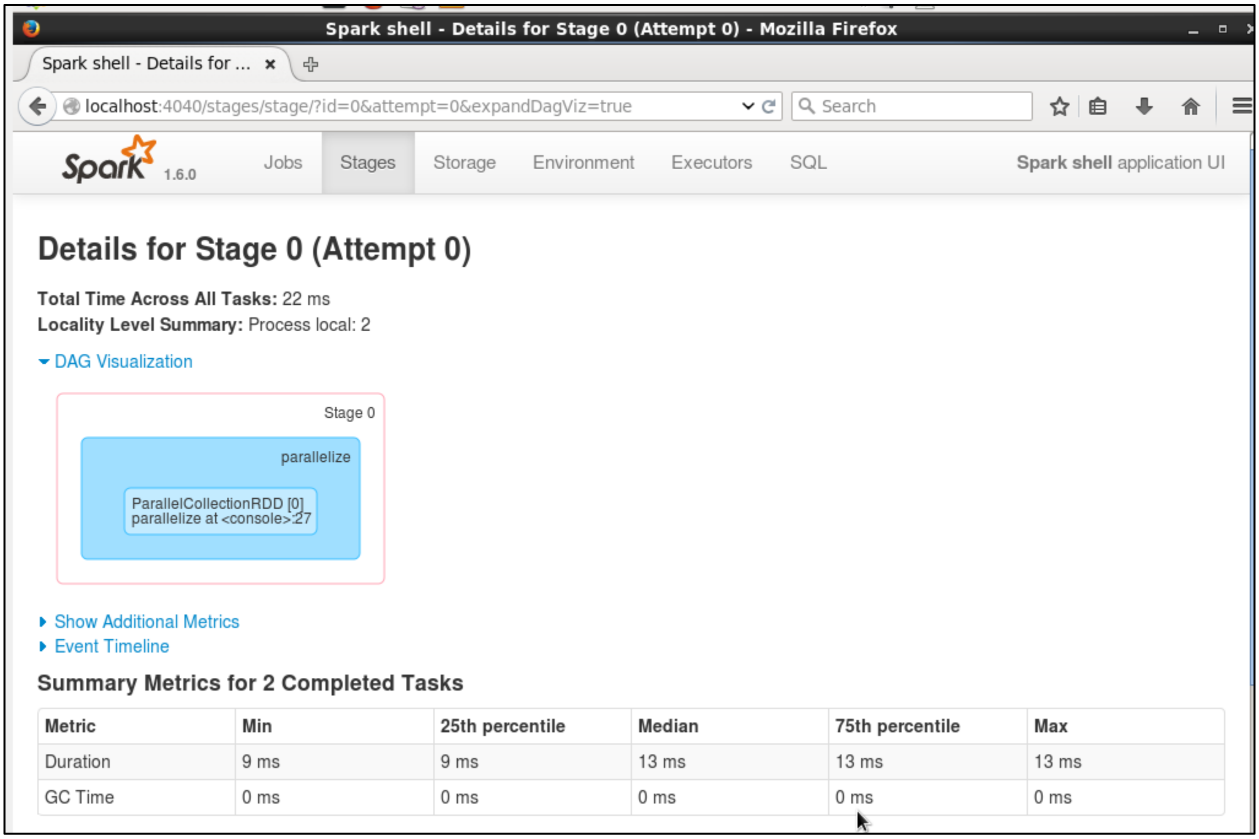
**Status:** SUCCEEDED
**Completed Stages:** 1

▸ Event Timeline
▾ DAG Visualization

### Completed Stages (1)

| Stage Id | Description | Submitted | Duration | Tasks: Succeeded/Total | Input | Output | Shuffle Read | Shuffle Write |
|---|---|---|---|---|---|---|---|---|
| 0 | collect at <console>:30 +details | 2017/09/12 22:40:05 | 0.3 s | 2/2 | | | | |

This action had only one node in the DAG. Detailed DAG is as follows:


Spark shell - Details for Stage 0 (Attempt 0) - Mozilla Firefox
Spark shell - Details for ...
localhost:4040/stages/stage/?id=0&attempt=0&expandDagViz=true
Search
Spark 1.6.0
Jobs
Stages
Storage
Environment
Executors
SQL
Spark shell application UI

Details for Stage 0 (Attempt 0)

**Total Time Across All Tasks:** 22 ms
**Locality Level Summary:** Process local: 2

▾ DAG Visualization

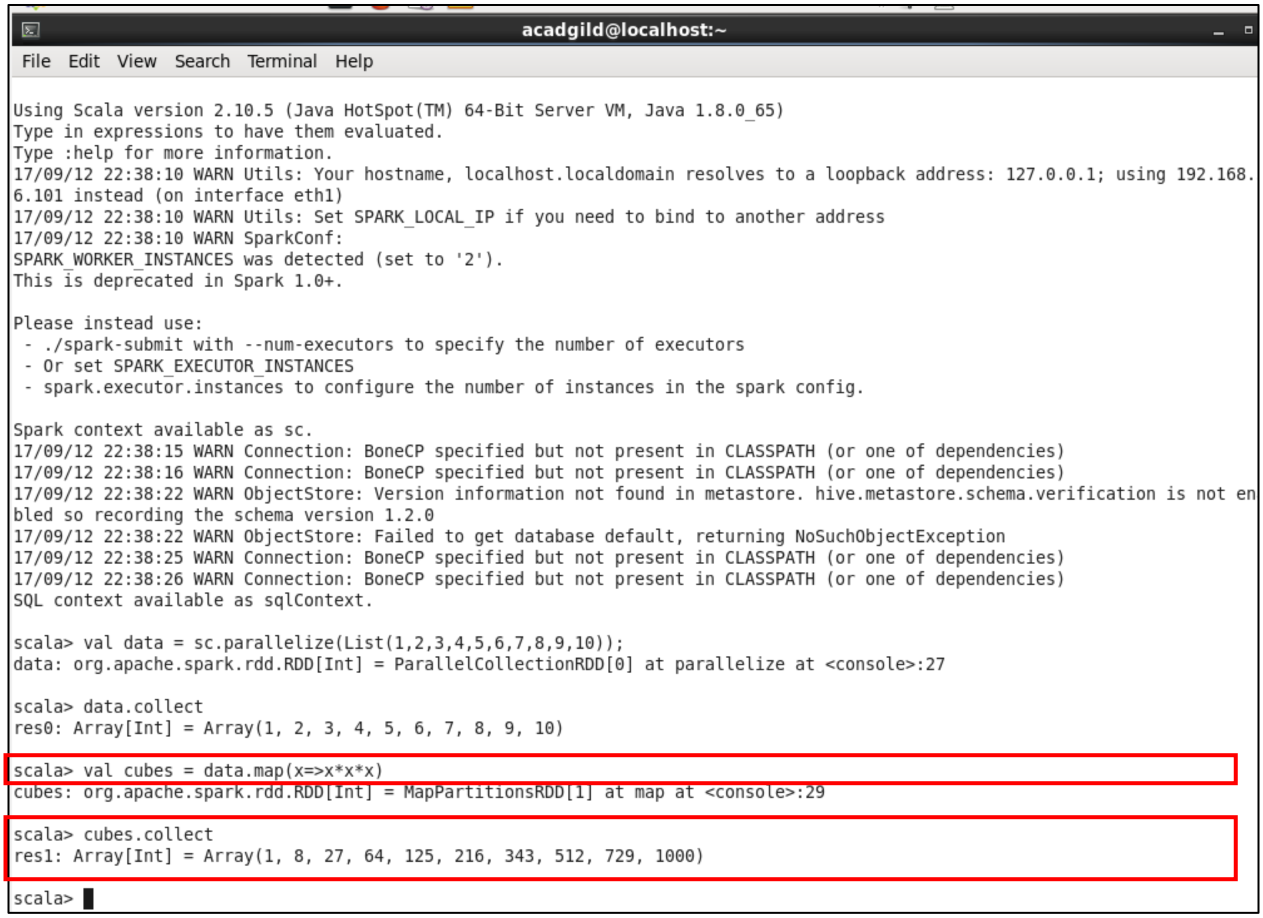
Stage 0
parallelize
ParallelCollectionRDD [0]
parallelize at <console>:27

▸ Show Additional Metrics
▸ Event Timeline

Summary Metrics for 2 Completed Tasks

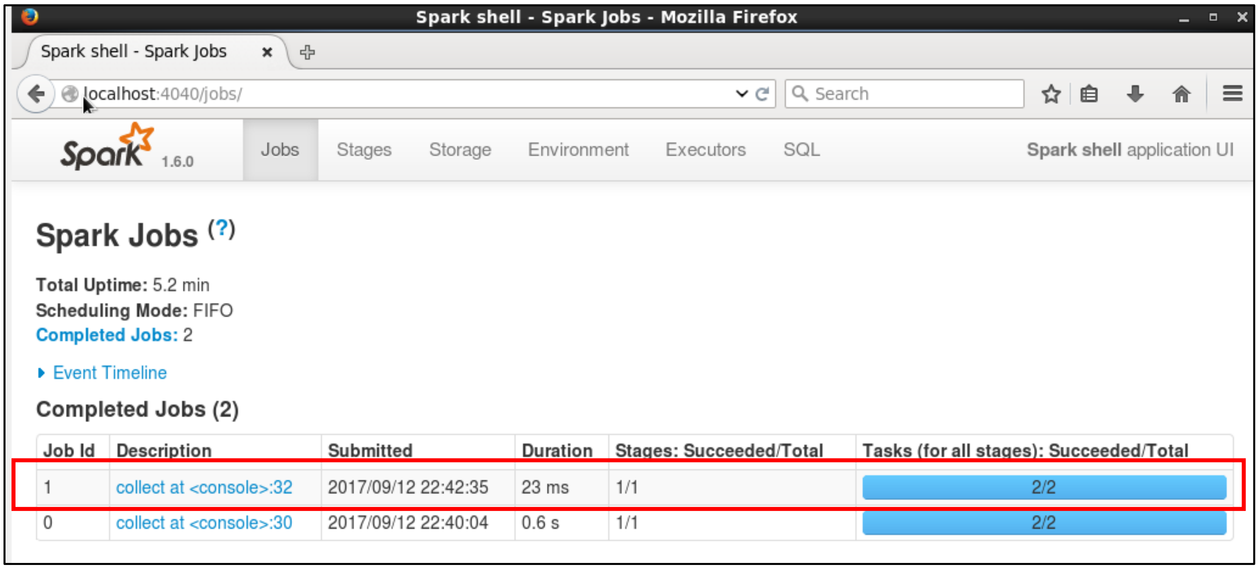
| Metric | Min | 25th percentile | Median | 75th percentile | Max |
|---|---|---|---|---|---|
| Duration | 9 ms | 9 ms | 13 ms | 13 ms | 13 ms |
| GC Time | 0 ms | 0 ms | 0 ms | 0 ms | 0 ms |

Step2:
Now, we use **map** transformation, wherein we will generate a cube of every element of the RDD as follows:
**val cubes =data.map( x=> x*x*x )**
This can be seen in the below snapshot as follows:



```
Using Scala version 2.10.5 (Java HotSpot(TM) 64-Bit Server VM, Java 1.8.0_65)
Type in expressions to have them evaluated.
Type :help for more information.
17/09/12 22:38:10 WARN Utils: Your hostname, localhost.localdomain resolves to a loopback address: 127.0.0.1; using 192.168.
6.101 instead (on interface eth1)
17/09/12 22:38:10 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
17/09/12 22:38:10 WARN SparkConf:
SPARK_WORKER_INSTANCES was detected (set to '2').
This is deprecated in Spark 1.0+.

Please instead use:
 - ./spark-submit with --num-executors to specify the number of executors
 - Or set SPARK_EXECUTOR_INSTANCES
 - spark.executor.instances to configure the number of instances in the spark config.

Spark context available as sc.
17/09/12 22:38:15 WARN Connection: BoneCP specified but not present in CLASSPATH (or one of dependencies)
17/09/12 22:38:16 WARN Connection: BoneCP specified but not present in CLASSPATH (or one of dependencies)
17/09/12 22:38:22 WARN ObjectStore: Version information not found in metastore. hive.metastore.schema.verification is not en
bled so recording the schema version 1.2.0
17/09/12 22:38:22 WARN ObjectStore: Failed to get database default, returning NoSuchObjectException
17/09/12 22:38:25 WARN Connection: BoneCP specified but not present in CLASSPATH (or one of dependencies)
17/09/12 22:38:26 WARN Connection: BoneCP specified but not present in CLASSPATH (or one of dependencies)
SQL context available as sqlContext.

scala> val data = sc.parallelize(List(1,2,3,4,5,6,7,8,9,10));
data: org.apache.spark.rdd.RDD[Int] = ParallelCollectionRDD[0] at parallelize at <console>:27

scala> data.collect
res0: Array[Int] = Array(1, 2, 3, 4, 5, 6, 7, 8, 9, 10)

scala> val cubes = data.map(x=>x*x*x)
cubes: org.apache.spark.rdd.RDD[Int] = MapPartitionsRDD[1] at map at <console>:29

scala> cubes.collect
res1: Array[Int] = Array(1, 8, 27, 64, 125, 216, 343, 512, 729, 1000)

scala> 
```

The above snapshot also shows the output on performing the **collect** action.
The execution of this action appeared in UI as follows.



## Spark Jobs (?)

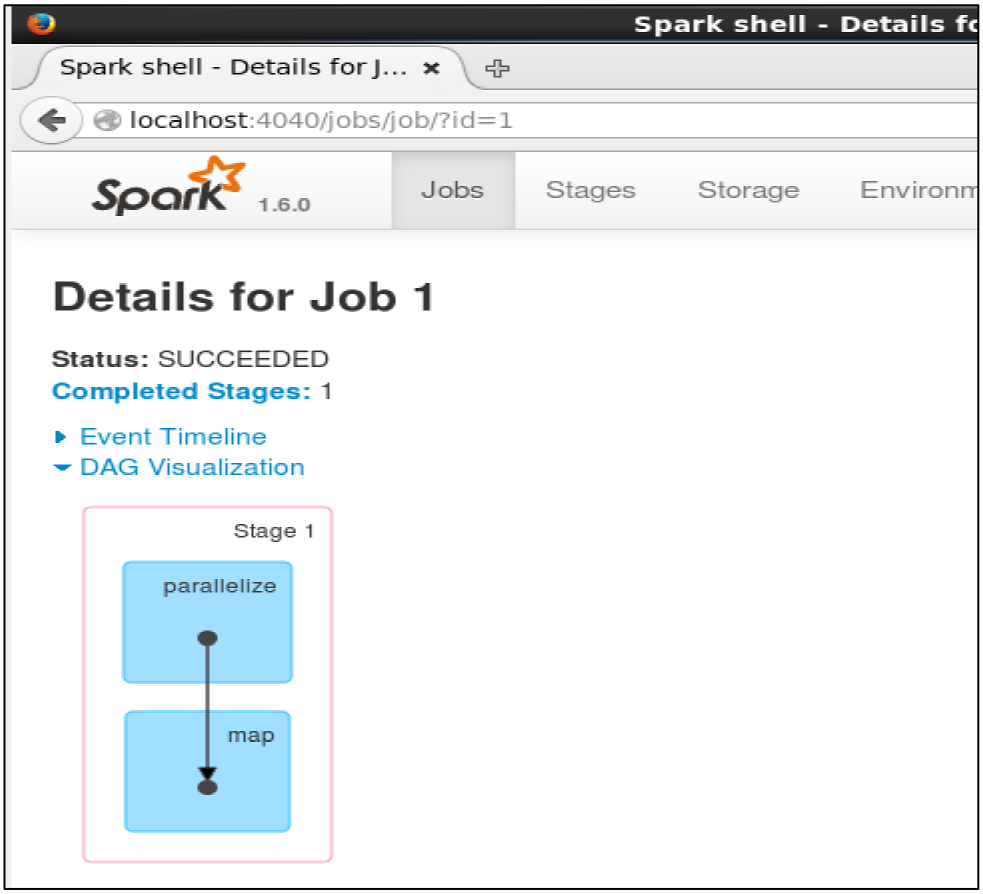
**Total Uptime:** 5.2 min
**Scheduling Mode:** FIFO
**Completed Jobs:** 2

▸ Event Timeline

### Completed Jobs (2)

| Job Id | Description | Submitted | Duration | Stages: Succeeded/Total | Tasks (for all stages): Succeeded/Total |
|---|---|---|---|---|---|
| 1 | collect at <console>:32 | 2017/09/12 22:42:35 | 23 ms | 1/1 | 2/2 |
| 0 | collect at <console>:30 | 2017/09/12 22:40:04 | 0.6 s | 1/1 | 2/2 |

Its DAG representation is as follows:



Spark shell - Details for Job 1

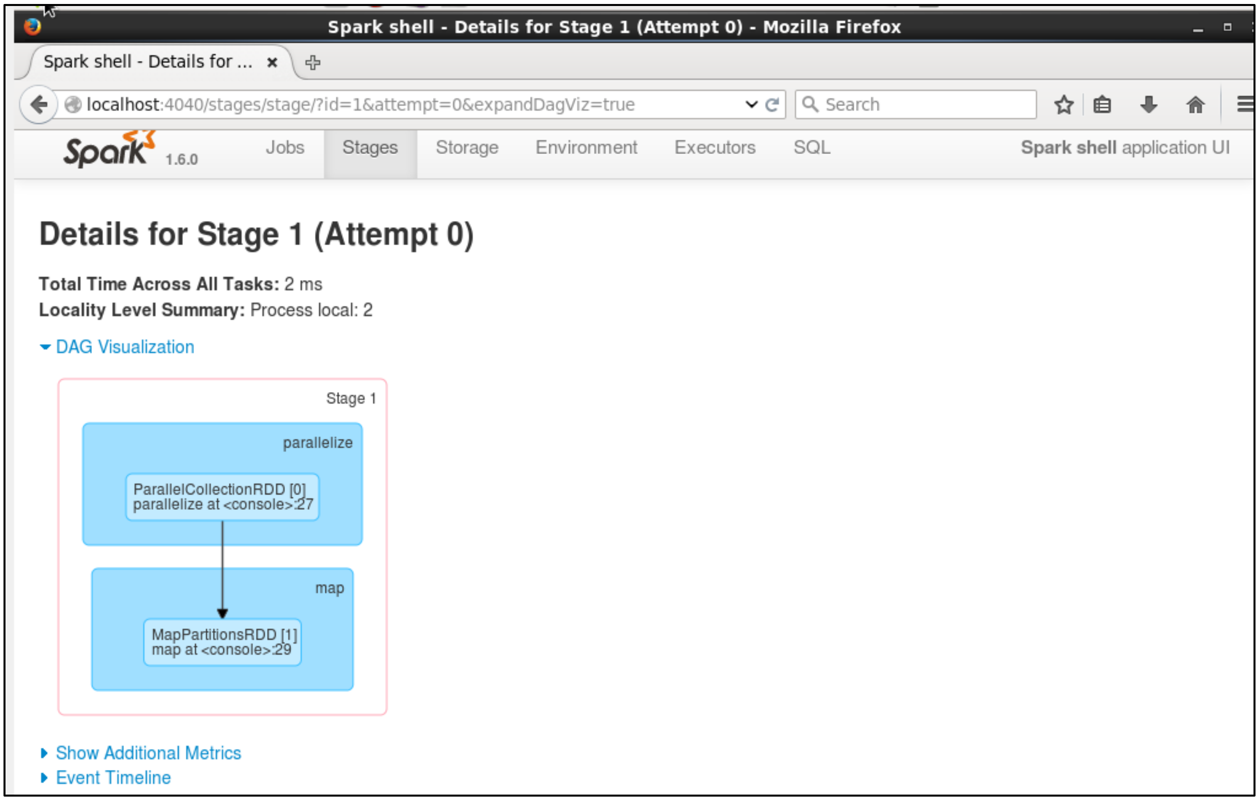
Details for Job 1

**Status:** SUCCEEDED

**Completed Stages:** 1

- Event Timeline
- DAG Visualization

Detailed DAG representation is as follows:


Spark shell - Details for Stage 1 (Attempt 0) - Mozilla Firefox
Spark shell - Details for ...
localhost:4040/stages/stage/?id=1&attempt=0&expandDagViz=true
Search
Spark 1.6.0
Jobs
Stages
Storage
Environment
Executors
SQL
Spark shell application UI

## Details for Stage 1 (Attempt 0)

**Total Time Across All Tasks:** 2 ms
**Locality Level Summary:** Process local: 2

▾ DAG Visualization

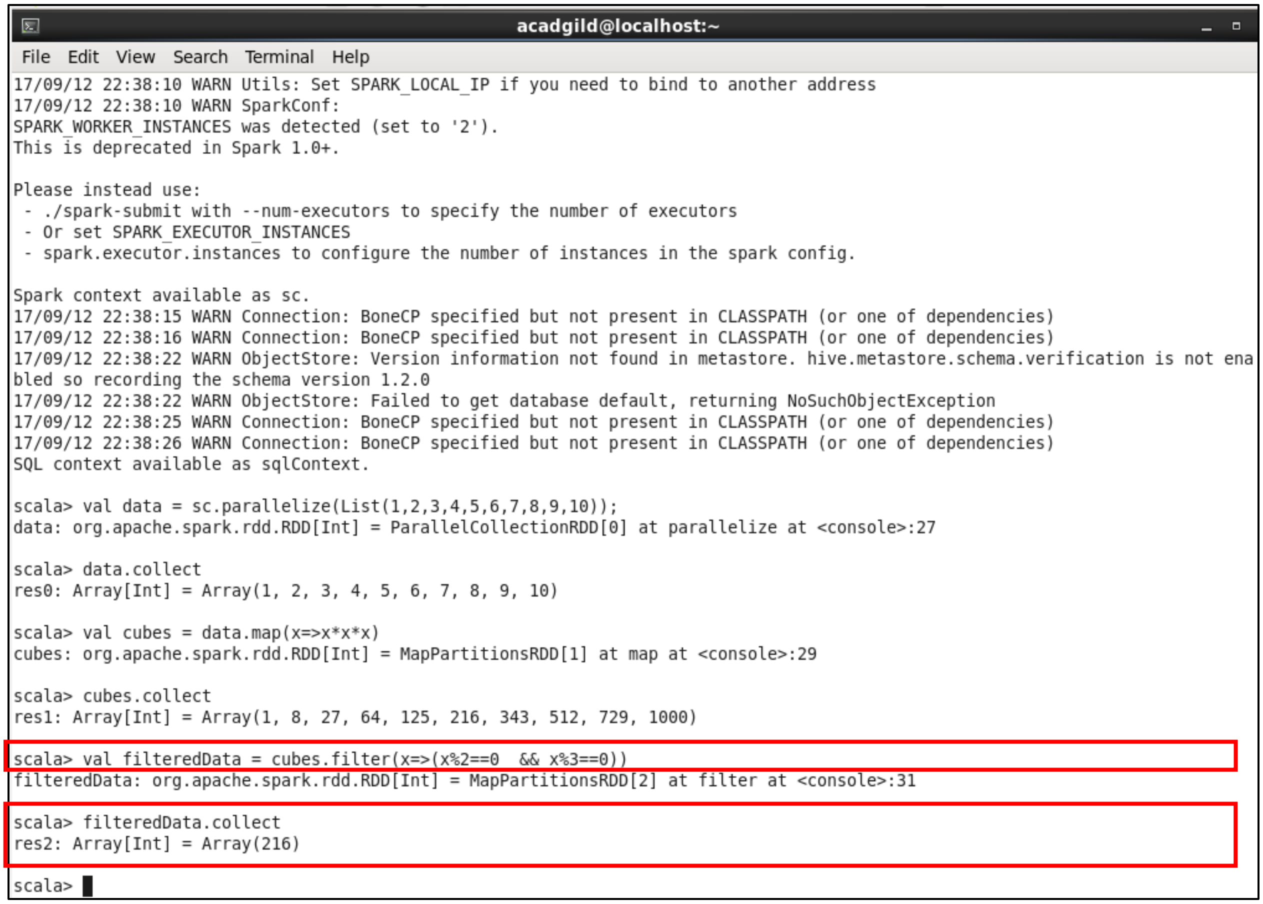
Stage 1
parallelize
ParallelCollectionRDD [0]
parallelize at <console>:27
map
MapPartitionsRDD [1]
map at <console>:29

▸ Show Additional Metrics
▸ Event Timeline

Step3:
Next is to filter the RDD. In this, we need only those elements of RDD that are divisible by 2 and 3. Hence, made use of **modulo** as follows:

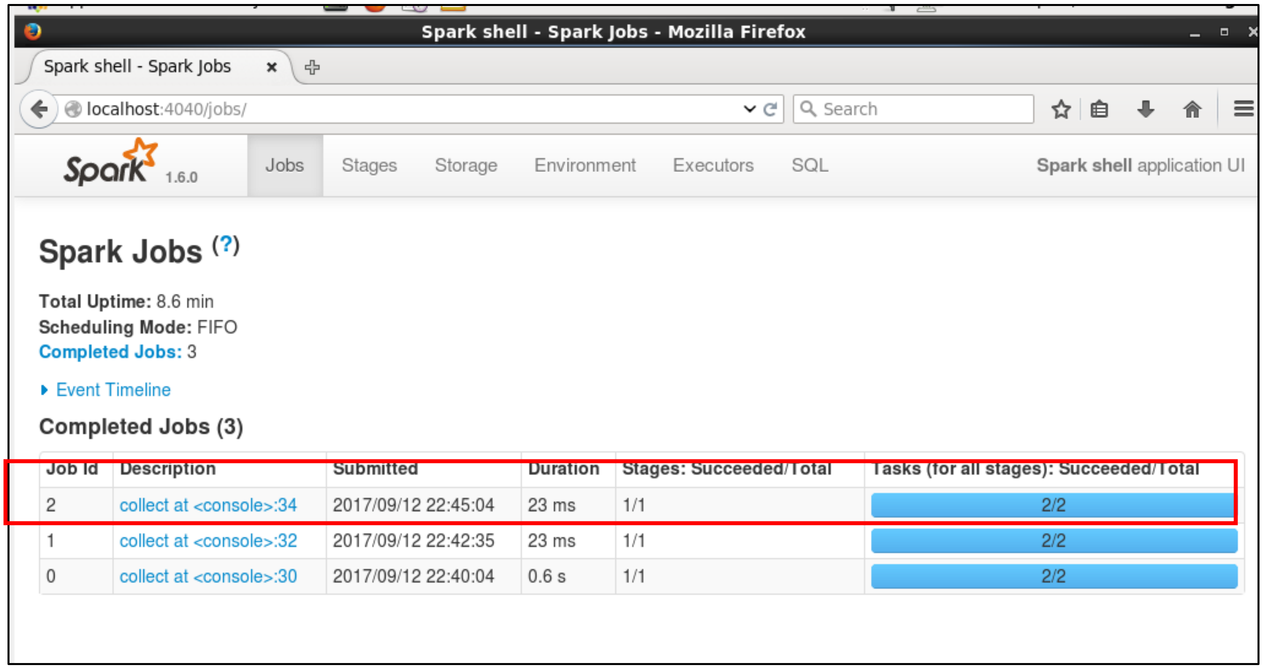
**val filteredData = cubes.filter( x=> ( x%2==0 && x%3==0 ) )**

Its execution is shown below as follows:



```
17/09/12 22:38:10 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
17/09/12 22:38:10 WARN SparkConf:
SPARK_WORKER_INSTANCES was detected (set to '2').
This is deprecated in Spark 1.0+.

Please instead use:
 - ./spark-submit with --num-executors to specify the number of executors
 - Or set SPARK_EXECUTOR_INSTANCES
 - spark.executor.instances to configure the number of instances in the spark config.

Spark context available as sc.
17/09/12 22:38:15 WARN Connection: BoneCP specified but not present in CLASSPATH (or one of dependencies)
17/09/12 22:38:16 WARN Connection: BoneCP specified but not present in CLASSPATH (or one of dependencies)
17/09/12 22:38:22 WARN ObjectStore: Version information not found in metastore. hive.metastore.schema.verification is not ena
bled so recording the schema version 1.2.0
17/09/12 22:38:22 WARN ObjectStore: Failed to get database default, returning NoSuchObjectException
17/09/12 22:38:25 WARN Connection: BoneCP specified but not present in CLASSPATH (or one of dependencies)
17/09/12 22:38:26 WARN Connection: BoneCP specified but not present in CLASSPATH (or one of dependencies)
SQL context available as sqlContext.

scala> val data = sc.parallelize(List(1,2,3,4,5,6,7,8,9,10));
data: org.apache.spark.rdd.RDD[Int] = ParallelCollectionRDD[0] at parallelize at <console>:27

scala> data.collect
res0: Array[Int] = Array(1, 2, 3, 4, 5, 6, 7, 8, 9, 10)

scala> val cubes = data.map(x=>x*x*x)
cubes: org.apache.spark.rdd.RDD[Int] = MapPartitionsRDD[1] at map at <console>:29

scala> cubes.collect
res1: Array[Int] = Array(1, 8, 27, 64, 125, 216, 343, 512, 729, 1000)

scala> val filteredData = cubes.filter(x=>(x%2==0  && x%3==0))
filteredData: org.apache.spark.rdd.RDD[Int] = MapPartitionsRDD[2] at filter at <console>:31

scala> filteredData.collect
res2: Array[Int] = Array(216)

scala> 
```

Performing the action **collect** on filtered RDD gives the results as shown above.
The execution of this action appeared in UI as follows.



Spark Jobs (?)

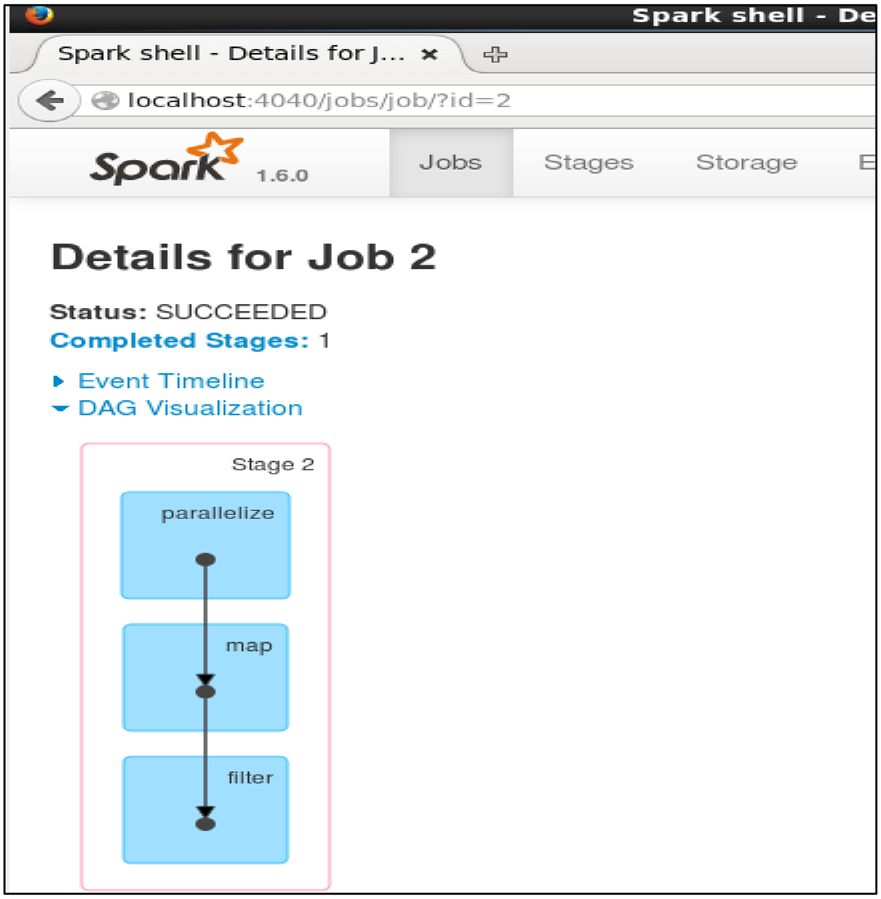
**Total Uptime:** 8.6 min
**Scheduling Mode:** FIFO
**Completed Jobs:** 3

Event Timeline

Completed Jobs (3)

| Job Id | Description | Submitted | Duration | Stages: Succeeded/Total | Tasks (for all stages): Succeeded/Total |
|---|---|---|---|---|---|
| 2 | collect at <console>:34 | 2017/09/12 22:45:04 | 23 ms | 1/1 | 2/2 |
| 1 | collect at <console>:32 | 2017/09/12 22:42:35 | 23 ms | 1/1 | 2/2 |
| 0 | collect at <console>:30 | 2017/09/12 22:40:04 | 0.6 s | 1/1 | 2/2 |

The DAG representation is as follows:

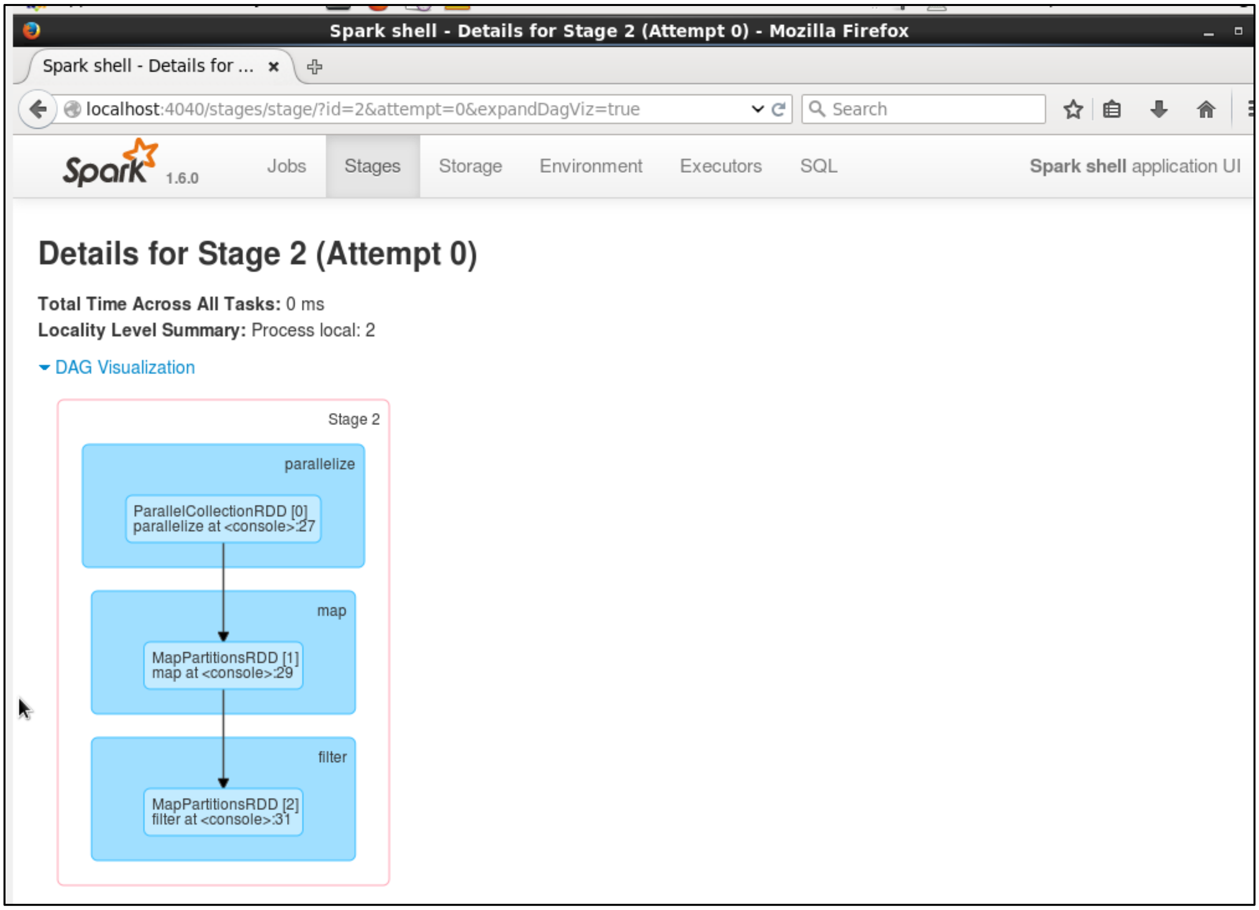


Details for Job 2

**Status:** SUCCEEDED
**Completed Stages:** 1

Event Timeline
DAG Visualization

Detailed representation of DAG is as follows:


Spark shell - Details for Stage 2 (Attempt 0) - Mozilla Firefox
Spark shell - Details for ...
localhost:4040/stages/stage/?id=2&attempt=0&expandDagViz=true
Search
Spark 1.6.0
Jobs
Stages
Storage
Environment
Executors
SQL
Spark shell application UI

Details for Stage 2 (Attempt 0)

**Total Time Across All Tasks:** 0 ms
**Locality Level Summary:** Process local: 2

DAG Visualization

Stage 2
parallelize
ParallelCollectionRDD [0]
parallelize at <console>:27
map
MapPartitionsRDD [1]
map at <console>:29
filter
MapPartitionsRDD [2]
filter at <console>:31

This is how we have performed required steps to solve the above mentioned problem statement.