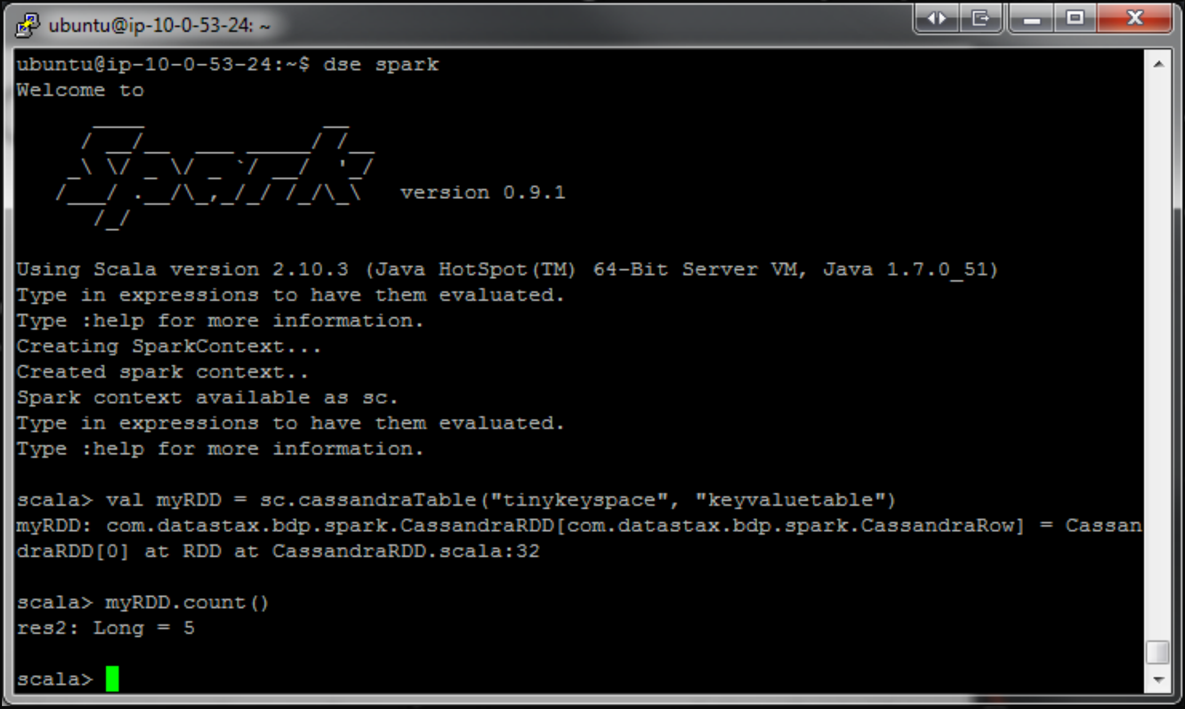
RDD FUNDAMENTALS

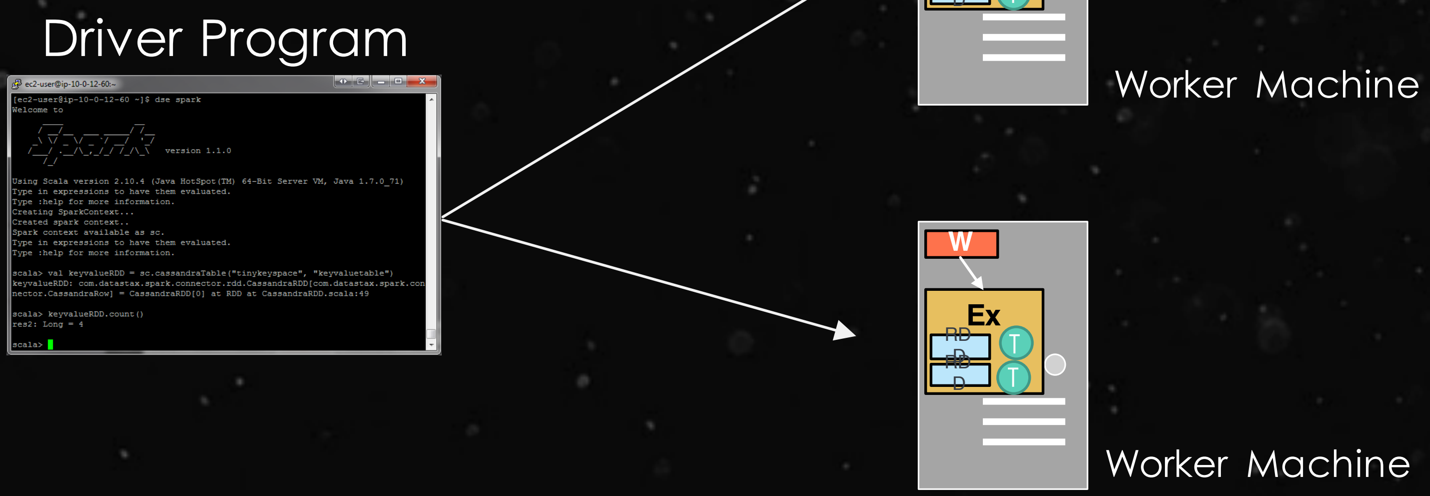
databricks

INTERACTIVE SHELL



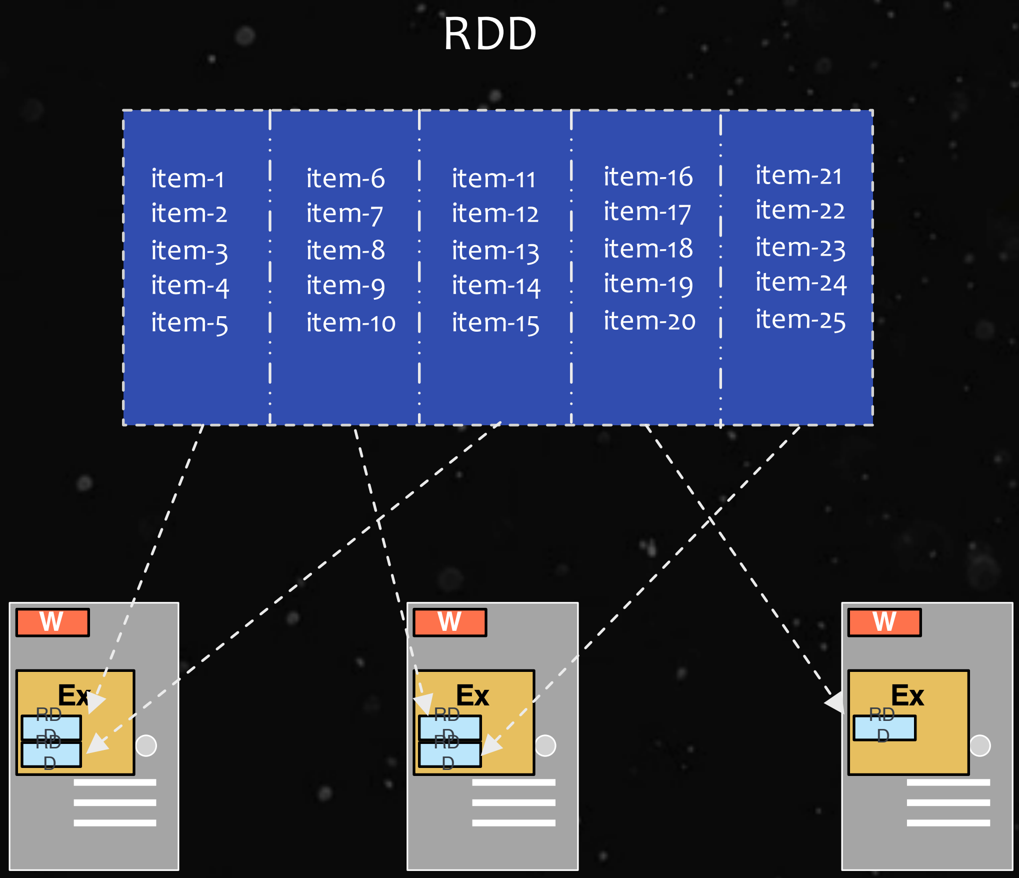
```
ubuntu@ip-10-0-53-24:~$ dse spark
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 0.9.1
      /_/

Using Scala version 2.10.3 (Java HotSpot(TM) 64-Bit Server VM, Java 1.7.0_51)
Type in expressions to have them evaluated.
Type :help for more information.
Creating SparkContext...
Created spark context..
Spark context available as sc.
Type in expressions to have them evaluated.
Type :help for more information.

scala> val myRDD = sc.cassandraTable("tinykeyspace", "keyvaluetable")
myRDD: com.datastax.bdp.spark.CassandraRDD[com.datastax.bdp.spark.CassandraRow] = Cassan
draRDD[0] at RDD at CassandraRDD.scala:32

scala> myRDD.count()
res2: Long = 5

scala>
```

(Scala, Python and R only)

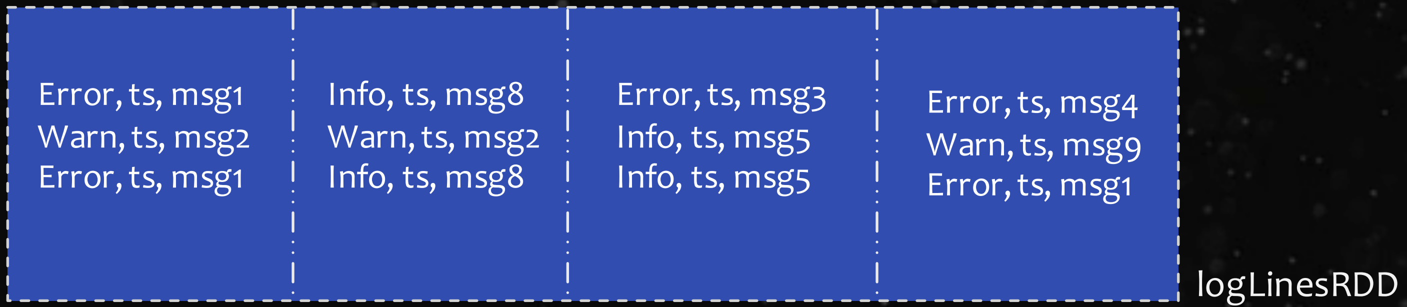


```
[ec2-user@ip-10-0-12-60 ~]$ dse spark
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 1.1.0
      /_/

Using Scala version 2.10.4 (Java HotSpot(TM) 64-Bit Server VM, Java 1.7.0_71)
Type in expressions to have them evaluated.
Type :help for more information.
Creating SparkContext...
Created spark context..
Spark context available as sc.
Type in expressions to have them evaluated.
Type :help for more information.

scala> val keyvalueRDD = sc.cassandraTable("tinykeyspace", "keyvaluetable")
keyvalueRDD: com.datastax.spark.connector.rdd.CassandraRDD[com.datastax.spark.con
nector.CassandraRow] = CassandraRDD[0] at RDD at CassandraRDD.scala:49

scala> keyvalueRDD.count()
res2: Long = 4

scala> 
```

more partitions = more parallelism



RDD w/ 4 partitions



A base RDD can be created 2 ways:

- Parallelize a collection
- Read data from an external source (S3, C*, HDFS, etc)

PARALLELIZE

```
# Parallelize in Python
wordsRDD = sc.parallelize([“fish", “cats“, “dogs”])
```

```
// Parallelize in Scala
val wordsRDD= sc.parallelize(List("fish", "cats", "dogs"))
```

```
// Parallelize in Java
JavaRDD<String> wordsRDD = sc.parallelize(Arrays.asList(“fish", “cats“, “dogs”));
```

- Take an existing in-memory collection and pass it to SparkContext's parallelize method
- Not generally used outside of prototyping and testing since it requires entire dataset in memory on one machine

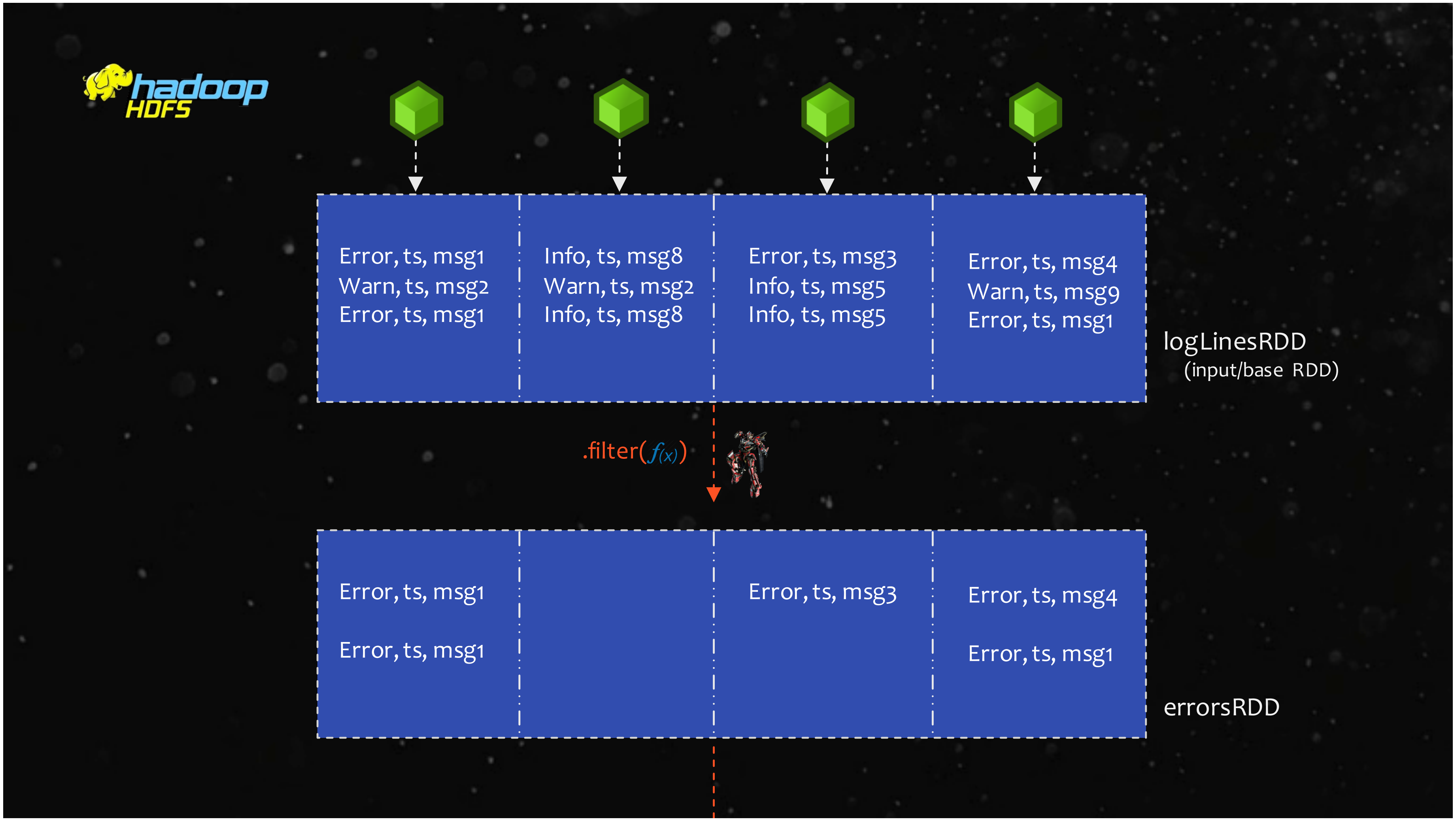
READ FROM TEXT FILE

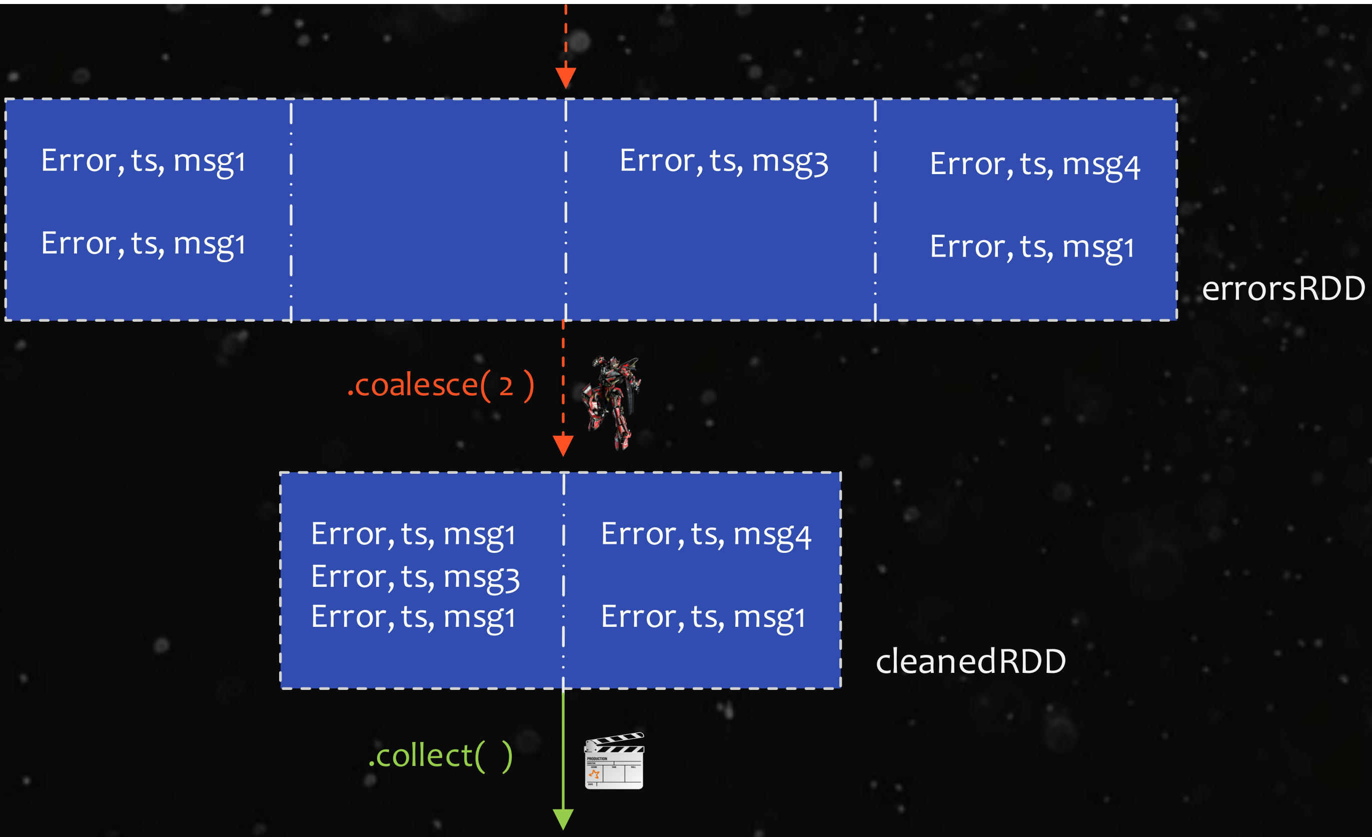
```
# Read a local txt file in Python
linesRDD = sc.textFile("/path/to/README.md")
```

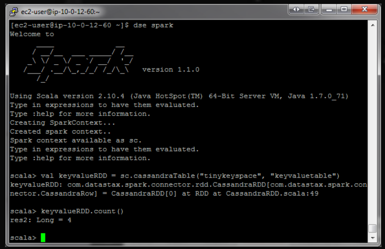
- There are other methods to read data from HDFS, C*, S3, HBase, etc.

```
// Read a local txt file in Scala
val linesRDD = sc.textFile("/path/to/README.md")
```

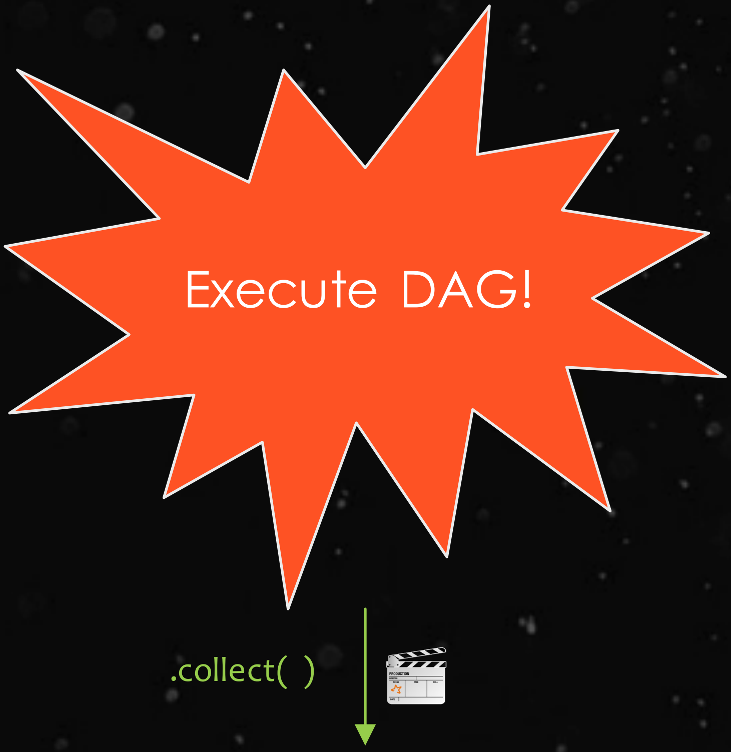
```
// Read a local txt file in Java
JavaRDD<String> lines = sc.textFile("/path/to/README.md");
```

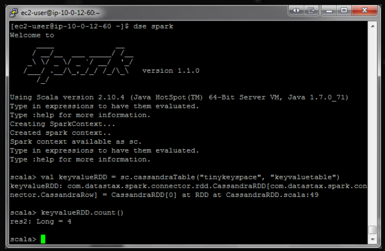

hadoop
HDFS
Error, ts, msg1
Warn, ts, msg2
Error, ts, msg1
Info, ts, msg8
Warn, ts, msg2
Info, ts, msg8
Error, ts, msg3
Info, ts, msg5
Info, ts, msg5
Error, ts, msg4
Warn, ts, msg9
Error, ts, msg1
logLinesRDD
(input/base RDD)
.filter(f(x))
Error, ts, msg1
Error, ts, msg1
Error, ts, msg3
Error, ts, msg4
Error, ts, msg1
errorsRDD





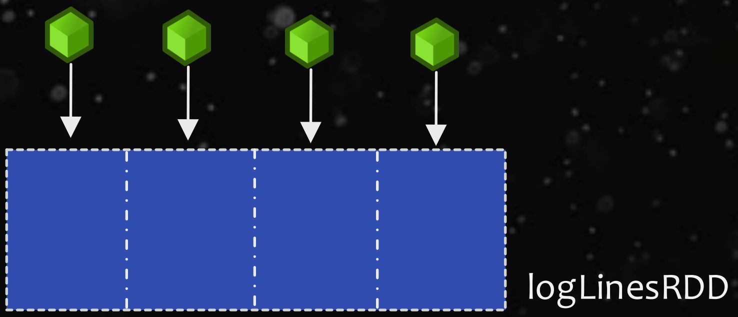
```
[ec2-user@ip-10-0-12-60 ~]$ dse spark
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 1.1.0
      /_/

Using Scala version 2.10.4 (Java HotSpot(TM) 64-Bit Server VM, Java 1.7.0_71)
Type in expressions to have them evaluated.
Type :help for more information.
Creating SparkContext...
Created spark context..
Spark context available as sc.
Type in expressions to have them evaluated.
Type :help for more information.

scala> val keyvalueRDD = sc.cassandraTable("tinykeyspace", "keyvaluetable")
keyvalueRDD: com.datastax.spark.connector.rdd.CassandraRDD[com.datastax.spark.con
nector.CassandraRow] = CassandraRDD[0] at RDD at CassandraRDD.scala:49

scala> keyvalueRDD.count()
res2: Long = 4

scala> 
```

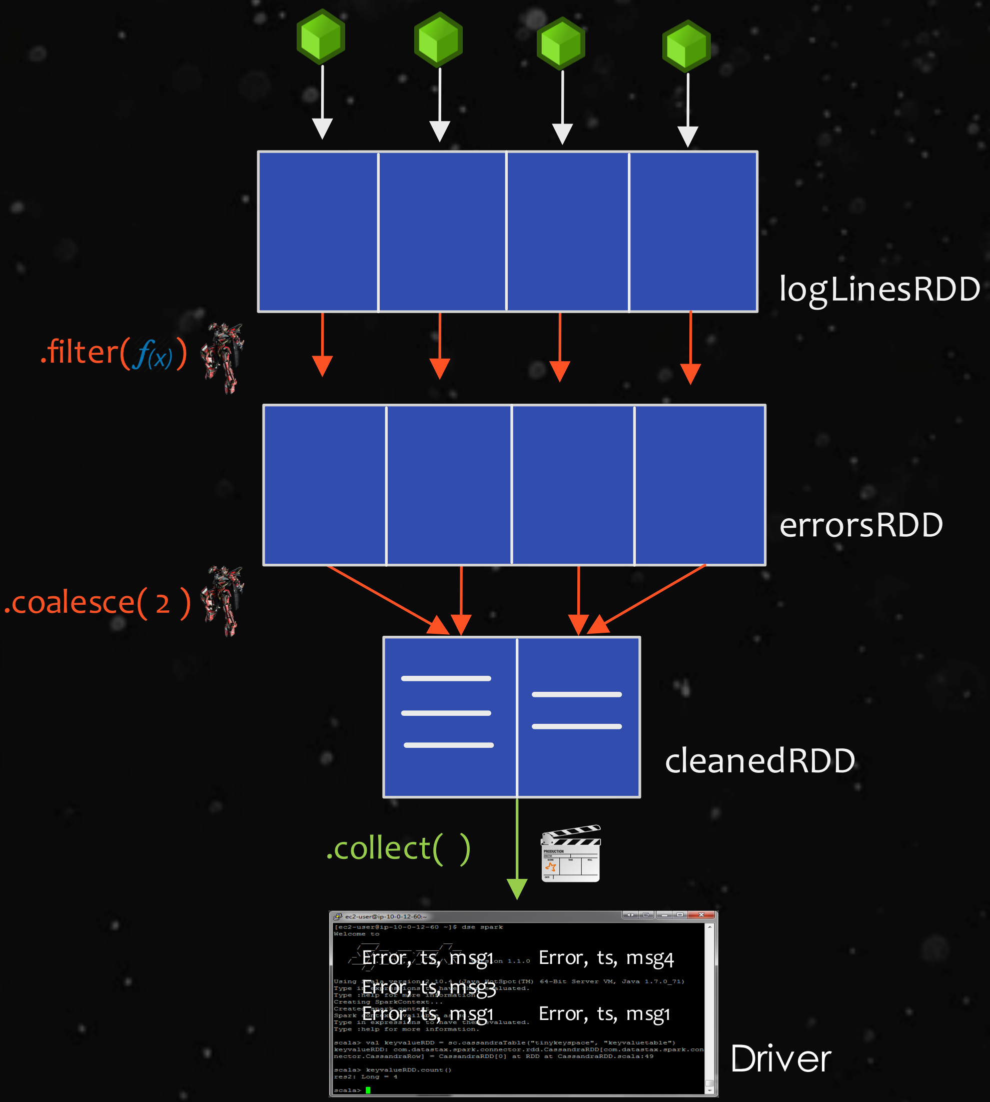




```
[ec2-user@ip-10-0-12-60 ~]$ dse spark
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 1.1.0
      /_/

Using Scala version 2.10.4 (Java HotSpot(TM) 64-Bit Server VM, Java 1.7.0_71)
Type in expressions to have them evaluated.
Type :help for more information.
Creating SparkContext...
Created spark context..
Spark context available as sc.
Type in expressions to have them evaluated.
Type :help for more information.

scala> val keyvalueRDD = sc.cassandraTable("tinykeyspace", "keyvaluetable")
keyvalueRDD: com.datastax.spark.connector.rdd.CassandraRDD[com.datastax.spark.con
nector.CassandraRow] = CassandraRDD[0] at RDD at CassandraRDD.scala:49

scala> keyvalueRDD.count()
res2: Long = 4

scala> 
```

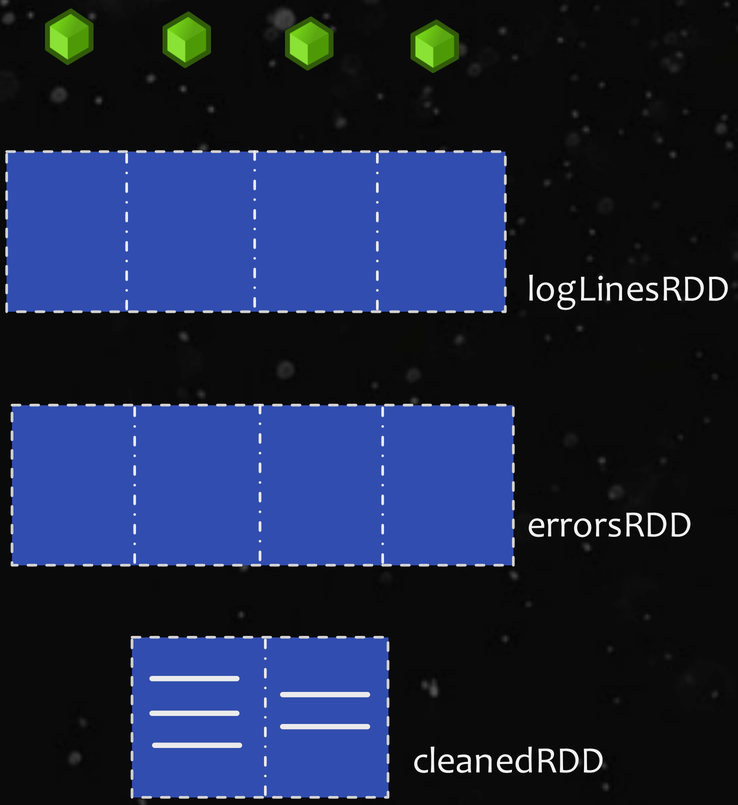
Driver

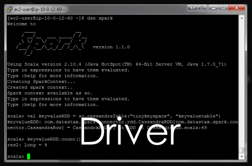

logLinesRDD



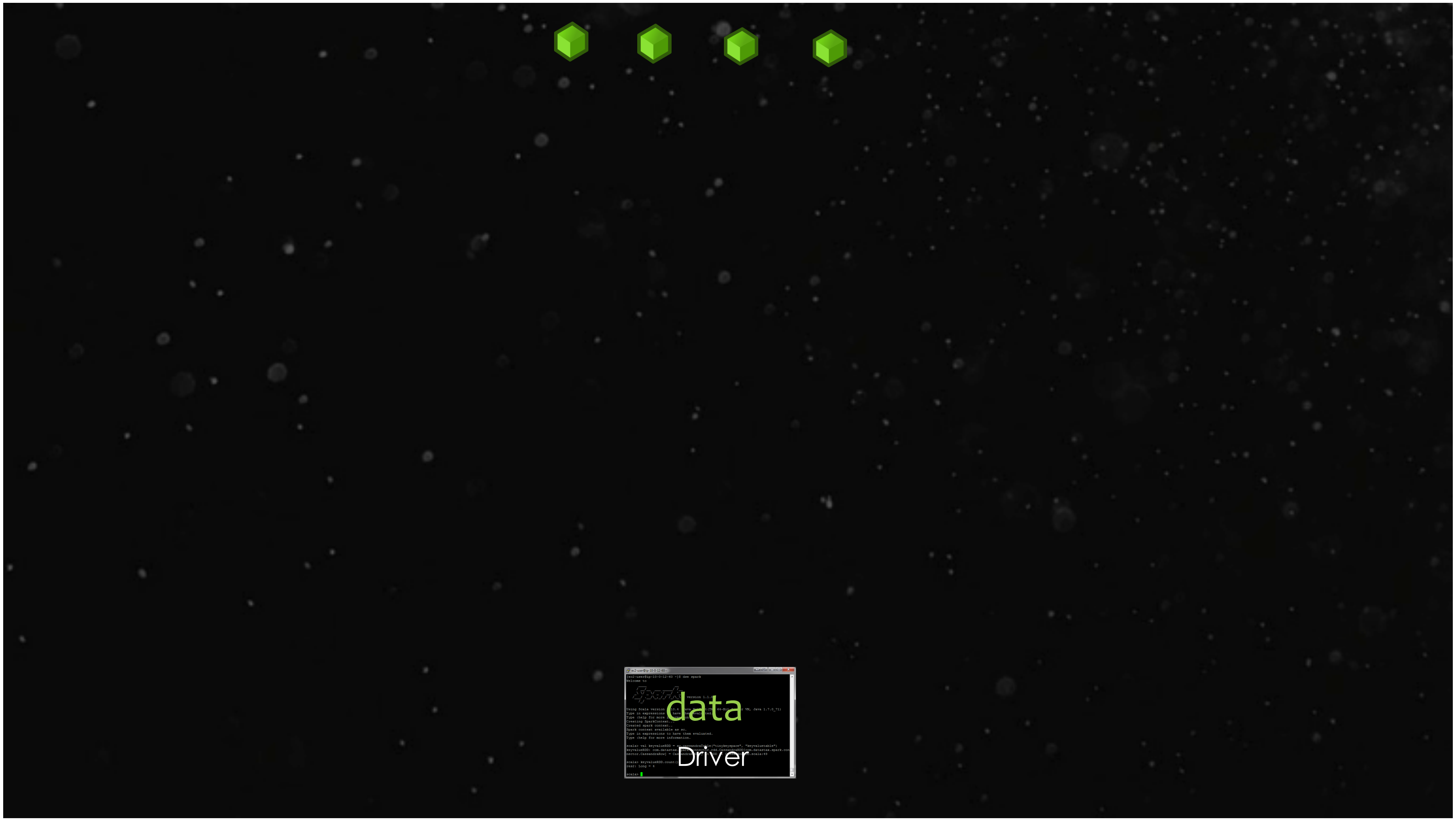
```
Error, ts, msg1     Error, ts, msg4
Error, ts, msg3
Error, ts, msg1     Error, ts, msg1
```

```
ec2-user@ip-10-0-12-60:~
[ec2-user@ip-10-0-12-60 ~]$ dse spark
Welcome to
                                        version 1.1.0

Using Scala version 2.10.4 (Java HotSpot(TM) 64-Bit Server VM, Java 1.7.0_71)
Type in expressions to have them evaluated.
Type :help for more information.
Creating SparkContext...
Created spark context..
Spark context available as sc.
Type in expressions to have them evaluated.
Type :help for more information.

scala> val keyvalueRDD = sc.cassandraTable("tinykeyspace", "keyvaluetable")
keyvalueRDD: com.datastax.spark.connector.rdd.CassandraRDD[com.datastax.spark.con
nector.CassandraRow] = CassandraRDD[0] at RDD at CassandraRDD.scala:49

scala> keyvalueRDD.count()
res2: Long = 4

scala> 
```

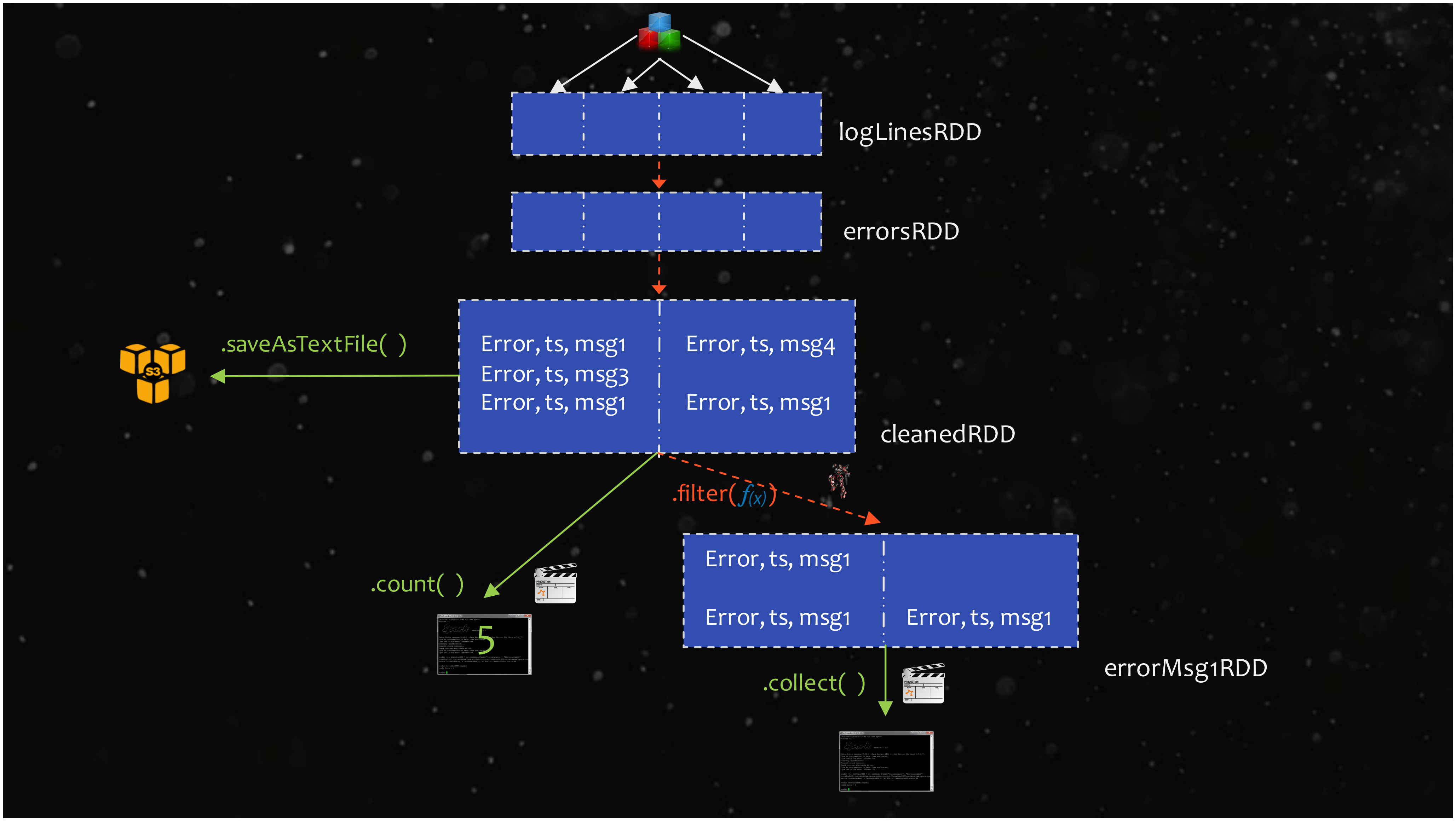


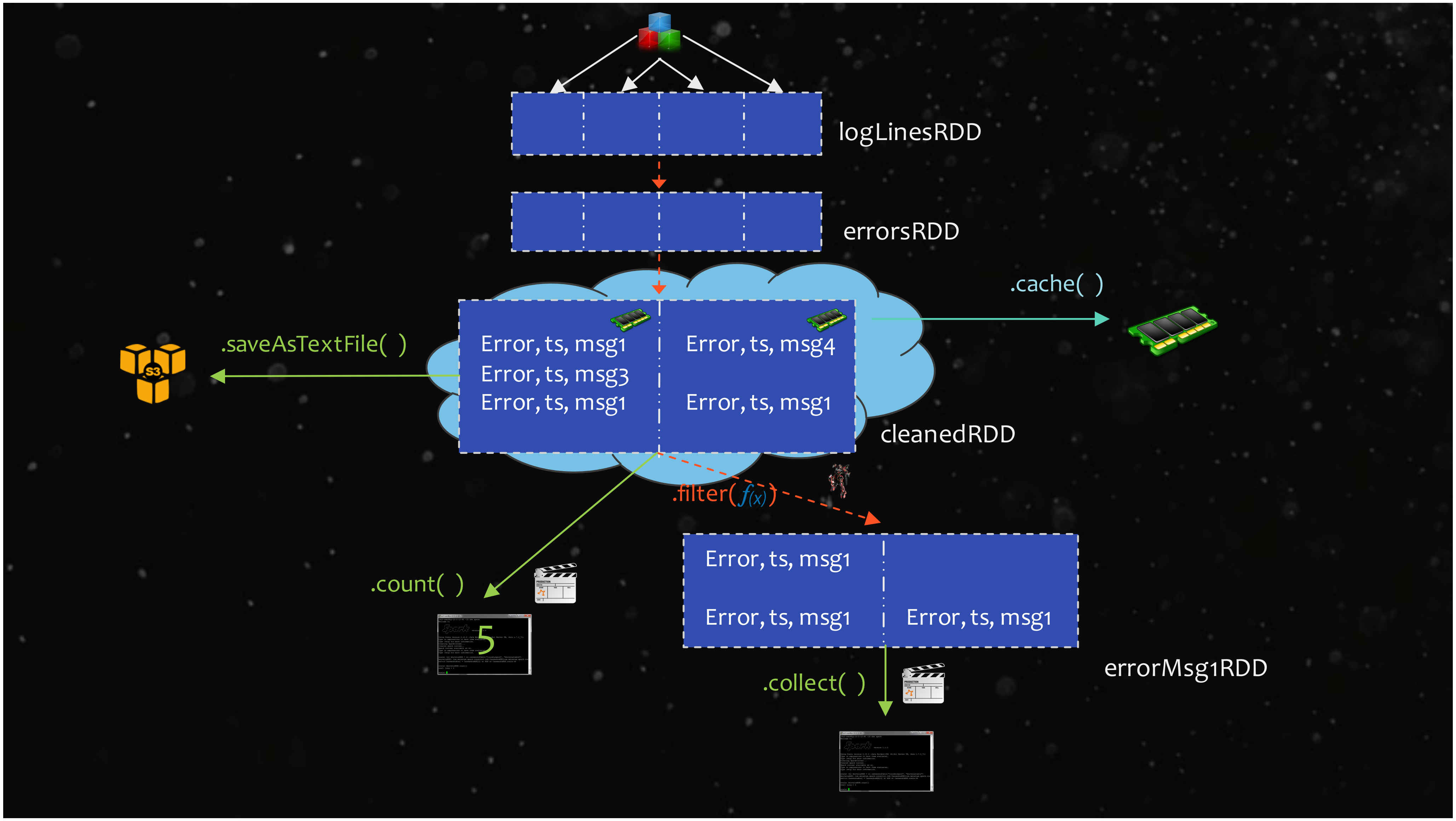


```
[ec2-user@ip-10-0-12-60 ~]$ dse spark
Welcome to
      Spark   version 1.1.0

Using Scala version 2.10.4 (Java HotSpot(TM) 64-Bit Server VM, Java 1.7.0_71)
Type in expressions to have them evaluated.
Type :help for more information.
Creating SparkContext...
Created spark context..
Spark context available as sc.
Type in expressions to have them evaluated.
Type :help for more information.

scala> val keyvalueRDD = sc.cassandraTable("tinykeyspace", "keyvaluetable")
keyvalueRDD: com.datastax.spark.connector.rdd.CassandraRDD[com.datastax.spark.con
nector.CassandraRow] = CassandraRDD[0] at RDD at CassandraRDD.scala:49

scala> keyvalueRDD.count()
res2: Long = 4

scala>
```



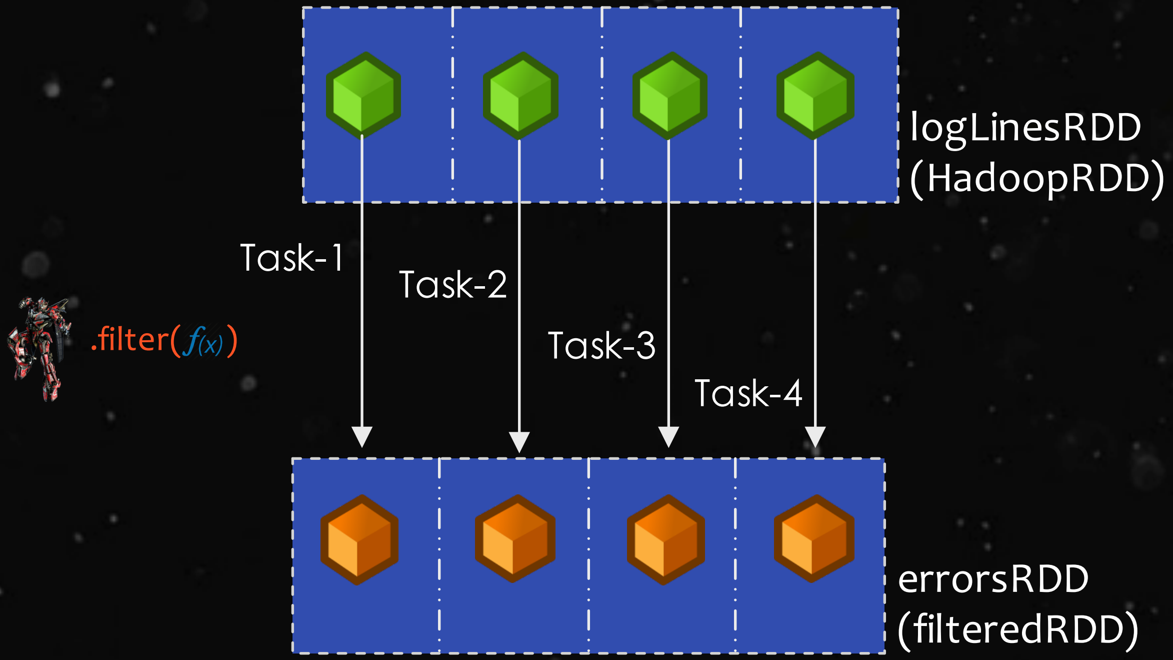
```
[ec2-user@ip-10-0-12-60 ~]$ dse spark
Welcome to
                                      version 1.1.0

Using Scala version 2.10.4 (Java HotSpot(TM) 64-Bit Server VM, Java 1.7.0_71)
Type in expressions to have them evaluated.
Type :help for more information.
Creating SparkContext...
Created spark context..
Spark context available as sc.
Type in expressions to have them evaluated.
Type :help for more information.

scala> val keyvalueRDD = sc.cassandraTable("tinykeyspace", "keyvaluetable")
keyvalueRDD: com.datastax.spark.connector.rdd.CassandraRDD[com.datastax.spark.con
nector.CassandraRow] = CassandraRDD[0] at RDD at CassandraRDD.scala:49

scala> keyvalueRDD.count()
res2: Long = 4

scala>
```


logLinesRDD
errorsRDD
.saveAsTextFile( )
Error, ts, msg1
Error, ts, msg3
Error, ts, msg1
Error, ts, msg4
Error, ts, msg1
cleanedRDD
.filter(f(x))
.count( )
5
Error, ts, msg1
Error, ts, msg1
Error, ts, msg1
.collect( )
errorMsg1RDD



logLinesRDD
errorsRDD
.cache( )
.saveAsTextFile( )
Error, ts, msg1
Error, ts, msg3
Error, ts, msg1
Error, ts, msg4
Error, ts, msg1
cleanedRDD
.filter(f(x))
.count( )
5
Error, ts, msg1
Error, ts, msg1
Error, ts, msg1
errorMsg1RDD
.collect( )

PARTITION -> TASK -> PARTITION



LIFECYCLE OF A SPARK PROGRAM

1) Create some input RDDs from external data or parallelize a collection in your driver program.

2) Lazily transform them to define new RDDs using transformations like `filter()` or `map()`

3) Ask Spark to `cache()` any intermediate RDDs that will need to be reused.

4) Launch actions such as `count()` and `collect()` to kick off a parallel computation, which is then optimized and executed by Spark.

TRANSFORMATIONS (lazy)

`map()`	`intersection()`	`cartesion()`
`flatMap()`	`distinct()`	`pipe()`
`filter()`	`groupByKey()`	`coalesce()`
`mapPartitions()`	`reduceByKey()`	`repartition()`
`mapPartitionsWithIndex()`	`sortByKey()`	`partitionBy()`
`sample()`	`join()`	...
`union()`	`cogroup()`	...

ACTIONS

```
reduce()                takeOrdered()

collect()               saveAsTextFile()

count()                 saveAsSequenceFile()

first()                 saveAsObjectFile()

take()                  countByKey()

takeSample()            foreach()

saveToCassandra()       ...
```

TYPES OF RDDS

- HadoopRDD
- FilteredRDD
- MappedRDD
- PairRDD
- ShuffledRDD
- UnionRDD
- PythonRDD

- DoubleRDD
- JdbcRDD
- JsonRDD
- VertexRDD
- EdgeRDD

- CassandraRDD *(DataStax)*
- GeoRDD *(ESRI)*
- EsSpark *(ElasticSearch)*