Executing Spark commands in Databricks Cc

```
import org.apache.spark.SparkContext

val sc = SparkContext.getOrCreate()  // Use exsiting SparkContext in
DataBricks Community
val rdd:RDD[String] = sc.textFile("/FileStore/tables/wikipedia.dat")
val containsWord = rdd.filter(x => x.contains("you")).persist()
val numberOfTimes = containsWord.count()  //Obtain number of sentences
val firstFew = containsWord.take(1)  //Obtain sentences containing chosen
word
```

```
import org.apache.spark.SparkContext
sc: org.apache.spark.SparkContext = org.apache.spark.SparkContext@34a07fac
rdd: org.apache.spark.rdd.RDD[String] = /FileStore/tables/wikipedia.dat MapP
artitionsRDD[12] at textFile at command-1513168191355444:4
containsWord: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[13] at fil
ter at command-1513168191355444:5
numberOfTimes: Long = 1799
firstFew: Array[String] = Array("<page><title>Wikipedia:WikiProject Spam/Lin
kReports/euchems-congress2010.de</title><text><!--Please do not comment or c
hange this page, it is bot generated and will be completely regenerated by t
he bot. If you want to comment, please do so on the talkpage.--> {{User:COIB
ot/Summary/LinkReports}} {{User:COIBot/linksaverdatabases|||||}}  <!-- tags
and categories -->{{NOINDEX}} == Links == * {{LinkSummary|euchems-congress20
10.de}} :* euchems-congress2010.de resolves to [//192.185.16.215 192.185.16.
215] :*: {{LinkSummary|192.185.16.215}} :* Link is not on the [[en:User:COIB
ot#Blacklist|blacklist]]. :* Link is not on the [[en:User:COIBot#Domainredli
st|domainredlist]]. :* Link is not on the [[en:User:COIBot#Monitorlist|Monit
orlist]]. :* None of the mentioned users is on the [[en:User:COIBot#Blacklis
t|blacklist]]. :* Link is not on the [[en:User:COIBot#Whitelist|whitelist]].
:* Link is not on the [[en:User:COIBot#Monitor list|monitor list]]. :* Link
is blacklisted by 1 on [//da.wikipedia.org/wiki/Mediawiki:Spam-blacklist da.
wikipedia.org] == Users == * {{IPSummary|89.165.165.14}} == Additions == {{U
ser:COIBot/Additionlist_top}} {{User:COIBot/EditSummary|id=32148914|lang=nl|
wikidomain=b|namespace=|pagename=Picking Simple Systems For clash of clans|u
sername=89.165.165.14|link=euchems-congress2010.de/?p=347|sortdomain=de.euch
ems-congress2010.|domain=euchems-congress2010.de|origdiff=http://nl.wikibook
s.org/w/index.php?oldid=271350&rcid=219394|resolved=X|isIP=1|date=2014-10-05
|time=01:26:02|wiki=nl.wikibooks.org|revid=271350|oldid=0|usercount=3|whitel
isted=0|blacklisteduser=0|whitereason=|blackreason=|deleted=0|top=0|there=1|
checked=1|coiflag=0|otherlinks={{User:COIBot/OtherLinks|link=www.google.de/s
earch?q=arguing+parties|domain=google.de|U=3|L=2120|UL=X|WUL=X|base=|basedom
ain=|baseip=}}{{User:COIBot/OtherLinks|link=euchems-congress2010.de/?p=347|d
omain=euchems-congress2010.de|U=3|L=1|UL=1|WUL=1|base=|basedomain=|baseip=}}
{{User:COIBot/OtherLinks|link=www.google.com/search?q=Mohammed&amp;btnI=luck
```

```
y|domain=google.com|U=3|L=-1|UL=X|WUL=X|base=|basedomain=|baseip=}}}} {{Use
r:COIBot/Additionlist_bottom}} * Displayed all 1 additions.</text></page>",)
```

Creating Spark dataframes

```
import org.apache.spark.sql.functions._
import org.apache.spark.sql.SparkSession

// Create spark session
val spark:SparkSession = SparkSession.builder()
                                     .appName("MyApp")
                                     .getOrCreate()

// Create spark context
import spark.implicits._

// Define dataset
case class Employee(id: Int, fname:String, lname:String, work:Array[Int],
city:String, age:Int)
val emp = Seq(Employee(12,"Joe","Smith",Array(38,67,89),"New York",34),
                Employee(645,"Slate","Markham",Array(28,3),"Sydney",2),
                Employee(12,"Sally","Owens",Array(48,1,0),"New York",12),
                Employee(221,"Joe","Walker",Array(21),"Sydney",89),
                Employee(12,"Joe","Runner",Array(21),"Sydney",89),
                Employee(645,"Slate","Ontario",Array(12,3),"Wellington",25))
// Convert dataset to RDD
val empRDD = spark.sparkContext.parallelize(emp)
// Convert dataset to Dataframe
val empDF = emp.toDF
// Convert Dataframe to RDD
val empRDDfromDF = empDF.rdd

// Visualize Dataframe
empDF.printSchema()
empDF.show()
```

```
root
 |-- id: integer (nullable = false)
 |-- fname: string (nullable = true)
 |-- lname: string (nullable = true)
 |-- work: array (nullable = true)
 |    |-- element: integer (containsNull = false)
 |-- city: string (nullable = true)
 |-- age: integer (nullable = false)

+---+-----+-------+------------+----------+---+
| id|fname|  lname|        work|      city|age|
+---+-----+-------+------------+----------+---+
| 12|  Joe|  Smith|[38, 67, 89]|  New York| 34|
|645|Slate|Markham|     [28, 3]|    Sydney|  2|
```

```
| 12|Sally|  Owens|  [48, 1, 0]|  New York| 12|
|221|  Joe| Walker|        [21]|    Sydney| 89|
| 12|  Joe| Runner|        [21]|    Sydney| 89|
|645|Slate|Ontario|     [12, 3]|Wellington| 25|
+---+-----+-------+------------+----------+---+
```

Transformations on dataframes

```
val selectedemp = empDF.select($"id",$"lname")
                       .where($"city" === "Sydney")
                       .orderBy($"id")
selectedemp.show()

val rankedemp = empDF.groupBy($"id")
                     .max("age")
rankedemp.show()

val rank = empDF.groupBy($"id",$"fname")
                .agg(count($"id"))
                .orderBy($"fname",$"count(id)".desc)

rank.show()
```

```
+---+-------+
| id|  lname|
+---+-------+
| 12| Runner|
|221| Walker|
|645|Markham|
+---+-------+

+---+--------+
| id|max(age)|
+---+--------+
| 12|      89|
|645|      25|
|221|      89|
+---+--------+

+---+-----+---------+
| id|fname|count(id)|
+---+-----+---------+
| 12|  Joe|        2|
|221|  Joe|        1|
```

Joins on dataframes

```
+---+--------------------+
| id|                   v|
```

```
+---+--------------------+
|101|        [Ruetli, AG]|
|102| [Brelaz, DemiTarif]|
|103|[Gress, DemiTarif...|
|104|[Schatten, DemiTa...|
+---+--------------------+

+---+---------+
| id|        v|
+---+---------+
|101|     Bern|
|101|     Thun|
|102| Lausanne|
|102|   Geneve|
|102|     Nyon|
|103|   Zurich|
|103|St-Gallen|
|103|     Chur|
+---+---------+
```