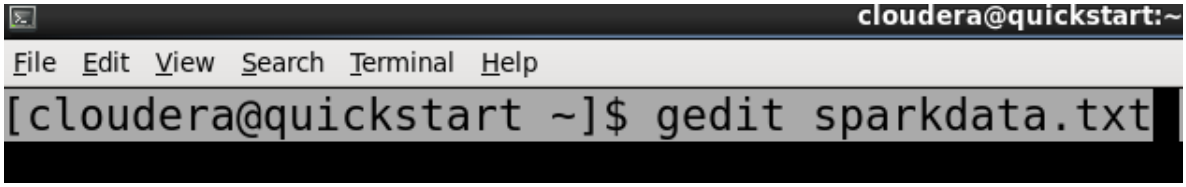
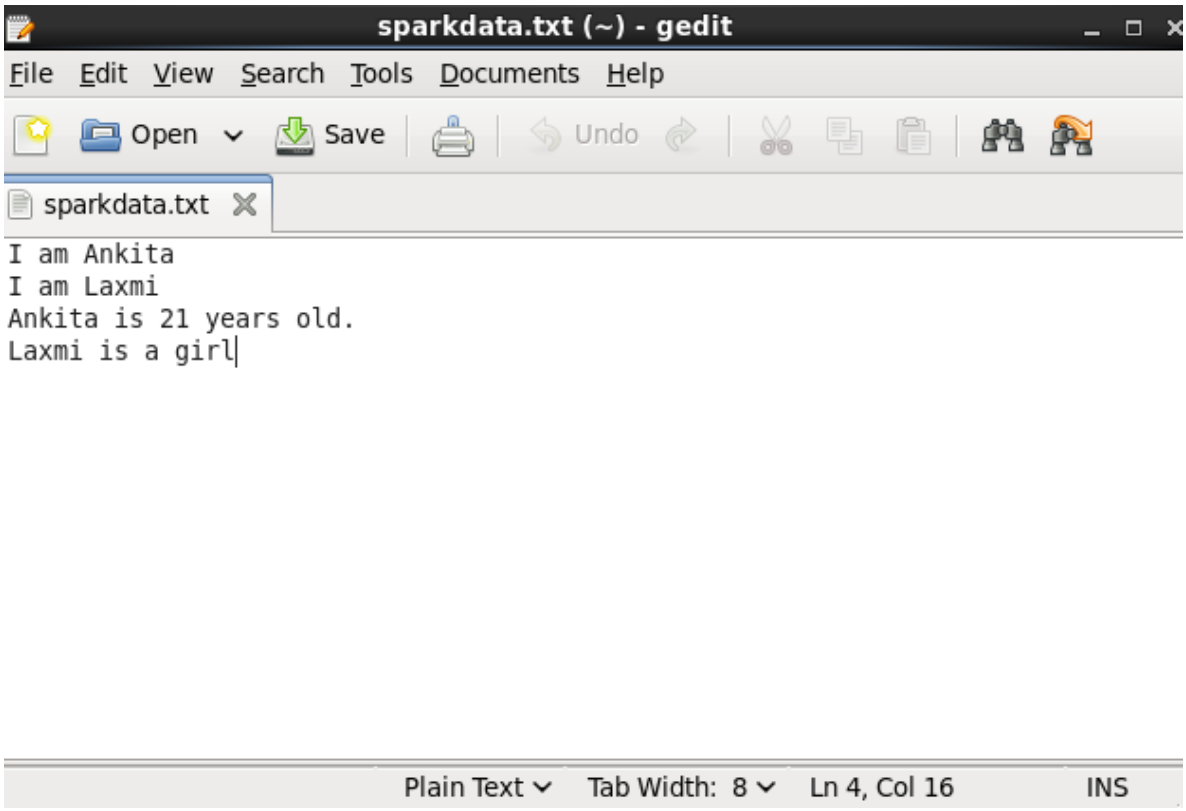
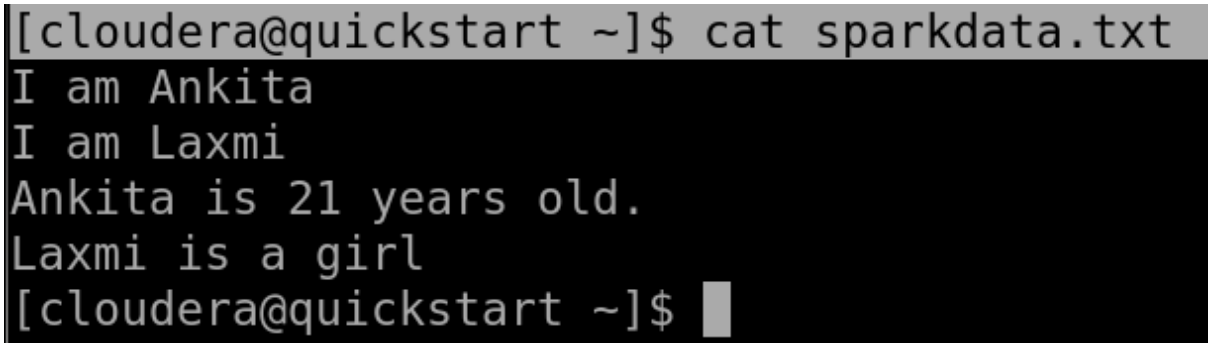
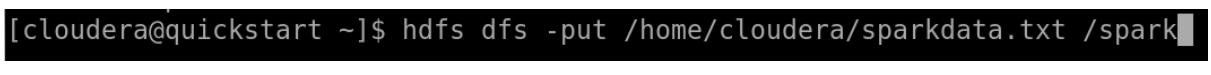

cloudera@quickstart:~
File Edit View Search Terminal Help
[cloudera@quickstart ~]$ gedit sparkdata.txt

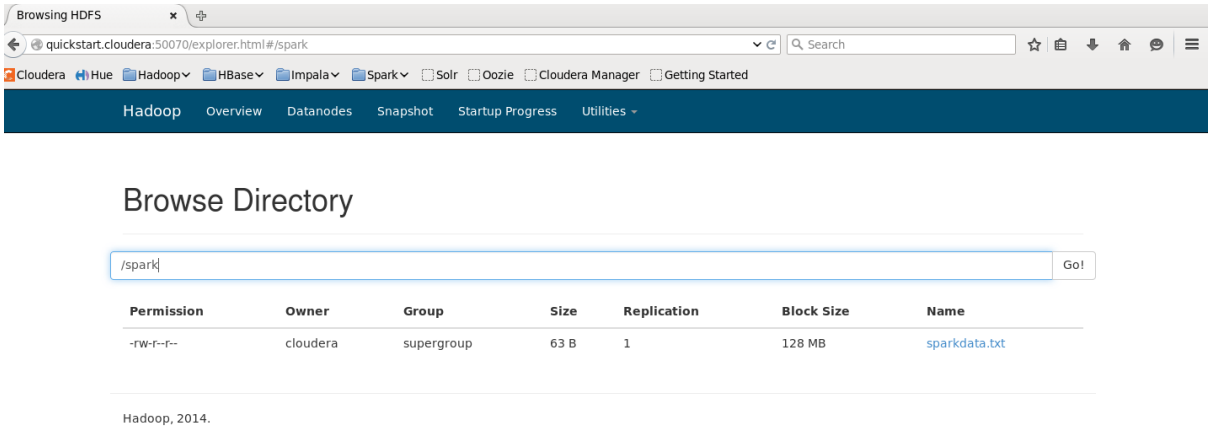

sparkdata.txt (~) - gedit
File Edit View Search Tools Documents Help
Open
Save
Undo
sparkdata.txt
I am Ankita
I am Laxmi
Ankita is 21 years old.
Laxmi is a girl
Plain Text
Tab Width: 8
Ln 4, Col 16
INS


[cloudera@quickstart ~]$ cat sparkdata.txt
I am Ankita
I am Laxmi
Ankita is 21 years old.
Laxmi is a girl
[cloudera@quickstart ~]$


[cloudera@quickstart ~]$ hdfs dfs -mkdir /spark


[cloudera@quickstart ~]$ hdfs dfs -put /home/cloudera/sparkdata.txt /spark

```
[cloudera@quickstart ~]$ hdfs dfs -ls /spark
Found 1 items
-rw-r--r--   1 cloudera supergroup         63 2023-09-01 00:10 /spark/sparkdata.
txt
[cloudera@quickstart ~]$
```




```
[cloudera@quickstart ~]$ spark-shell
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/zookeeper/lib/slf4j-log4j12-1.7.5.jar
!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/jars/slf4j-log4j12-1.7.5.jar!/org/slf4j/i
mpl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
```


```
scala> val data=sc.textFile("/spark/sparkdata.txt")
23/09/01 00:19:21 INFO storage.MemoryStore: ensureFreeSpace(277083) called with curMem=622147, maxMem
23/09/01 00:19:21 INFO storage.MemoryStore: Block broadcast_3 stored as values in memory (estimated s
23/09/01 00:19:21 INFO storage.MemoryStore: ensureFreeSpace(21083) called with curMem=899230, maxMem=
23/09/01 00:19:21 INFO storage.MemoryStore: Block broadcast_3_piece0 stored as bytes in memory (estim
23/09/01 00:19:21 INFO storage.BlockManagerInfo: Added broadcast_3_piece0 in memory on localhost:4389
23/09/01 00:19:21 INFO storage.BlockManagerMaster: Updated info of block broadcast_3_piece0
23/09/01 00:19:21 INFO spark.SparkContext: Created broadcast 3 from textFile at <console>:21
data: org.apache.spark.rdd.RDD[String] = /spark/sparkdata.txt MapPartitionsRDD[7] at textFile at <con
```

```
scala> data.collect;
23/09/01 00:19:25 INFO mapred.FileInputFormat: Total input paths to process : 1
23/09/01 00:19:25 INFO spark.SparkContext: Starting job: collect at <console>:24
23/09/01 00:19:25 INFO scheduler.DAGScheduler: Got job 0 (collect at <console>:24) with 1 output part
23/09/01 00:19:25 INFO scheduler.DAGScheduler: Final stage: Stage 0(collect at <console>:24)
23/09/01 00:19:25 INFO scheduler.DAGScheduler: Parents of final stage: List()
```

```
23/09/01 00:19:26 INFO scheduler.DAGScheduler: Job 0 finished: collect at <console>:24, took 1.2744
res3: Array[String] = Array(I am Ankita, I am Laxmi, Ankita is 21 years old., Laxmi is a girl)
```

```
scala> val splitdata=data.flatMap(line=>line.split(" "));
splitdata: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD
[10] at flatMap at <console>:23

scala> splitdata.collect;
```

```
 collect at <console>:26, took 0.058715 s
res5: Array[String] = Array(I, am, Ankita, I, am, Laxmi, Ankit
a, is, 21, years, old., Laxmi, is, a, girl)
```

```
scala> val mapdata = splitdata.map(word=>(word,1));
mapdata: org.apache.spark.rdd.RDD[(String, Int)] = MapPartitio
nsRDD[11] at map at <console>:25

scala> mapdata.collect;
```

```
 collect at <console>:28, took 0.048377 s
res6: Array[(String, Int)] = Array((I,1), (am,1), (Ankita,1),
(I,1), (am,1), (Laxmi,1), (Ankita,1), (is,1), (21,1), (years,1
), (old.,1), (Laxmi,1), (is,1), (a,1), (girl,1))
```

```
scala> val reducedata = mapdata.reduceByKey(_+_);
reducedata: org.apache.spark.rdd.RDD[(String, Int)] = Shuffled
RDD[12] at reduceByKey at <console>:27

scala> reducedata.collect;
```

```
res7: Array[(String, Int)] = Array((Ankita,2), (Laxmi,2), (21,
1), (is,2), (old.,1), (am,2), (girl,1), (a,1), (I,2), (years,1
))
```