

Using Apache Spark : With Cloudera quickstart_tutorial

<wordcount>

로컬 파일 확인

```
> ls  
1
```

Downloads txt file to execute wordcount example

```
> wget http://inst.eecs.berkeley.edu/~cs61a/fa11/shakespeare.txt  
2
```

```
> ls  
3
```

하둡에 폴더 만들기

```
> sudo -u hdfs hadoop fs -mkdir /spark_exam/
```

하둡으로 복사

```
> hadoop fs -copyFromLocal ./shakespeare.txt /spark_exam
```

확인

```
> hadoop fs -ls /spark_exam  
4
```

스파크 띄우기

```
> spark-shell --jars /usr/lib/avro/avro-mapred.jar \ --conf  
spark.serializer=org.apache.spark.serializer.KryoSerializer  
5,6
```

```
> val file = sc.textFile("hdfs://quickstart.cloudera/spark_exam/  
shaekspeare.txt")  
7
```

```
> val counts = file.flatMap(line => line.split(" ")).map(word => (word,  
1)).reduceByKey(_ + _)  
8
```

```
> counts.saveAsTextFile("hdfs://quickstart.cloudera/spark_exam/wordcount")  
9,10
```

```
> counts.count()  
11,12
```

```
> counts.toArray().foreach(println)  
13
```

```
> exit
```

```
>hadoop fs -ls /spark_exam/wordcount  
14
```

<estimating Pi>

```
>spark-shell --jars /usr/lib/avro/avro-mapred.jar \  
--conf spark.serializer=org.apache.spark.serializer.KryoSerializer
```

```
> val count = sc.parallelize(1 to NUM_SAMPLE).map{i =>  
    val x = Math.random()  
    val y = Math.random()  
    if(x*x + y*y <1) 1 else 0  
}.reduce(_ + _)  
1
```

```
> println("Pi is roughly " + 4.0*count / NUM_SAMPLE)  
2
```

* NUM_SAMPLE : the number of throwing dart.

> ls

> wget

확인

> ls

> sudo -u hdfs hadoop fs -mkdir /spark_exam/logistics_data

하둡으로 복사

> hadoop fs -copyFromLocal ./sample_libsvm_data.txt /spark_exam/
logistics_data

확인

> hadoop fs -ls /spark_exam/linear_data

스파크 띄우기

> spark-shell --jars /usr/lib/avro/avro-mapred.jar \ --conf
spark.serializer=org.apache.spark.serializer.KryoSerializer

>