Apache Spark

Practical-10

In Python and Scala Interface

CloudxLAB (You can install it on your local machine too.)

Main Focus on PYTHON Interface

Example: “Average Number of Friends by Age.”

CloudxLab

Must Know How to Open both the interface and run the code.

By default Interface is – Scala Language

Dataset :

FAKEFRIENDS.CSV

WHAT IS IN IT ? LET'S SEE !!

Dataset :

userID	name	age	friends
0	Will	33	385
1	Jean-Luc	26	2
2	Hugh	55	221
3	Deanna	40	465
4	Quark	68	21
5	Weyoun	59	318
6	Gowron	37	220

Spark Functions:

01

redeceBykey(): combine values with same keys

02

GroupBykey(): group values with the same keys

03

SortBykey(): sort RDD by key values.

Setting up the spark configurations:

```
from pyspark import SparkConf, SparkContext

conf = SparkConf().setMaster("local").setAppName("FriendsByAge")
sc = SparkContext(conf = conf)
```

Code:

- *lines = sc.textFile("file:///Spark/fakefriends.csv")*

- *rdd = lines.map(parseLine)*

Code:

userID	name	age	friends
0	Will	33	385
1	Jean-Luc	26	2
2	Hugh	55	221
3	Deanna	40	465
4	Quark	68	21
5	Weyoun	59	318
6	Gowron	37	220

```
def parseLine(line):
    fields = line.split(',')
    age = int(fields[2])
    numFriends = int(fields[3])
    return (age, numFriends)
```

Output:

- 33,385

  33,2

  55,221

  40,465

.....

Code:

- totalsByAge = rdd.mapValues(lambda x: (x, 1)).reduceByKey(lambda x, y: (x[0] + y[0], x[1] + y[1]))

Line-1:

rdd.mapValues(lambda x: (x, 1)).

rdd.mapValues(lambda x: (x, 1))

reduceByKey(lambda x, y: (x[0] + y[0], x[1] + y[1]))

Code:

- averagesByAge = totalsByAge.mapValues(lambda x: x[0] / x[1])

Code:

```
results = averagesByAge.collect()
for result in results:
   print(result)
```

Computing an Average: (scala)

Computing an Average: (scala)

```
var rdd = sc.parallelize(array(1.0,2,3,4,5,6,7),3);

var rdd_count = rdd.map((_,1))

var(sum,count)= rdd_count.reduce((x,y)=>x._1+y._1,x._2+y._2)

var avg  = sum/count
```

How to create an rdd ?

1. Loading the data file.

2. Distributing Object and Parallelize it

What describes Apache Pig best?

- ○ An SQL query processor
- ○ A database to store data on HDFS
- ○ An engine for executing data flows in parallel on Hadoop

What do Pigs fly mean?

- ○ Pig is designed for performance
- ○ Pig script is very light weight
- ○ Pig script is very fragile

Which statement is not part of Pig philosophy?

- ○ Pigs are domestic animals
- ○ Pigs fly
- ○ Pigs eat anything
- ○ Pigs live anywhere
- ○ Pigs oink

Pig converts most of its queries into sequences of MapReduce tasks and executes them

- ○ True
- ○ False

The command to start Pig is?

- ○ pig shell
- ○ pig
- ○ pig-cli
- ○ pig-client

How to run Pig in local mode?

- ○ pig -local
- ○ pig local
- ○ pig -x local

What does Spark streaming do?

- ○ It helps us in using our own programs as transformation or action.
- ○ It is an API on top of spark which processes continuous stream of data.
- ○ It lets us stream video or audio to a remote server
- ○ There is no such thing as spark streaming

When do Spark Streaming Applications End?

- ○ When the work is done
- ○ When the data in input streams is no longer available
- ○ When it is killed by the operating system or user.
- ○ It is automatically stopped if it runs beyond certain time.

Which of the following is not true about Spark Streaming?

- ○ Spark Streaming uses Apache Spark underneath to process data
- ○ Spark Streaming converts the incoming data into RDDs
- ○ Spark Streaming can not read from HDFS because HDFS is a static source not stream
- ○ Spark Streaming can save data into HDFS

Will start at 10.5 a.m.