Graphx

Cmd 1

```
1 import org.apache.spark._
2 import org.apache.spark.graphx._
3 // To make some of the examples work we will also need RDD
4 import org.apache.spark.rdd.RDD
```

```
import org.apache.spark._
import org.apache.spark.graphx._
import org.apache.spark.rdd.RDD
```

Command took 1.25 seconds -- by louhy1128@gmail.com at 2017/7/3 下午4:46:07 on My Cluster

Cmd 2

```
1 val users: RDD[(VertexId, (String, String))] =
2   sc.parallelize(Array((3L, ("rxin", "student")), (7L, ("jgonzal",
  "postdoc")),
3                        (5L, ("franklin", "prof")), (2L, ("istoica",
  "prof"))))
```

```
users: org.apache.spark.rdd.RDD[(org.apache.spark.graphx.VertexId, (String, Stri
ng))] = ParallelCollectionRDD[2040] at parallelize at <console>:43
```

Command took 0.40 seconds -- by louhy1128@gmail.com at 2017/7/3 下午4:46:16 on My Cluster

Cmd 3

```
1 val relationships: RDD[Edge[String]] =
2   sc.parallelize(Array(Edge(3L, 7L, "collab"),    Edge(5L, 3L, "advisor"),
3                        Edge(2L, 5L, "colleague"), Edge(5L, 7L, "pi")))
```

```
relationships: org.apache.spark.rdd.RDD[org.apache.spark.graphx.Edge[String]] =
 ParallelCollectionRDD[2041] at parallelize at <console>:43
```

Command took 0.15 seconds -- by louhy1128@gmail.com at 2017/7/3 下午4:46:25 on My Cluster

Cmd 4

```
1 val defaultUser = ("John Doe", "Missing")
```

```
defaultUser: (String, String) = (John Doe,Missing)
```

Command took 0.07 seconds -- by louhy1128@gmail.com at 2017/7/3 下午4:46:36 on My Cluster

Cmd 5

```
1 val graph = Graph(users, relationships, defaultUser)
```

```
graph: org.apache.spark.graphx.Graph[(String, String),String] = org.apache.spar
k.graphx.impl.GraphImpl@59f1de4a
```

Command took 0.14 seconds -- by louhy1128@gmail.com at 2017/7/3 下午4:46:44 on My Cluster

Cmd 6

```
1 graph.vertices.filter { case (id, (name, pos)) => pos == "postdoc" }.count
```

▸ (1) Spark Jobs

```
res0: Long = 1
```

Command took 0.26 seconds -- by louhy1128@gmail.com at 2017/7/3 下午4:47:55 on My Cluster

Cmd 7

```
1 graph.vertices.filter { case (id, (name, pos)) => pos == "postdoc" }
```

```
res1: org.apache.spark.graphx.VertexRDD[(String, String)] = VertexRDDImpl[2058] at RDD at VertexRDD.scala:55
```

Command took 0.09 seconds -- by louhy1128@gmail.com at 2017/7/3 下午4:48:33 on My Cluster

Cmd 8

```
1 graph.edges.filter(e => e.srcId > e.dstId).count
```

▸ (1) Spark Jobs

```
res2: Long = 1
```

Command took 0.12 seconds -- by louhy1128@gmail.com at 2017/7/3 下午4:49:10 on My Cluster

Cmd 9

```
1 graph.edges.filter(e => e.srcId > e.dstId).collect()
```

▸ (1) Spark Jobs

```
res3: Array[org.apache.spark.graphx.Edge[String]] = Array(Edge(5,3,advisor))
```

Command took 0.12 seconds -- by louhy1128@gmail.com at 2017/7/3 下午4:49:49 on My Cluster

Cmd 10

```
1 graph.edges.filter { case Edge(src, dst, prop) => src > dst }.count
```

▸ (1) Spark Jobs

```
res4: Long = 1
```

Command took 0.11 seconds -- by louhy1128@gmail.com at 2017/7/3 下午4:51:20 on My Cluster

Cmd 11

```
1 val facts: RDD[String] =
2   graph.triplets.map(triplet =>
3     triplet.srcAttr._1 + " is the " + triplet.attr + " of " +
  triplet.dstAttr._1)
4 facts.collect.foreach(println(_))
```

▸ (1) Spark Jobs

```
rxin is the collab of jgonzal
franklin is the advisor of rxin
istoica is the colleague of franklin
franklin is the pi of jgonzal
```

```
facts: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[2067] at map at <cons
ole>:51
```

Command took 0.25 seconds -- by louhy1128@gmail.com at 2017/7/3 下午4:51:23 on My Cluster

Cmd 12

```
1 val inDegrees: VertexRDD[Int] = graph.inDegrees
```

```
inDegrees: org.apache.spark.graphx.VertexRDD[Int] = VertexRDDImpl[2071] at RDD a
t VertexRDD.scala:55
```

Command took 0.15 seconds -- by louhy1128@gmail.com at 2017/7/3 下午4:51:53 on My Cluster

Cmd 13

```
1 inDegrees.collect.foreach(println)
```

▸ (1) Spark Jobs

```
(3,1)
(5,1)
(7,2)
```

Command took 0.16 seconds -- by louhy1128@gmail.com at 2017/7/3 下午4:52:12 on My Cluster

Cmd 14

```
1 val outDegrees: VertexRDD[Int] = graph.outDegrees
2
```

```
outDegrees: org.apache.spark.graphx.VertexRDD[Int] = VertexRDDImpl[2076] at RDD
 at VertexRDD.scala:55
```

Command took 0.09 seconds -- by louhy1128@gmail.com at 2017/7/3 下午4:53:56 on My Cluster

Cmd 15

```
1 outDegrees.collect.foreach(println)
```

▸ (1) Spark Jobs

```
(2,1)
(3,1)
(5,2)
```

Command took 0.16 seconds -- by louhy1128@gmail.com at 2017/7/3 下午4:53:59 on My Cluster

Cmd 16

```
1 graph.triplets.map(
2   triplet => triplet.srcAttr._1 + " is the " + triplet.attr + " of " +
  triplet.dstAttr._1
3 ).collect.foreach(println(_))
```

▸ (1) Spark Jobs

```
rxin is the collab of jgonzal
franklin is the advisor of rxin
```

```
istoica is the colleague of franklin
franklin is the pi of jgonzal
```

Command took 0.13 seconds -- by louhy1128@gmail.com at 2017/7/3 下午4:55:08 on My Cluster

Cmd 17

```
1 val validGraph = graph.subgraph(vpred = (id, attr) => attr._2 != "Missing")
```

```
validGraph: org.apache.spark.graphx.Graph[(String, String),String] = org.apache.spark.graphx.impl.GraphImpl@36601179
```

Command took 0.17 seconds -- by louhy1128@gmail.com at 2017/7/3 下午4:55:15 on My Cluster

Cmd 18

```
1 validGraph.vertices.collect.foreach(println(_))
```

▸ (1) Spark Jobs

```
(2,(istoica,prof))
(3,(rxin,student))
(5,(franklin,prof))
(7,(jgonzal,postdoc))
```

Command took 0.13 seconds -- by louhy1128@gmail.com at 2017/7/3 下午4:55:27 on My Cluster

Cmd 19

```
1 validGraph.triplets.map(
2   triplet => triplet.srcAttr._1 + " is the " + triplet.attr + " of " + triplet.dstAttr._1
3 ).collect.foreach(println(_))
```

▸ (1) Spark Jobs

```
rxin is the collab of jgonzal
franklin is the advisor of rxin
istoica is the colleague of franklin
franklin is the pi of jgonzal
```

Command took 0.15 seconds -- by louhy1128@gmail.com at 2017/7/3 下午4:55:40 on My Cluster

Cmd 20

```
1 import org.apache.spark.graphx.GraphLoader
2 import sys.process._
3 "wget -P /tmp http://www.utdallas.edu/~axn112530/cs6350/data/followers.txt" !!
4
5 import sys.process._
6 "wget -P /tmp http://www.utdallas.edu/~axn112530/cs6350/data/users.txt" !!
```

```
--2017-07-03 22:00:34--  http://www.utdallas.edu/~axn112530/cs6350/data/followers.txt
Resolving www.utdallas.edu (www.utdallas.edu)... 104.16.43.54, 104.16.44.54, 2400:cb00:2048:1::6810:2b36, ...
```

```
Connecting to www.utdallas.edu (www.utdallas.edu)|104.16.43.54|:80... connected.
HTTP request sent, awaiting response... 200 OK
Length: 32 [text/plain]
Saving to: ‘/tmp/followers.txt’

     0K                                                       100% 6.66M=0s

2017-07-03 22:00:34 (6.66 MB/s) - ‘/tmp/followers.txt’ saved [32/32]

--2017-07-03 22:00:34--  http://www.utdallas.edu/~axn112530/cs6350/data/users.tx
t
Resolving www.utdallas.edu (www.utdallas.edu)... 104.16.44.54, 104.16.43.54, 240
0:cb00:2048:1::6810:2c36, ...
Connecting to www.utdallas.edu (www.utdallas.edu)|104.16.44.54|:80... connected.
HTTP request sent, awaiting response... 200 OK
Length: 169 [text/plain]
Saving to: ‘/tmp/users.txt’

     0K                                                       100% 35.0M=0s

2017-07-03 22:00:35 (35.0 MB/s) - ‘/tmp/users.txt’ saved [169/169]

warning: there were 2 feature warning(s); re-run with -feature for details
import org.apache.spark.graphx.GraphLoader
import sys.process._
import sys.process._
res15: String = ""
```

Command took 0.36 seconds -- by louhy1128@gmail.com at 2017/7/3 下午5:00:34 on My Cluster

Cmd 21

```
1 val graph = GraphLoader.edgeListFile(sc, "file:/tmp/followers.txt")
```

▸ (1) Spark Jobs

```
graph: org.apache.spark.graphx.Graph[Int,Int] = org.apache.spark.graphx.impl.Gra
phImpl@2ef8ef1e
```

Command took 0.13 seconds -- by louhy1128@gmail.com at 2017/7/3 下午5:03:43 on My Cluster

Cmd 22

```
1 graph.inDegrees.collect()
```

▸ (1) Spark Jobs

```
res19: Array[(org.apache.spark.graphx.VertexId, Int)] = Array((6,1), (2,1), (1,
2), (3,2), (7,2))
```

Command took 0.14 seconds -- by louhy1128@gmail.com at 2017/7/3 下午5:03:46 on My Cluster

Cmd 23

```
graph.outDegrees.collect()
```

▸ (1) Spark Jobs

```
res20: Array[(org.apache.spark.graphx.VertexId, Int)] = Array((4,1), (6,2), (2,
1), (1,1), (3,1), (7,2))
```

Command took 0.13 seconds -- by louhy1128@gmail.com at 2017/7/3 下午5:03:49 on My Cluster

Cmd 24

```
val ranks = graph.pageRank(0.0001).vertices
```

▸ (1) Spark Jobs

```
ranks: org.apache.spark.graphx.VertexRDD[Double] = VertexRDDImpl[3075] at RDD at
VertexRDD.scala:55
```

Command took 7.01 seconds -- by louhy1128@gmail.com at 2017/7/3 下午5:03:51 on My Cluster

Cmd 25

```
ranks.collect()
```

▸ (1) Spark Jobs

```
res21: Array[(org.apache.spark.graphx.VertexId, Double)] = Array((4,0.15), (6,0.
7013599933629602), (2,1.390049198216498), (1,1.4588814096664682), (3,0.999344203
8507723), (7,1.2973176314422592))
```

Command took 0.20 seconds -- by louhy1128@gmail.com at 2017/7/3 下午5:04:00 on My Cluster

Cmd 26

```
val users = sc.textFile("file:/tmp/users.txt").map { line =>
  val fields = line.split(",")
  (fields(0).toLong, fields(1))
}
```

```
users: org.apache.spark.rdd.RDD[(Long, String)] = MapPartitionsRDD[3080] at map
 at <console>:51
```

Command took 0.14 seconds -- by louhy1128@gmail.com at 2017/7/3 下午5:04:22 on My Cluster

Cmd 27

```
val ranksByUsername = users.join(ranks).map {
  case (id, (username, rank)) => (username, rank)
}
// Print the result
println(ranksByUsername.collect().mkString("\n"))
```

▸ (1) Spark Jobs

```
(justinbieber,0.15)
(matei_zaharia,0.7013599933629602)
(ladygaga,1.390049198216498)
```

```
(BarackObama,1.4588814096664682)
(jeresig,0.9993442038507723)
(odersky,1.2973176314422592)
ranksByUsername: org.apache.spark.rdd.RDD[(String, Double)] = MapPartitionsRDD[3
084] at map at <console>:57
```

Command took 0.30 seconds -- by louhy1128@gmail.com at 2017/7/3 下午5:04:30 on My Cluster

Cmd 28

```
1 //above is basic pagerank
```

Command took 0.09 seconds -- by louhy1128@gmail.com at 2017/7/3 下午5:04:54 on My Cluster

Cmd 29

```
1 // output pagerank with usernames
2 import org.apache.spark.graphx.GraphLoader
```

```
import org.apache.spark.graphx.GraphLoader
```

Command took 0.08 seconds -- by louhy1128@gmail.com at 2017/7/3 下午5:06:18 on My Cluster

Cmd 30

```
1 val users = (sc.textFile("file:/tmp/users.txt")
2   .map(line => line.split(",")).map( parts => (parts.head.toLong,
  parts.tail) ))
```

```
users: org.apache.spark.rdd.RDD[(Long, Array[String])] = MapPartitionsRDD[3088]
 at map at <console>:53
```

Command took 0.17 seconds -- by louhy1128@gmail.com at 2017/7/3 下午5:06:28 on My Cluster

Cmd 31

```
1 val followerGraph = GraphLoader.edgeListFile(sc, "file:/tmp/followers.txt")
```

▸ (1) Spark Jobs

```
followerGraph: org.apache.spark.graphx.Graph[Int,Int] = org.apache.spark.graphx.
impl.GraphImpl@36c0ef3d
```

Command took 0.12 seconds -- by louhy1128@gmail.com at 2017/7/3 下午5:06:36 on My Cluster

Cmd 32

```
1 // Attach the user attributes
2 val graph = followerGraph.outerJoinVertices(users) {
3   case (uid, deg, Some(attrList)) => attrList
4   // Some users may not have attributes so we set them as empty
5   case (uid, deg, None) => Array.empty[String]
6 }
```

```
graph: org.apache.spark.graphx.Graph[Array[String],Int] = org.apache.spark.graph
x.impl.GraphImpl@1ff7b5de
```

```
Command took 0.13 seconds -- by louhy1128@gmail.com at 2017/7/3 下午5:07:55 on My Cluster
```

Cmd 33

```
1 // Restrict the graph to users with usernames and names
2 val subgraph = graph.subgraph(vpred = (vid, attr) => attr.size == 2)
```

```
subgraph: org.apache.spark.graphx.Graph[Array[String],Int] = org.apache.spark.graphx.impl.GraphImpl@d26c51b
Command took 0.12 seconds -- by louhy1128@gmail.com at 2017/7/3 下午5:08:07 on My Cluster
```

Cmd 34

```
1 val pagerankGraph = subgraph.pageRank(0.001)
```

▸ (1) Spark Jobs

```
pagerankGraph: org.apache.spark.graphx.Graph[Double,Double] = org.apache.spark.graphx.impl.GraphImpl@6f231f75
Command took 4.04 seconds -- by louhy1128@gmail.com at 2017/7/3 下午5:09:19 on My Cluster
```

Cmd 35

```
userInfoWithPageRank: org.apache.spark.graphx.Graph[(Double, List[String]),Int] = org.apache.spark.graphx.impl.GraphImpl@2f54579
Command took 0.30 seconds -- by louhy1128@gmail.com at 2017/7/3 下午5:09:27 on My Cluster
```