Technical Report Of FastCD in Spark GraphX

We implement FastCD in Spark Graphx-1.5.0. The details of implementation are as follows:

1. How to build a Spark cluster.

This part introduces steps to build a Spark cluster.

2.Run FastCD in Spark Graphx.

This part introduces steps to use FastCD.

3.Build FastCD Source in Eclipse.

Basic Environment Description

OS:	Ubuntu 14.04
JAVA version:	Jdk 1.7
Hadoop version :	2.6.0
Scala version:	2.10
Spark version:	1.5.0

Part 1 : Build a Spark cluster

1) Configure SSH to login each slave without password on master.

- ssh-keygen -t dsa -P '' -f ~/.ssh/id_dsa
- cat ~/.ssh/id_dsa.pub >>~/.ssh/authorized_keys
- scp ~/.ssh/authorized_keys each slave:~/.ssh/

2) Download Hadoop 2.6.0 (http://www.apache.org/dyn/closer.cgi/hadoop/common/hadoop-2.6.0/hadoop-2.6.0.tar.gz) and install it.

- Change file core-site.xml . Add properties:fs.default.name and hadoop.tmp.dir .
- Change file hdfs-site.xml . Add properties:dfs.namenode.secondary.http-address , dfs.namenode.name.dir , dfs.datanode.data.dir, dfs.replication and dfs.webhdfs.enabled.
- Add the hostname of each slave to file slaves.
- Execute hadoop namenode -format and start-dfs.sh .
- As the follows in Figure 1, execute hdfs -put somefile / and hdfs dfs -ls / to make sure it's in there.

```
root@master:~# hdfs dfs -put test.txt /
root@master:~# hdfs dfs -ls /

Found 1 items
-rw-r--r--   3 root supergroup         39 2016-07-22 21:15 /test.txt
root@master:~#
root@master:~#
```

Figure 1:Test for HDFS

3) Download Spark 1.5.0 (http://spark.apache.org/downloads.html) and install it.

- Add the hostname of each slave to file slaves..
- Change file spark-env.sh .Add the follows contents:

```
export SCALC_HOME=/root/scala
export JAVA_HOME=/root/java
export SPARK_LOCAL_DIRS=/spark/spark-1.2.0-bin-hadoop2.4/tmp
export SPARK_MASTER_IP=166.111.141.3
export SPARK_MASTER_PORT=8070
export SPARK_MASTER_WEBUI_PORT=8090
export SPARK_WORKER_PORT=8092
export SPARK_WORKER_MEMORY=4G
export SPARK_WORKER_CORES=4
```

- Execute $SPARK_HOME/sbin/start-all.sh .

Part 2 : Run FastCD in Spark Graphx

Run command:

```
./bin/spark-submit \
  --class    org.spark.graphx.test.FastCDTest\
  --master spark://<MasterIP>:<Port> \
  fastCD.jar \
  <edgesFile> <verticeFile> <partition>
```

The edgesFile records the message of edges ,and the verticeFile records the message of the vertices.The partition denote the number of partition in the Spark Graphx.

Note:All files must save in HDFS at first.

Part 3 : Build FastCD Source in Eclipse

1)Download and unzip FastCD source code.

2)Import Source code and configure the project.

"File -> Import->Existing Projects into Workspace".

Add the spark-assembly-*.jar for the project. Don't forget remove scala in spark-assembly-*.jar."

Build Path -> Add External Archives".