

User Documentation

CMPUT 391 BigData Project Group 4
Xiaohui Ma, Shenshu Zhou, Yue Yin
April 22, 2016

1. Requirements (on remote)

- a. Cassandra Version 2.2.5
- b. Python 2
- c. gcc

2. Setup Environment

a. Local

- i. Download source code:

git clone <https://github.com/C391BigDataT4/CMPUT391Project.git>

- ii. Set up SSH tunnel. Add the following to ".ssh/conf":

Host 10.2.3.31

ProxyCommand ssh -i 391winter.pem -q ubuntu@162.246.157.170 -W %h:%p

Host 10.2.3.30

ProxyCommand ssh -i 391winter.pem -q ubuntu@162.246.157.170 -W %h:%p

Host 10.2.3.5

ProxyCommand ssh -i 391winter.pem -q ubuntu@162.246.157.170 -W %h:%p

b. Remote

- i. Download Cassandra and untar the file

www.apache.org/dyn/closer.lua/cassandra/2.2.5/apache-cassandra-2.2.5-bin.tar.gz

- ii. Add Cassandra path to ".bashrc":

export CASSANDRA_HOME=/home/ubuntu/cassandra/apache-cassandra

export PATH=\$PATH:\$CASSANDRA_HOME/bin

3. Initialization

- a. Stop remote server:

sh sh/stop_remote_server.sh

- b. Upload necessary files and configuration to clusters:

sh sh/initialize.sh

- c. Start remote server:

sh sh/start_remote_server.sh

- d. Create table:

sh sh/create_table.sh

- e. Check nodes status:

sh sh/check_server_status.sh

4. Generate Random Data

- a. SSH into Node 1 (10.2.3.31):
`ssh -i 391winter.pem ubuntu@10.2.3.31`
- b. Run generator with one argument - number of days. One day will have 10,000 rows (about 50MB in size). Output folder is /mystorage/
In the ssh window, run command:
`~/generate 20 &` to generate 200,000 rows of data (about 1GB)

5. Insert Data

- a. Run the Python code on Node 1 (10.2.3.31) for inserting data, followed by the csv file to insert:
`python insert.py /mystorage/data.csv &`

6. Test Queries

- a. Run the Python code on Node 1 (10.2.3.31) to test the queries:
In the ssh window, run `python query.py`
- b. The output will look like this:
Query 1 finished in 4.21734189987 seconds.
Query 2 finished in 0.920250892639 seconds.
Query 3 finished in 1.88912391663 seconds.
Query 4 finished in 7.66732597351 seconds.
Query 5 finished in 3.91390299797 seconds.