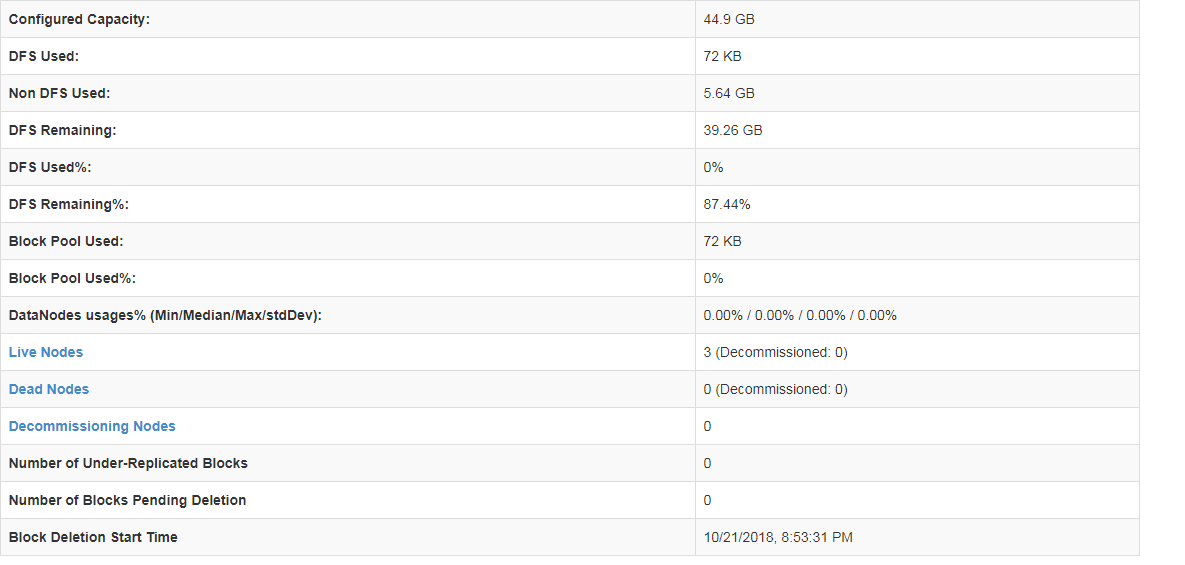
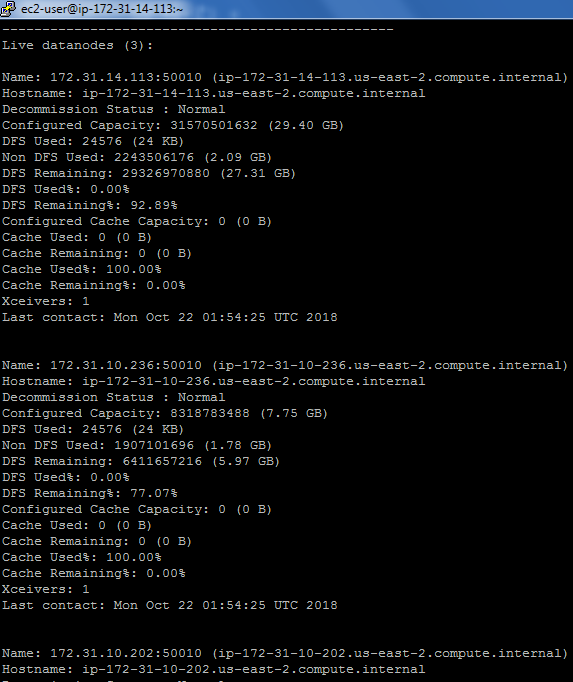
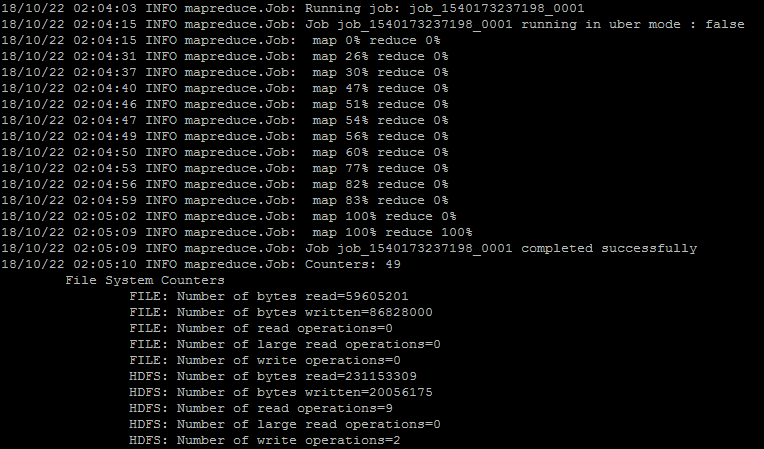
William Chirciu  
CSC 555 Project Phase 1

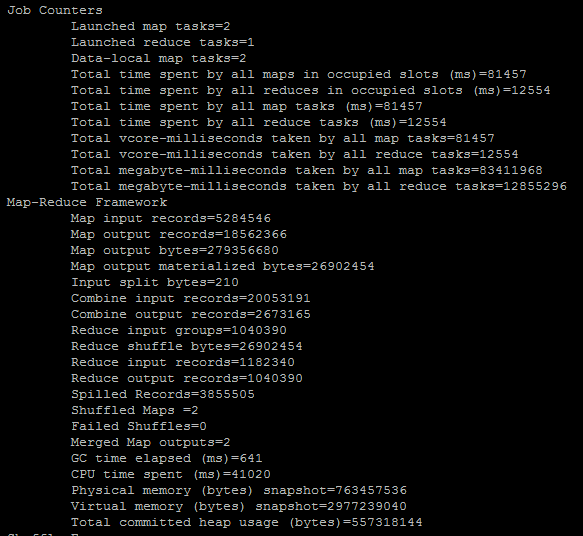
Cluster Status:

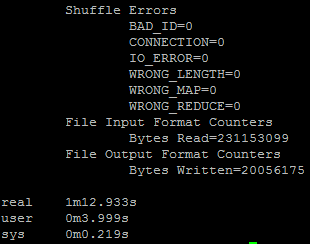




Running word count on bioproject.xml with our node cluster



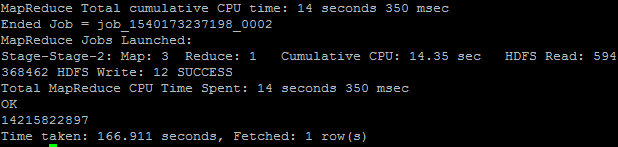


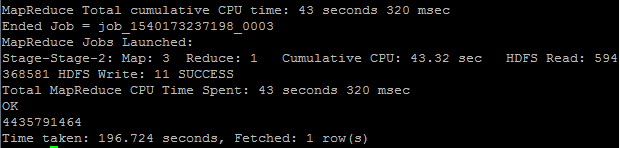


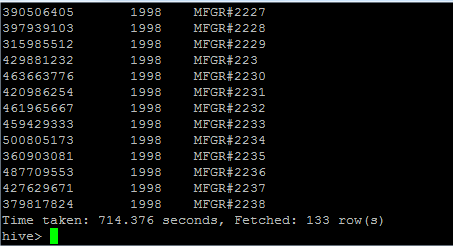
This job took 1 minute and 13 seconds with 2 map tasks and 1 reduce task.



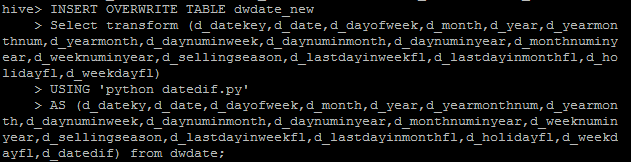
**Part 2: Hive**

Query 1.2  


Query 1.3  


Query 2.2  


HIVE COMMAND FOR TRANSFORMING dwdate table



ASSOCIATED PYTHON CODE:

import sys

from datetime import date

for line in sys.stdin:

line = line.strip()

vals = line.split('\t')

#vals[8] = day number in month

#vals[10] = month num in year

#vals[4] = year

d1 = date(1994,11,1)

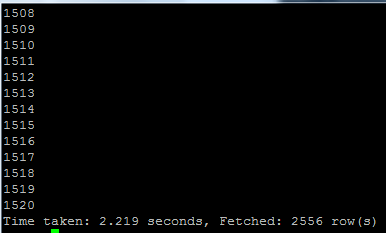
d2 = date(int(vals[4]),int(vals[10]),int(vals[8]))

diff = str((d2-d1).days)

vals.append(diff)

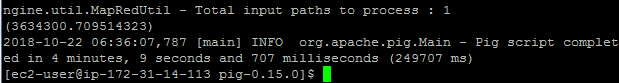
print '\t'.join(vals)

Some output of d\_datediff column: used SELECT d\_datediff from dwdate\_new

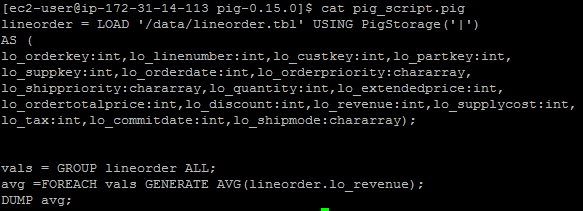


**Part 3: PIG**

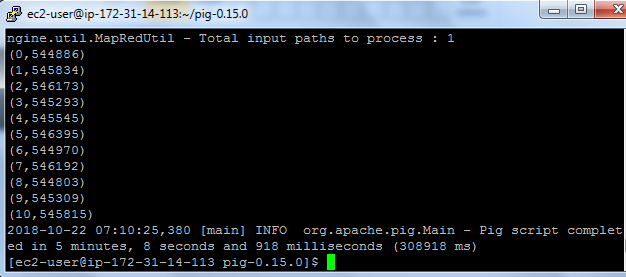
**Query 0.1**



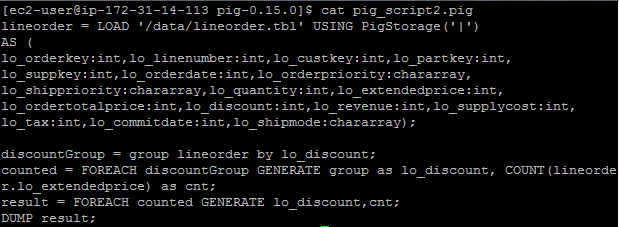
**Pig commands:**

****

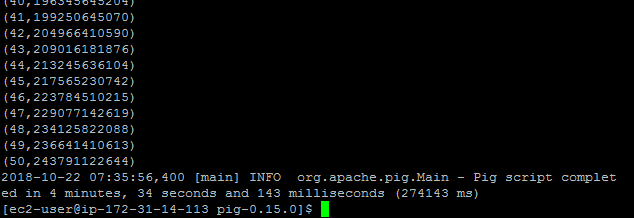
**Query 0.2**

****

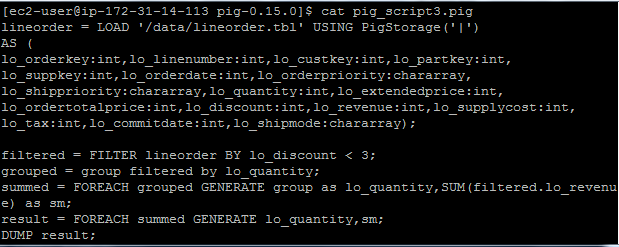
**Pig Commands:**

****

**Query 0.3**

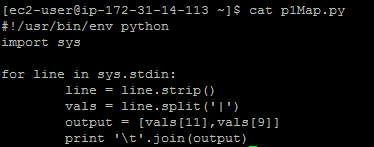
****

**Pig Commands:**

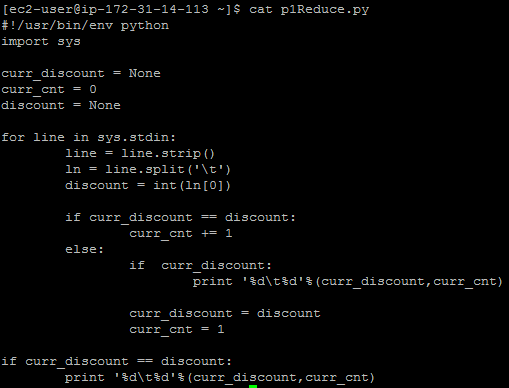
****

**Part 4: Hadoop Streaming**

**Mapper:**

****

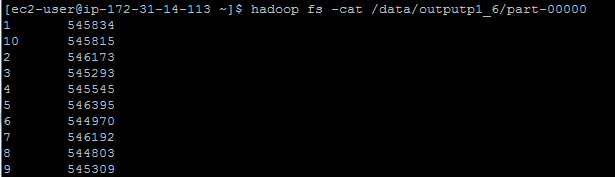
**Reducer:**

****

**Hadoop Streaming Command:**

****

**Output:**

****