



SAN JOSÉ STATE UNIVERSITY

GROUP PROJECT

CMPE 274– Business Intelligence

(Spring Semester, 2014)

Project Name: Music Data Analysis

Submitted To: Dr. Weider Yu

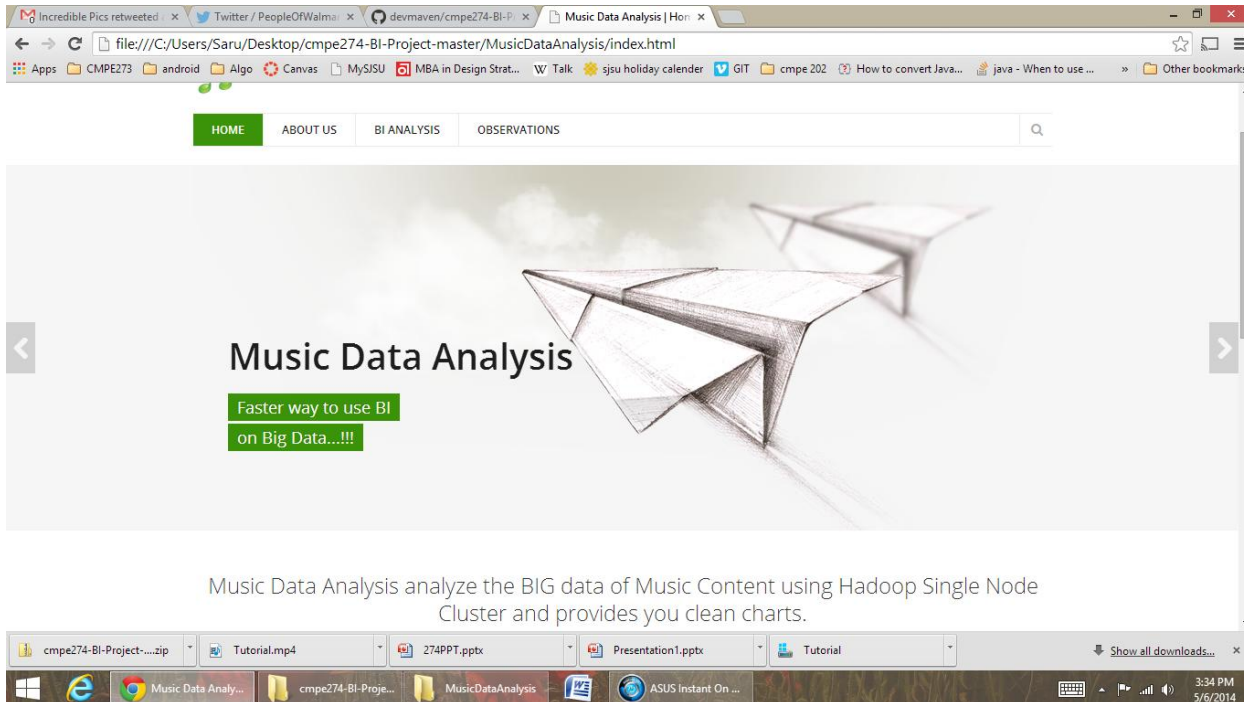
Date of Submission: 05/6/2014

Submitted By: Team 9

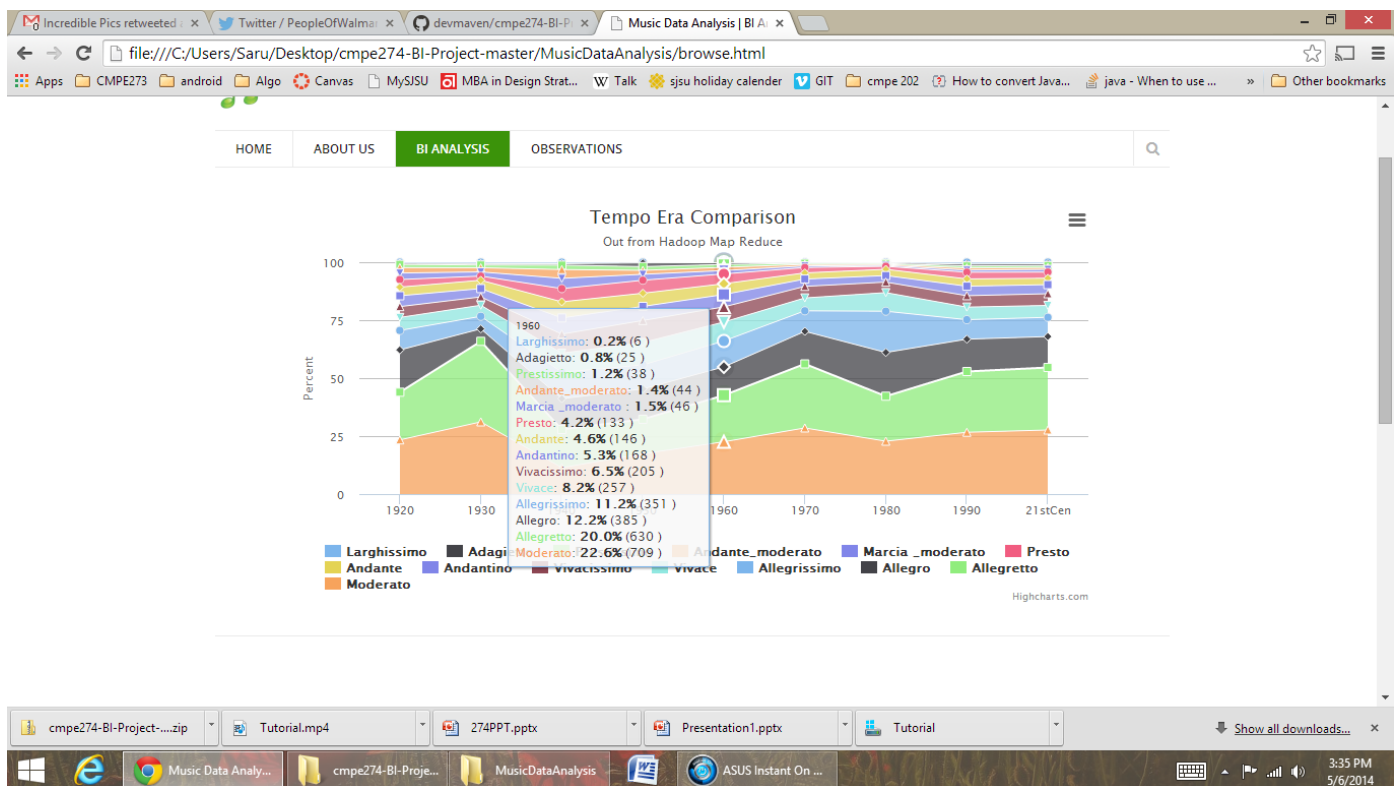
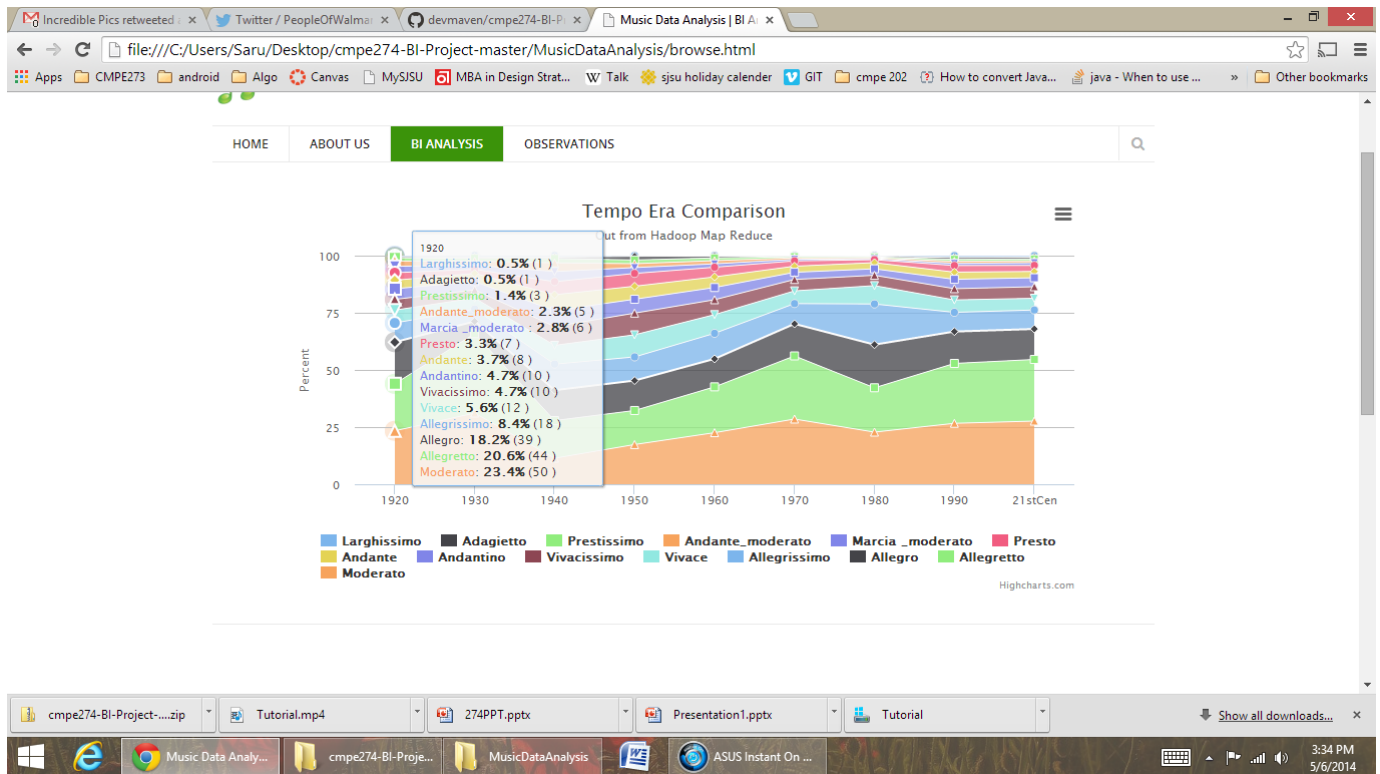
S.No	Project Member	SJSU ID	Email ID	Phone No
1	Sarvagya Jain	009269377	sarvagya.jain@sjsu.edu	408-326-9743
2	Swapnil Pancholi	009303684	pancholi.swapnil@gmail.com	408-831-8162
3	Arun Malik	009304607	malik.mgm@gmail.com	408-834-6060
4	Pooja Kasu	009305829	pooja.kasu@gmail.com	479-200-1056
5	Gaurav Kesarwani	009278815	gaurav.kesarwani2@gmail.com	408-442-8559

Snapshots of Demo

The website developed to represent our data

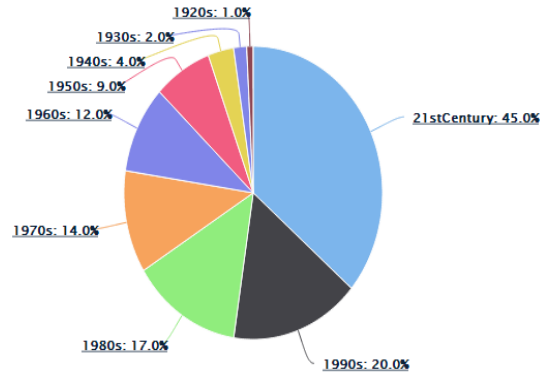


BI Analysis (Dynamic Graphs)



TempoEra Comparison

Click the slices to view tempo details




Highcharts.com

Observations


Kind	% Complete	Num Tasks	Pending	Running	Complete	Killed	Failed/Killed Task Attempts
map	100.00%	2	0	0	2	0	0 / 0
reduce	100.00%	1	0	0	1	0	0 / 0

Observations

Map Completion Graph - [close](#)



Reduce Completion Graph - [close](#)



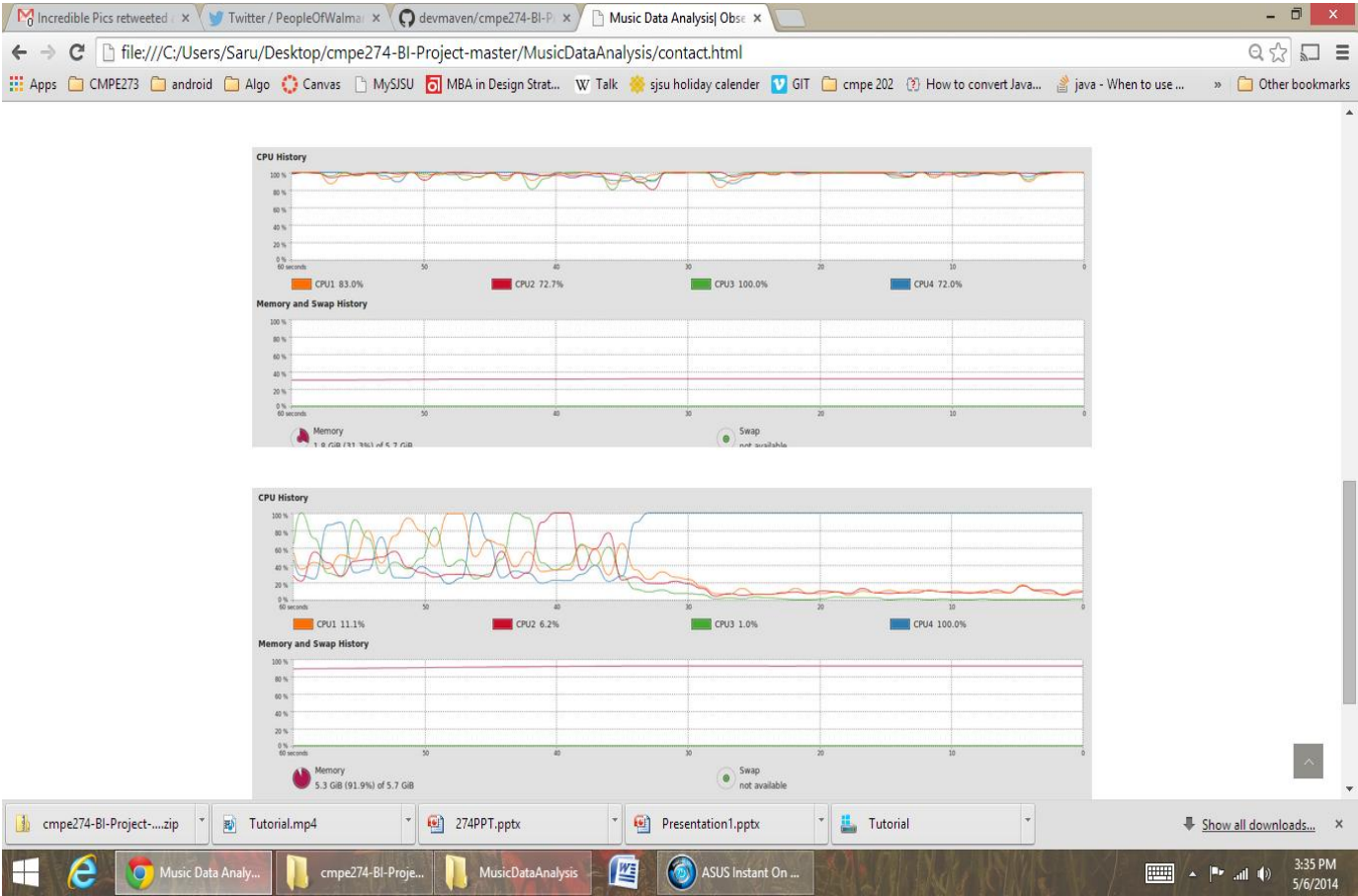
cmpe274-BI-Project-....zip Tutorial.mp4 274PPT.pptx Presentation1.pptx Tutorial

Show all downloads...

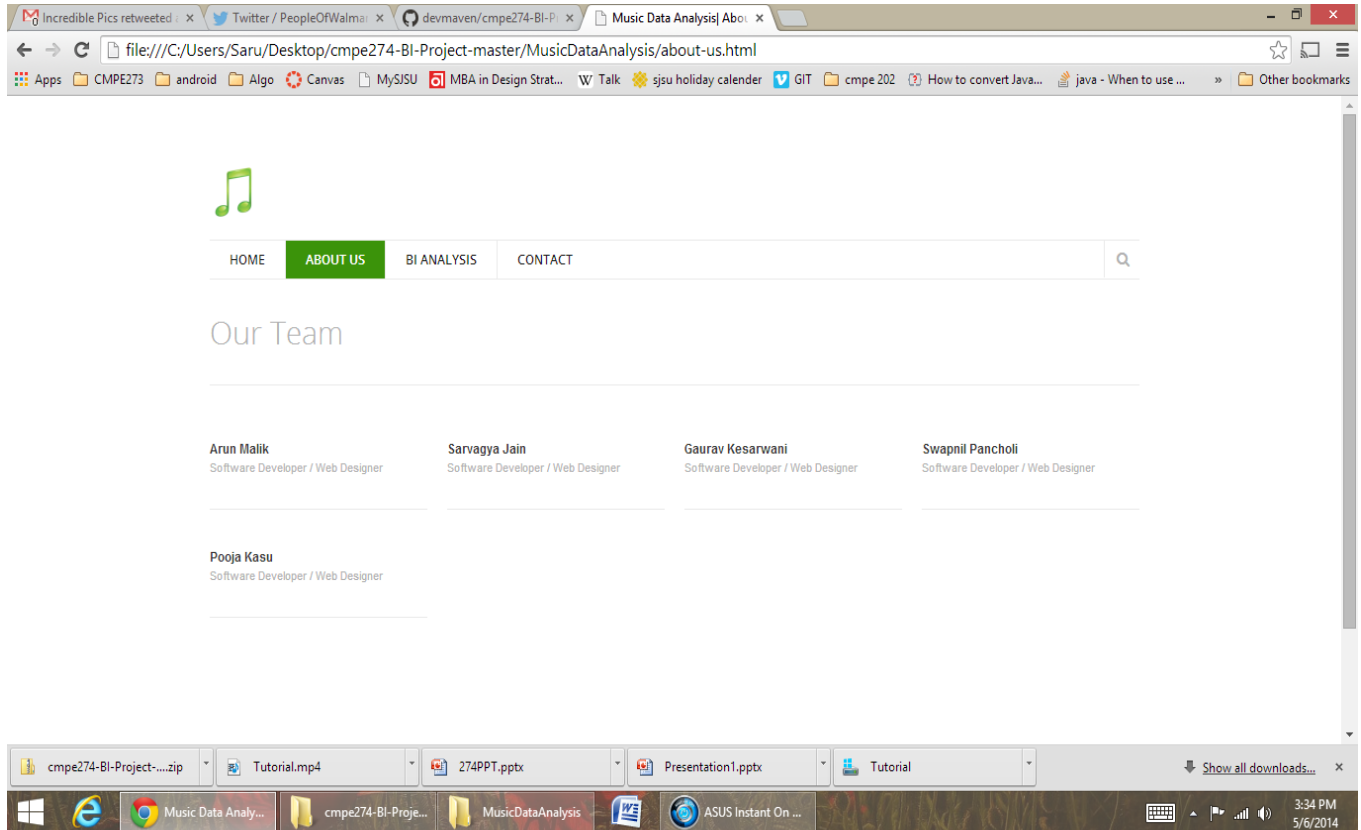
Music Data Analy... cmpe274-BI-Proje... MusicDataAnalysis ASUS Instant On ...

3:35 PM 5/6/2014

Observation : CPU Utilization



About Us



Map Reduce Scripts

Mapper

```
*C:\Users\Saru\Downloads\mapper.py - Notepad++

File Edit Search View Encoding Language Settings Macro Run Plugins Window ?

text_hear.c Locations.xml new.py mapper.py reducer.py

1  #!/usr/bin/env python
2  #The script reads each track from the input file, uses bubble sort to find out its tempo category and maps it to its respective category
3  import sys
4
5  def read_input(file):
6      for line in file:
7          # split the line into words
8          track, tempo, loudness = line.split()
9          yield track, tempo
10
11 #bubble sort
12 def sort_bubble(tempo):
13     tempoRange = [19, 40, 45, 50, 55, 65, 69, 72, 77, 83, 85, 97, 109, 132, 140, 150, 167, 177]
14     tempocategory = ['Larghissimo', 'Grave', 'Lento', 'Largo', 'Larghetto', 'Adagio', 'Adagietto', 'Andante_moderato', 'Andante', 'Andantino', 'Marcia_moderato', 'Moderato', 'Allegretto',
15                     'Allegro', 'Vivace', 'Vivacissimo', 'Allegro_smo', 'Presto', 'Prestissimo']
16     min_index=0
17     max_index=len(tempoRange)-1
18     while(min_index<max_index):
19         #determining midvalue
20         midindex=(min_index+max_index)/2
21         #comparison
22         if(tempoRange[midindex]==tempo):
23             return tempocategory[midindex]
24         elif(tempoRange[midindex] > tempo):
25             max_index=midindex-1
26         elif(tempoRange[midindex]<tempo):
27             min_index=midindex+1
28     return tempocategory[max_index+1]
29
30 def main():
31     # input comes from STDIN (standard input) and then prints the values of trackid and the category of track's tempo
32     data = read_input(sys.stdin)
33     for trackid, tempo in data:
34         tempo = float(tempo)
35         category = sort_bubble(tempo)
36         print "%s,%s" %(category, trackid)
37
38 if __name__ == '__main__':
```


Reducer

```
*C:\Users\Saru\Downloads\reducer.py - Notepad++
File Edit Search View Encoding Language Settings Macro Run Plugins Window ?
text_heart.c Locations.xml new.py mapper.py reducer.py
8 eracategorytrack = {}
9 # Create a Dictionary with artistId as the key and list of artist ratings as the values
10 for line in sys.stdin:
11     line = line.strip()
12     category, trackid = line.split(",")
13     trackrepository[trackid] = category
14 #Import Era, category data from track_metadata using sqlite3
15 try:
16     import sqlite3
17 except ImportError:
18     print 'you need sqlite3 installed to use this program'
19     sys.exit(0)
20 dbfile = '/home/swap/Downloads/track_metadata.db'
21 conn= sqlite3.connect(dbfile)
22 c=conn.cursor()
23 q = "select era, track_id from songs where era!=0 order by era"
24 c.execute(q)
25 #data = c.fetchone()
26 rows = c.fetchall()
27 c.close()
28 conn.close()
29 # Make era and category as a composite key to get the Trackids for that particular Era in a particular category
30 for row in rows:
31     #print "%s,%s" %(row[1], row[0])
32     key = row[0] + "," + trackrepository[row[1]]
33     if(eracategorytrack.has_key(key)):
34         eracategorytrack[key].append(row[1])
35     else:
36         eracategorytrack[key] = [row[1]]
37 #print era , category and the No of tracks in the particular era and category
38 for key in eracategorytrack:
39     era, category = key.split(',')
40     trackidsample = eracategorytrack[key]
41     trackids = str(trackidsample)
42     trackids = trackids.strip('[]')
43     trackids = trackids.replace('u'', '')
44     trackids = trackids.replace('\\', '')
45     print "%s\t%s\t%s" %(era, category, len(eracategorytrack[key]))
```

Python file length: 1484 lines: 48 Ln: 13 Col: 38 Sel: 0|0 UNIX ANSI as UTF-8 INS

Windows taskbar: 4:58 PM 5/6/2014

Snapshots of running Map reduce Scripts on Hadoop

Starting Hadoop Services

```
Activities | Terminal | Tue May 6, 13:02
hduser@swap-Dell-System-XPS-L502X: ~/hadoop/bin

hduser@swap-Dell-System-XPS-L502X: ~/hadoop/bin$ cd bin
hduser@swap-Dell-System-XPS-L502X: ~/hadoop/bin$ ./stop-all.sh
stopping jobtracker
localhost: stopping tasktracker
stopping namenode
localhost: stopping datanode
localhost: stopping secondarynamenode
hduser@swap-Dell-System-XPS-L502X: ~/hadoop/bin$ ./start-all.sh
starting namenode, logging to /home/hduser/hadoop/libexec/../logs/hadoop-hduser-namenode-swap-Dell-System-XPS-L502X.out
localhost: starting datanode, logging to /home/hduser/hadoop/libexec/../logs/hadoop-hduser-datanode-swap-Dell-System-XPS-L502X.out
localhost: starting secondarynamenode, logging to /home/hduser/hadoop/libexec/../logs/hadoop-hduser-secondarynamenode-swap-Dell-System-XPS-L502X.out
starting jobtracker, logging to /home/hduser/hadoop/libexec/../logs/hadoop-hduser-jobtracker-swap-Dell-System-XPS-L502X.out
localhost: starting tasktracker, logging to /home/hduser/hadoop/libexec/../logs/hadoop-hduser-tasktracker-swap-Dell-System-XPS-L502X.out
hduser@swap-Dell-System-XPS-L502X: ~/hadoop/bin$
```

(JPS) Check Hadoop Service Status

```
Activities | Terminal v Tue May 6, 13:02
hduser@swap-Dell-System-XPS-L502X: ~/hadoop/bin
File Edit View Search Terminal Tabs Help
hduser@swap-Dell-System-XPS-L502X: ~/hadoop$ cd bin
hduser@swap-Dell-System-XPS-L502X: ~/hadoop/bin$ ./stop-all.sh
stopping jobtracker
localhost: stopping tasktracker
stopping namenode
localhost: stopping datanode
localhost: stopping secondarynamenode
hduser@swap-Dell-System-XPS-L502X: ~/hadoop/bin$ ./start-all.sh
starting namenode, logging to /home/hduser/hadoop/libexec/../logs/hadoop-hduser-namenode-swap-Dell-System-XPS-L502X.out
localhost: starting datanode, logging to /home/hduser/hadoop/libexec/../logs/hadoop-hduser-datanode-swap-Dell-System-XPS-L502X.out
localhost: starting secondarynamenode, logging to /home/hduser/hadoop/libexec/../logs/hadoop-hduser-secondarynamenode-swap-Dell-System-XPS-L502X.out
starting jobtracker, logging to /home/hduser/hadoop/libexec/../logs/hadoop-hduser-jobtracker-swap-Dell-System-XPS-L502X.out
localhost: starting tasktracker, logging to /home/hduser/hadoop/libexec/../logs/hadoop-hduser-tasktracker-swap-Dell-System-XPS-L502X.out
hduser@swap-Dell-System-XPS-L502X: ~/hadoop/bin$ jps
14289 JobTracker
14523 Jps
14444 TaskTracker
13879 NameNode
14031 DataNode
14196 SecondaryNameNode
hduser@swap-Dell-System-XPS-L502X: ~/hadoop/bin$
```

Execute Map-Reduce Scripts

```
Activities Terminal v Tue May 6, 13:02
hduser@swap-Dell-System-XPS-L502X: ~/hadoop
File Edit View Search Terminal Tabs Help
hduser@swap-Dell-System-XPS-L502X: ~/hadoop$ cd bin
hduser@swap-Dell-System-XPS-L502X: ~/hadoop/bin$ ./stop-all.sh
stopping jobtracker
localhost: stopping tasktracker
stopping namenode
localhost: stopping datanode
localhost: stopping secondarynamenode
hduser@swap-Dell-System-XPS-L502X: ~/hadoop/bin$ ./start-all.sh
starting namenode, logging to /home/hduser/hadoop/libexec/../logs/hadoop-hduser-namenode-swap-Dell-System-XPS-L502X.out
localhost: starting datanode, logging to /home/hduser/hadoop/libexec/../logs/hadoop-hduser-datanode-swap-Dell-System-XPS-L502X.out
localhost: starting secondarynamenode, logging to /home/hduser/hadoop/libexec/../logs/hadoop-hduser-secondarynamenode-swap-Dell-System-XPS-L502X.out
starting jobtracker, logging to /home/hduser/hadoop/libexec/../logs/hadoop-hduser-jobtracker-swap-Dell-System-XPS-L502X.out
localhost: starting tasktracker, logging to /home/hduser/hadoop/libexec/../logs/hadoop-hduser-tasktracker-swap-Dell-System-XPS-L502X.out
hduser@swap-Dell-System-XPS-L502X: ~/hadoop/bin$ jps
14289 JobTracker
14523 Jps
14444 TaskTracker
13879 NameNode
14031 DataNode
14196 SecondaryNameNode
hduser@swap-Dell-System-XPS-L502X: ~/hadoop/bin$ cd ..
hduser@swap-Dell-System-XPS-L502X: ~/hadoop$ bin/hadoop jar contrib/streaming/hadoop-streaming-1.2.1.jar \-file /home/hduser/mapper.py \-mapper /home/hduser/mapper.py \-file /home/hduser/reducer.py \-reducer /home/hduser/reducer.py \-input /user/hduser/gutenberg/Tenpofull.txt \-output /user/hduser/output/04
```

```
hduser@swap-Dell-System-XPS-L502X: ~/hadoop$ ./start-all.sh
stopping jobtracker
localhost: stopping tasktracker
stopping namenode
localhost: stopping datanode
localhost: stopping secondarynamenode
hduser@swap-Dell-System-XPS-L502X:~/hadoop$ ./start-all.sh
starting namenode, logging to /home/hduser/hadoop/libexec/../logs/hadoop-hduser-namenode-swap-Dell-System-XPS-L502X.out
localhost: starting datanode, logging to /home/hduser/hadoop/libexec/../logs/hadoop-hduser-datanode-swap-Dell-System-XPS-L502X.out
localhost: starting secondarynamenode, logging to /home/hduser/hadoop/libexec/../logs/hadoop-hduser-secondarynamenode-swap-Dell-System-XPS-L502X.out
starting jobtracker, logging to /home/hduser/hadoop/libexec/../logs/hadoop-hduser-jobtracker-swap-Dell-System-XPS-L502X.out
localhost: starting tasktracker, logging to /home/hduser/hadoop/libexec/../logs/hadoop-hduser-tasktracker-swap-Dell-System-XPS-L502X.out
hduser@swap-Dell-System-XPS-L502X:~/hadoop$ jps
14289 JobTracker
14523 Jps
14444 TaskTracker
13879 NameNode
14031 DataNode
14196 SecondaryNameNode
hduser@swap-Dell-System-XPS-L502X:~/hadoop$ cd ..
hduser@swap-Dell-System-XPS-L502X:~/hadoop$ bin/hadoop jar contrib/streaming/hadoop-streaming-1.2.1.jar \-file /home/hduser/mapper.py \-mapper /home/hduser/mapper.py \-file /home/hduser/reducer.py \-reducer /home/hduser/reducer.py \-input /user/hduser/gutenberg/Tenpofull.txt \-output /user/hduser/output/04
packageJobJar: [/home/hduser/mapper.py, /home/hduser/reducer.py, /home/hduser/hadoop/tnp/hadoop-unjar7717210963780518093/] [] /tnp/streamjob6568796048
838685046.jar tmpDir=null
14/05/06 13:02:54 INFO util.NativeCodeLoader: Loaded the native-hadoop library
14/05/06 13:02:54 WARN snappy.LoadSnappy: Snappy native library not loaded
14/05/06 13:02:54 INFO mapred.FileInputFormat: Total input paths to process : 1
14/05/06 13:02:55 INFO streaming.StreamJob: getLocalDirs(): [/home/hduser/hadoop/tnp/napred/local]
14/05/06 13:02:55 INFO streaming.StreamJob: Running job: job_201405061302_0001
14/05/06 13:02:55 INFO streaming.StreamJob: To kill this job, run:
14/05/06 13:02:55 INFO streaming.StreamJob: /home/hduser/hadoop/libexec/../bin/hadoop job -Dmapred.job.tracker=localhost:54311 -kill job_201405061302_0001
14/05/06 13:02:55 INFO streaming.StreamJob: Tracking URL: http://localhost:50030/jobdetails.jsp?jobId=job_201405061302_0001
14/05/06 13:02:56 INFO streaming.StreamJob: map 0% reduce 0%
```

HDFS Job Status

Activities google-chrome-stable v Tue May 6, 13:03

localhost:8080/ Inbox (1) - pan download can TechSmith | Cr video master Goad video n Simple Screen O Ope! Google Hadoop job...

localhost:50030/jobdetails.jsp?jobid=job_201405061302_0001

Apps Android B CHPE274 CHPE275 CHPE277 Linux DS Sem1 SJSU Live LiveCr Watch Game of Th... movies

	Launched map tasks	0	0	2
	Data-local map tasks	0	0	2
	SLOTS_MILLIS_REDUCES	0	0	15,277
File Output Format Counters	Bytes Written	0	0	3,353
FileSystemCounters	FILE_BYTES_READ	31,755,127	31,755,109	63,510,236
	HDFS_BYTES_READ	33,904,036	0	33,904,036
	FILE_BYTES_WRITTEN	63,629,318	31,814,529	95,443,847
	HDFS_BYTES_WRITTEN	0	3,353	3,353
Map-Reduce Framework	Map output materialized bytes	0	0	31,755,115
	Map input records	0	0	1,000,000
	Reduce shuffle bytes	0	0	31,755,115
	Spilled Records	0	0	3,000,000
	Map output bytes	0	0	29,755,103
	Total committed heap usage (bytes)	0	0	451,936,256
	CPU time spent (ms)	0	0	27,640
	Map input bytes	0	0	33,900,940
	SPLIT_RAW_BYTES	222	0	222
	Combine input records	0	0	0
	Reduce input records	0	0	1,000,000
	Reduce input groups	0	0	1,000,000
	Combine output records	0	0	0

Export Output from HDFS to Local

```
Activities | Terminal | Tue May 6, 13:04
hduser@swap-Dell-System-XPS-L502X: ~/hadoop

File Edit View Search Terminal Tabs Help

hduser@swap-Dell-System-XPS-L502X: ~/hadoop$
hduser@swap-Dell-System-XPS-L502X: ~/hadoop$
starting jobtracker, logging to /home/hduser/hadoop/libexec/../logs/hadoop-hduser-jobtracker-swap-Dell-System-XPS-L502X.out
localhost: starting tasktracker, logging to /home/hduser/hadoop/libexec/../logs/hadoop-hduser-tasktracker-swap-Dell-System-XPS-L502X.out
hduser@swap-Dell-System-XPS-L502X: ~/hadoop/bin$ jps
14289 JobTracker
14523 Jps
14444 TaskTracker
13879 NameNode
14031 DataNode
14196 SecondaryNameNode
hduser@swap-Dell-System-XPS-L502X: ~/hadoop/bin$ cd ..
hduser@swap-Dell-System-XPS-L502X: ~/hadoop$ bin/hadoop jar contrib/streaming/hadoop-streaming-1.2.1.jar \-file /home/hduser/mapper.py \-mapper /home
/hduser/mapper.py \-file /home/hduser/reducer.py \-reducer /home/hduser/reducer.py \-input /user/hduser/gutenberg/Tenpofull.txt \-output /user/hduser
/output/o4
packageJobJar: [/home/hduser/mapper.py, /home/hduser/reducer.py, /home/hduser/hadoop/tnp/hadoop-unjar7717210963780518093/] [] /tnp/streamjob6568796048
638085046.jar tmpDir=null
14/05/06 13:02:54 INFO util.NativeCodeLoader: Loaded the native-hadoop library
14/05/06 13:02:54 WARN snappy.LoadSnappy: Snappy native library not loaded
14/05/06 13:02:54 INFO mapred.FileInputFormat: Total input paths to process : 1
14/05/06 13:02:55 INFO streaming.StreamJob: getLocalDirs(): [/home/hduser/hadoop/tnp/napred/local]
14/05/06 13:02:55 INFO streaming.StreamJob: Running job: job_201405061302_0001
14/05/06 13:02:55 INFO streaming.StreamJob: To kill this job, run:
14/05/06 13:02:55 INFO streaming.StreamJob: /home/hduser/hadoop/libexec/../bin/hadoop job \-Dmapred.job.tracker=localhost:54311 \-kill job_201405061302_
0001
14/05/06 13:02:55 INFO streaming.StreamJob: Tracking URL: http://localhost:50030/jobdetails.jsp?jobid=job_201405061302_0001
14/05/06 13:02:56 INFO streaming.StreamJob: map 0% reduce 0%
14/05/06 13:03:07 INFO streaming.StreamJob: map 47% reduce 0%
14/05/06 13:03:08 INFO streaming.StreamJob: map 96% reduce 0%
14/05/06 13:03:09 INFO streaming.StreamJob: map 100% reduce 0%
14/05/06 13:03:17 INFO streaming.StreamJob: map 100% reduce 33%
14/05/06 13:03:20 INFO streaming.StreamJob: map 100% reduce 100%
14/05/06 13:03:27 INFO streaming.StreamJob: Job complete: job_201405061302_0001
14/05/06 13:03:27 INFO streaming.StreamJob: Output: /user/hduser/output/o4
hduser@swap-Dell-System-XPS-L502X: ~/hadoop$ bin/hadoop dfs \-getmerge /user/hduser/output/o1 /home/hduser/output.csv
```

Resulting Dataset

output.csv - LibreOffice Calc

File Edit View Insert Format Tools Data Window Help

LibreOffice Calc

1950s

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	1950s	Adagio	59													
2	1940s	Andante	33													
3	1920s	Vivacissimo	11													
4	1930s	Adagio	6													
5	21st century	Vivace	23315													
6	1980s	Allegretto	5061													
7	1970s	Allegro	2600													
8	1920s	Vivace	12													
9	1960s	Largo	46													
10	1950s	Lento	27													
11	21st century	Largo	709													
12	1920s	Andante	10													
13	1920s	Adagio	1													
14	1990s	Adagio	1088													
15	1990s	Andante	257													
16	1980s	Vivace	3485													
17	1940s	Allegro	66													
18	1980s	Andante moderato	269													
19	1990s	Largo	271													
20	1940s	Largo	1													
21	1950s	Allegretto	424													
22	1940s	Vivace	13													
23	1970s	Moderato	2937													
24	1930s	Lento	2													
25	1990s	Vivace	9057													
26	1940s	Grave	3													
27	1980s	Largo	75													
28	1960s	Larghetto	38													
29	1930s	Grave	1													

Sheet 1 / 1

Default

Sum10

100%