

Assignment 2

Student Name: Akhil Kumar Gour
SJSU ID: 012455586

Report

Web crawling task:

- We had to crawl WikiCFP for big data, data mining, database, and AI conferences and their location every year.
- We wrote our own Java code and build the crawler based on that.
- When searching per category, WikiCFP allows navigation until page 20, so we crawled all 20 pages for big data, data mining, databases, and artificial intelligence. The output of the crawling was in the tab separated format: conference_acronym (event), conference_name(title), and conference_location(place).
- I had a separate output file for all the different categories.

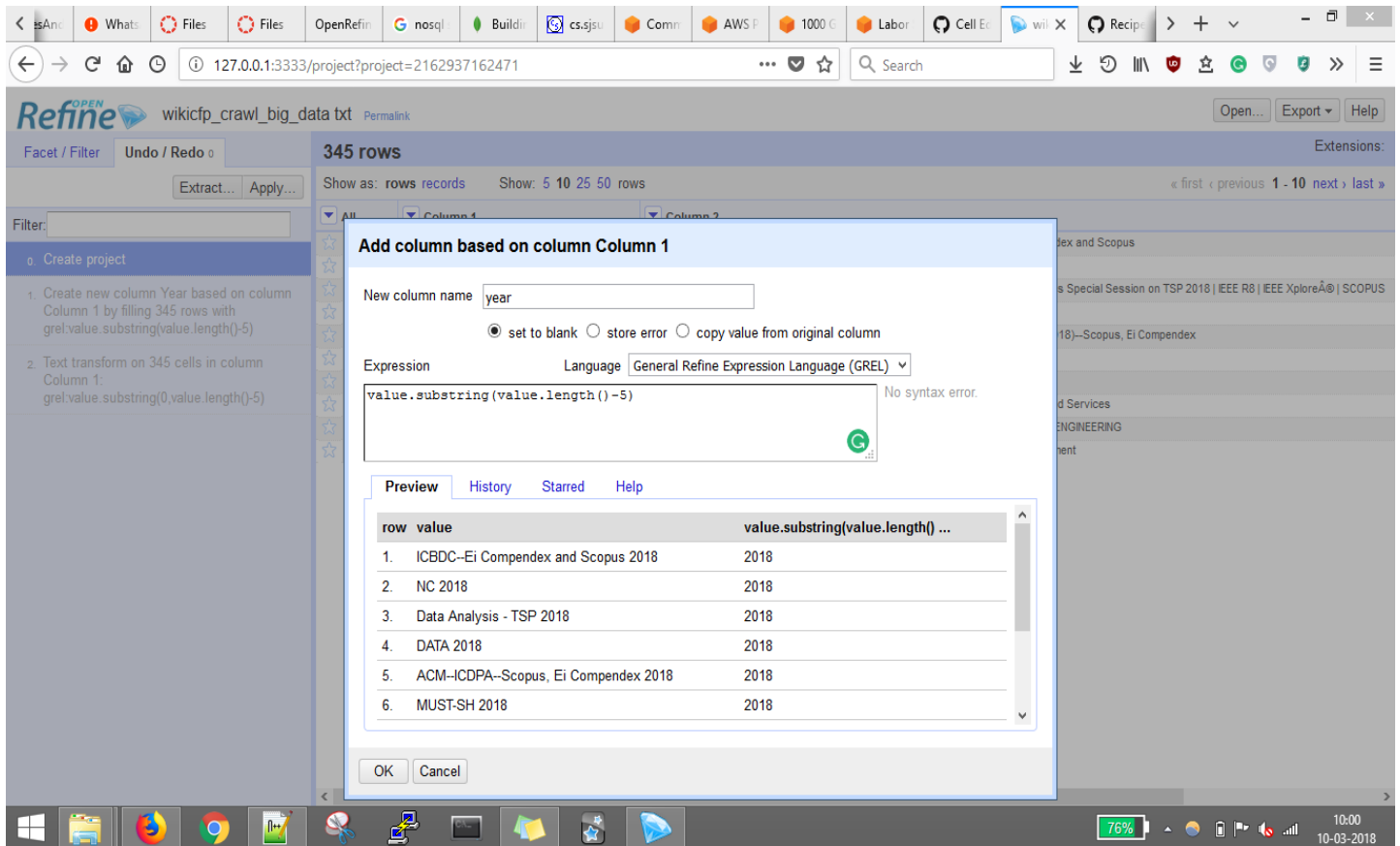
Data cleaning:

- I used OpenRefine to clean each of these files (files for each category).
- I removed the unwanted fields, aggregated and segregated fields as per the requirement.
- I segregated the event acronym and the event year into two separate columns.
- I trimmed the country name from the location field as we only needed results according to the cities.
- Then I merged all the files together and created one single file of the crawled data.

The screenshot shows the OpenRefine web interface. The browser address bar indicates the URL: 127.0.0.1:3333/project?project=2162937162471. The OpenRefine header shows the project name 'wikicfp_crawl_big_data.txt' and a 'Permalink' button. The main table displays 345 rows of data. The first column is labeled 'All', the second 'Column 1', and the third 'Column 2'. The table lists various conferences and their locations. The bottom task pane shows a list of tasks, including 'Create project' and 'Create new column year based on column Column 1 by filling 345 rows with `rel.value.substring(value.length()-5)`'.

	Column 1	Column 2	Column 3
1.	ICBDC--EI Compendex and Scopus 2018	2018 International Conference on Big Data and Computing (ICBDC 2018)--EI Compendex and Scopus	Shenzhen, China
2.	NC 2018	6th International Conference of Networks and Communications	Vienna, Austria
3.	Data Analysis - TSP 2018	Multivariate Data Analysis and Knowledge Discovery â€” From Theory to Applications Special Session on TSP 2018 IEEE R8 IEEE XploreÂ® SCOPUS	Athens, Greece
4.	DATA 2018	7th International Conference on Data Science, Technology and Applications	Porto, Portugal
5.	ACM--ICDPA--Scopus, Ei Compendex 2018	ACM--2018 International Conference on Data Processing and Applications (ICDPA 2018)--Scopus, Ei Compendex	Guangzhou, China
6.	MUST-SH 2018	Multimedia Services and Technologies for Smart-health (two special issues)	San Diego, USA
7.	BGNT 2018	International Workshop on Big Data and Networks Technologies	Gran Canaria, Spain
8.	ACCSE 2018	The Third International Conference on Advances in Computation, Communications and Services	Barcelona, Spain
9.	IMDA 2018	INTELLIGENT METHODS FOR DATA ANALYSIS AND COMPUTER AIDED SOFTWARE ENGINEERING	Vilnius, Lithuania
10.	IMMM 2018	The Eighth International Conference on Advances in Information Mining and Management	Barcelona, Spain

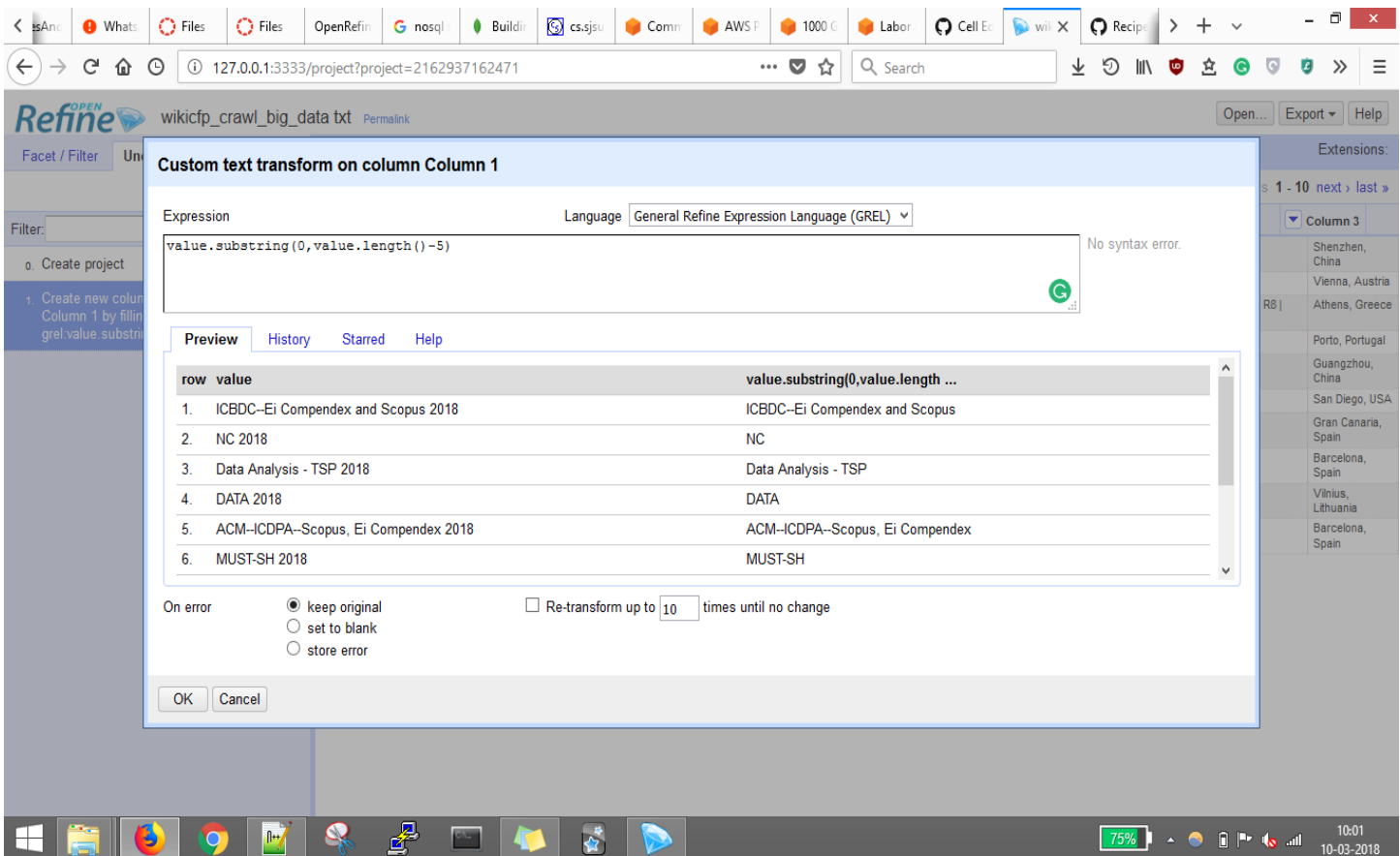
Splitting column1 to get the year column:



Refine tool interface showing the 'Add column based on column Column 1' dialog. The dialog is configured to create a new column named 'year' by applying the expression `value.substring(value.length()-5)` to the first column. The 'set to blank' radio button is selected. The preview table shows the results of the expression on six rows of data.

row	value	value.substring(value.length() ...
1.	ICBDC-Ei Compendex and Scopus 2018	2018
2.	NC 2018	2018
3.	Data Analysis - TSP 2018	2018
4.	DATA 2018	2018
5.	ACM-ICDPA-Scopus, Ei Compendex 2018	2018
6.	MUST-SH 2018	2018

Removing unwanted data:



Refine tool interface showing the 'Custom text transform on column Column 1' dialog. The dialog is configured to apply the expression `value.substring(0,value.length()-5)` to the first column. The 'On error' section has 'keep original' selected. The preview table shows the results of the expression on six rows of data.

row	value	value.substring(0,value.length ...
1.	ICBDC-Ei Compendex and Scopus 2018	ICBDC-Ei Compendex and Scopus
2.	NC 2018	NC
3.	Data Analysis - TSP 2018	Data Analysis - TSP
4.	DATA 2018	DATA
5.	ACM-ICDPA-Scopus, Ei Compendex 2018	ACM-ICDPA-Scopus, Ei Compendex
6.	MUST-SH 2018	MUST-SH

Final data (after all the other operations):

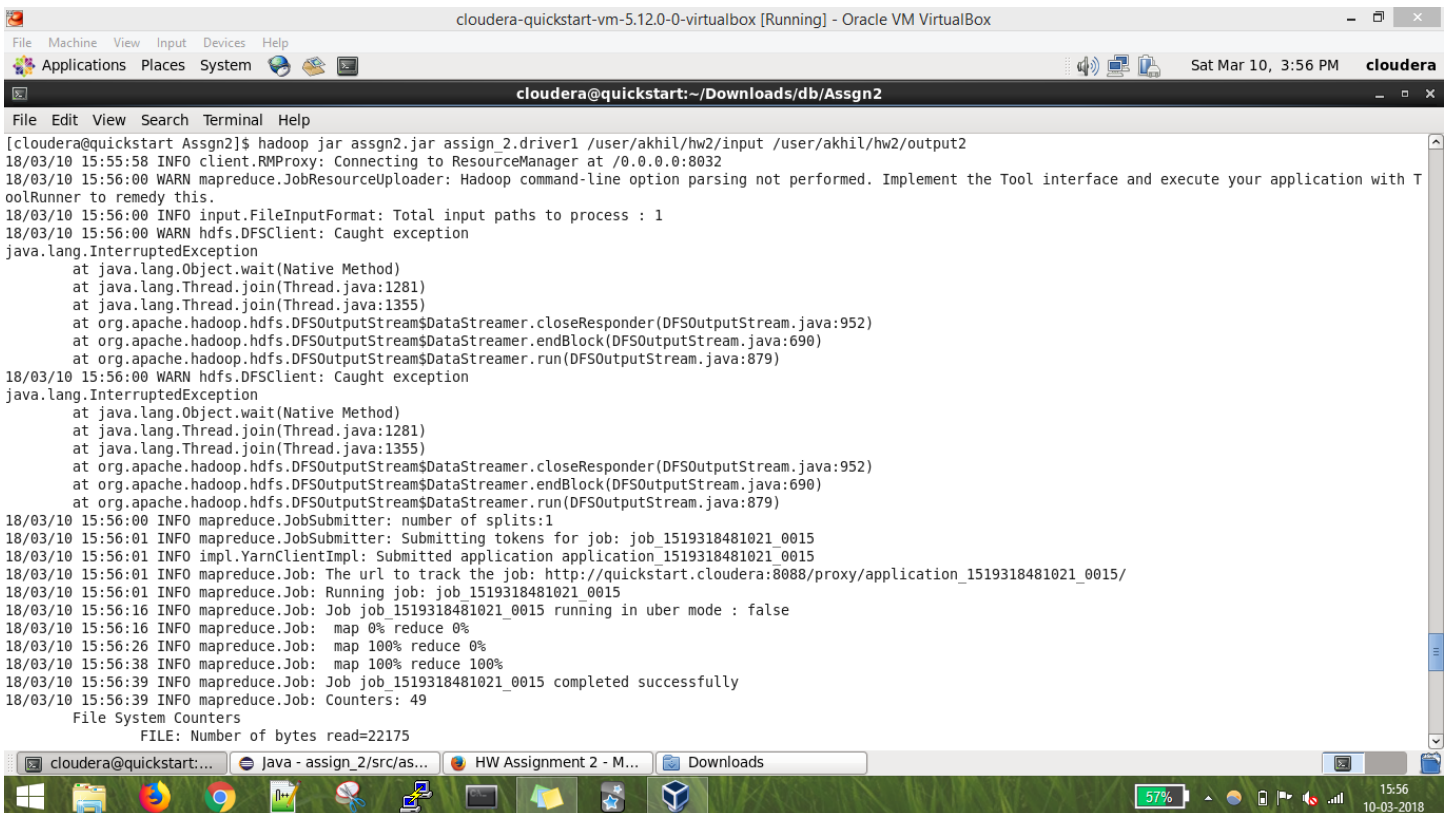
The screenshot displays the Refine Open web application interface. At the top, there's a browser window showing the URL `127.0.0.1:3333/project?project=2162937162471`. Below the browser, the Refine Open logo is visible. The main area shows a data table with 345 rows. The table has columns for `event`, `year`, `title`, `city`, and `country`. The left sidebar shows a list of operations performed on the data, including creating a project, adding columns, and renaming columns. The bottom of the screen shows a Windows taskbar with various application icons and system status indicators.

	event	year	title	city	country
1.	ICBDC—Ei Compindex and Scopus	2018	2018 International Conference on Big Data and Computing (ICBDC 2018)—Ei Compindex and Scopus	Shenzhen	China
2.	NC	2018	6th International Conference of Networks and Communications	Vienna	Austria
3.	Data Analysis - TSP	2018	Multivariate Data Analysis and Knowledge Discovery 86 From Theory to Applications Special Session on TSP 2018 IEEE R8 IEEE XploreA® SCOPUS	Athens	Greece
4.	DATA	2018	7th International Conference on Data Science, Technology and Applications	Porto	Portugal
5.	ACM—ICDPA—Scopus, Ei Compindex	2018	ACM—2018 International Conference on Data Processing and Applications (ICDPA 2018)—Scopus, Ei Compindex	Guangzhou	China
6.	MUST-SH	2018	Multimedia Services and Technologies for Smart-health (two special issues)	San Diego	USA
7.	BIGNT	2018	International Workshop on Big Data and Networks Technologies	Gran Canaria	Spain
8.	ACCSE	2018	The Third International Conference on Advances in Computation, Communications and Services	Barcelona	Spain
9.	IMDA	2018	INTELLIGENT METHODS FOR DATA ANALYSIS AND COMPUTER AIDED SOFTWARE ENGINEERING	Vilnius	Lithuania
10.	IMMM	2018	The Eighth International Conference on Advances in Information Mining and Management	Barcelona	Spain

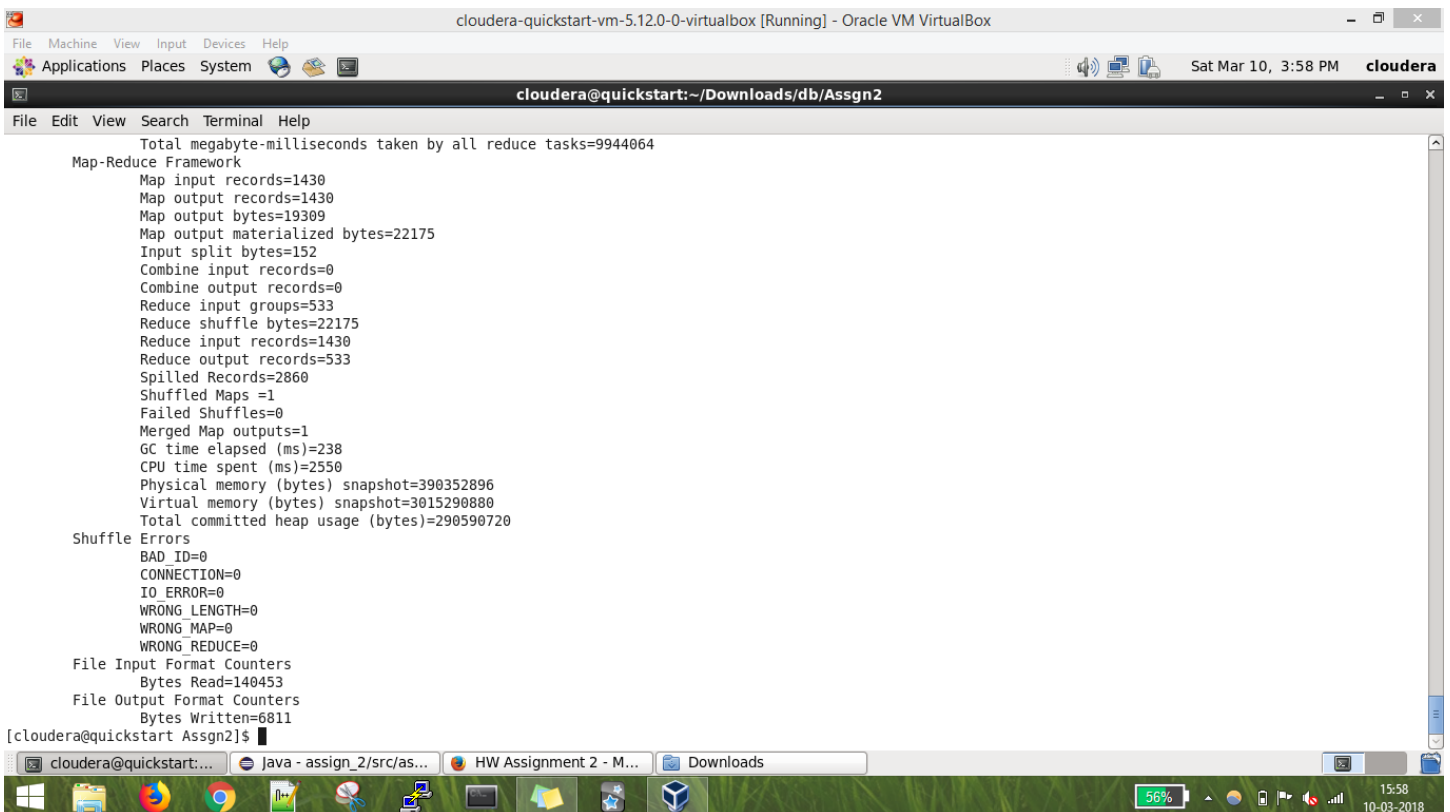
Tasks: In this part, we used Hadoop on the data we just crawled to compute various statistics.

Task I: Compute and plot the number of conferences per city. Which are the top 10 locations?

- For this task, we wrote a Hadoop MapReduce program.
 - Before that, we place the crawled data file in the HDFS using `–put` command.
 - The driver class created a job and called the Mapper and Reducer classes.
 - The Mapper class mapped all the similar cities together and passed it as the outputKey, along with an `IntWritable` variable (1) as outputValue which is used to sum up the count, to the reducer.
 - The reducer then gets the city name and sum up the count 1 by 1 for each city.
 - We export this package as a jar and use the following command to run it as Hadoop jar.
- ```
hadoop jar jarname.jar package.driver /inputpath /outputpath
```



```
cloudera-quickstart-vm-5.12.0-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
cloudera@quickstart:~/Downloads/db/Assign2
File Edit View Search Terminal Help
[cloudera@quickstart Assign2]$ hadoop jar assign2.jar assign 2.driver1 /user/akhil/hw2/input /user/akhil/hw2/output2
18/03/10 15:55:58 INFO client.RMPProxy: Connecting to ResourceManager at /0.0.0.0:8032
18/03/10 15:56:00 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
18/03/10 15:56:00 INFO input.FileInputFormat: Total input paths to process : 1
18/03/10 15:56:00 WARN hdfs.DFSCClient: Caught exception
java.lang.InterruptedException
 at java.lang.Object.wait(Native Method)
 at java.lang.Thread.join(Thread.java:1281)
 at java.lang.Thread.join(Thread.java:1355)
 at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.closeResponder(DFSOutputStream.java:952)
 at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutputStream.java:690)
 at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFSOutputStream.java:879)
18/03/10 15:56:00 WARN hdfs.DFSCClient: Caught exception
java.lang.InterruptedException
 at java.lang.Object.wait(Native Method)
 at java.lang.Thread.join(Thread.java:1281)
 at java.lang.Thread.join(Thread.java:1355)
 at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.closeResponder(DFSOutputStream.java:952)
 at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutputStream.java:690)
 at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFSOutputStream.java:879)
18/03/10 15:56:00 INFO mapreduce.JobSubmitter: number of splits:1
18/03/10 15:56:01 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1519318481021_0015
18/03/10 15:56:01 INFO impl.YarnClientImpl: Submitted application application_1519318481021_0015
18/03/10 15:56:01 INFO mapreduce.Job: The url to track the job: http://quickstart.cloudera:8088/proxy/application_1519318481021_0015/
18/03/10 15:56:01 INFO mapreduce.Job: Running job: job_1519318481021_0015
18/03/10 15:56:16 INFO mapreduce.Job: Job job_1519318481021_0015 running in uber mode : false
18/03/10 15:56:16 INFO mapreduce.Job: map 0% reduce 0%
18/03/10 15:56:26 INFO mapreduce.Job: map 100% reduce 0%
18/03/10 15:56:38 INFO mapreduce.Job: map 100% reduce 100%
18/03/10 15:56:39 INFO mapreduce.Job: Job job_1519318481021_0015 completed successfully
18/03/10 15:56:39 INFO mapreduce.Job: Counters: 49
File System Counters
 FILE: Number of bytes read=22175
```



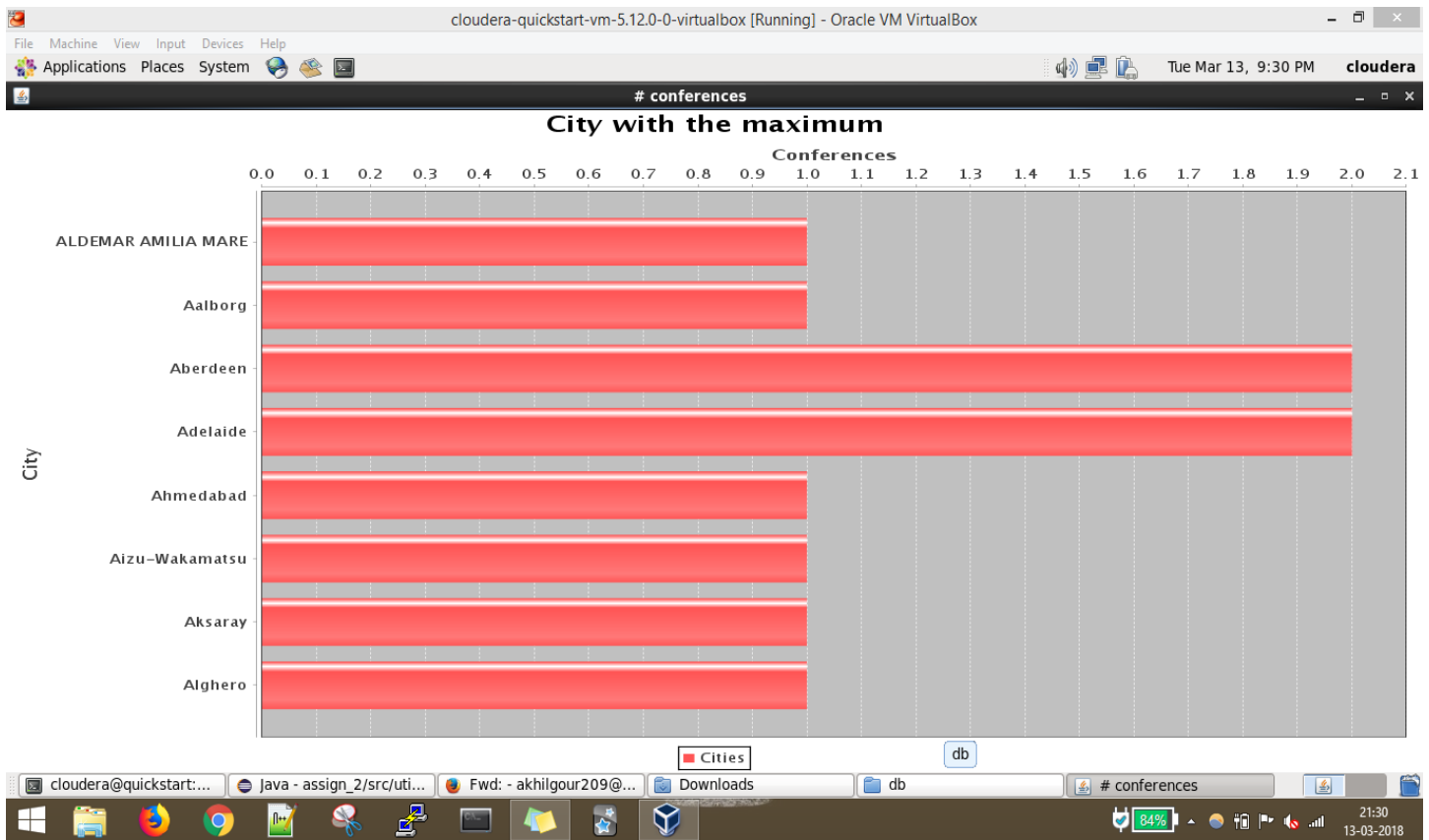
```
cloudera-quickstart-vm-5.12.0-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
cloudera@quickstart:~/Downloads/db/Assign2
File Edit View Search Terminal Help
Total megabyte-milliseconds taken by all reduce tasks=9944064
Map-Reduce Framework
 Map input records=1430
 Map output records=1430
 Map output bytes=19309
 Map output materialized bytes=22175
 Input split bytes=152
 Combine input records=0
 Combine output records=0
 Reduce input groups=533
 Reduce shuffle bytes=22175
 Reduce input records=1430
 Reduce output records=533
 Spilled Records=2860
 Shuffled Maps =1
 Failed Shuffles=0
 Merged Map outputs=1
 GC time elapsed (ms)=238
 CPU time spent (ms)=2550
 Physical memory (bytes) snapshot=390352896
 Virtual memory (bytes) snapshot=3015290880
 Total committed heap usage (bytes)=290590720
Shuffle Errors
 BAD_ID=0
 CONNECTION=0
 IO_ERROR=0
 WRONG_LENGTH=0
 WRONG_MAP=0
 WRONG_REDUCE=0
File Input Format Counters
 Bytes Read=140453
File Output Format Counters
 Bytes Written=6811
[cloudera@quickstart Assign2]$
```

Then we get the output file to our local file system or we can choose to cat the output from the Hadoop itself.

## Output: Top cities in descending order:

```
cloudera-quickstart-vm-5.12.0-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System Mon Mar 12, 10:03 PM cloudera
cloudera@quickstart:~/Downloads/db/Assgn2
File Edit View Search Terminal Help
[cloudera@quickstart Assgn2]$ hadoop fs -cat /user/akhil/hw2/output1/p*| awk '{FS=" ";$0=$0; print $NF|" "$0}'| sort -n -r|cut -d"| " -f2
Halifax 35
Melbourne 29
Dubai 22
Chengdu 22
Beijing 22
Barcelona 19
New Orleans 17
Shanghai 15
Vienna 14
San Diego 14
New York 13
Athens 13
Lyon 12
Seoul 11
Porto 11
Bangkok 11
Tokyo 10
Sydney 10
Seattle 10
Munich 10
Genoa 10
Rio de Janeiro 9
Macau 9
Istanbul 9
Hanoi 9
Brisbane 9
Boston 9
Vancouver 8
Skopje 8
Rome 8
Prague 8
Paris 8
Montreal 8
```

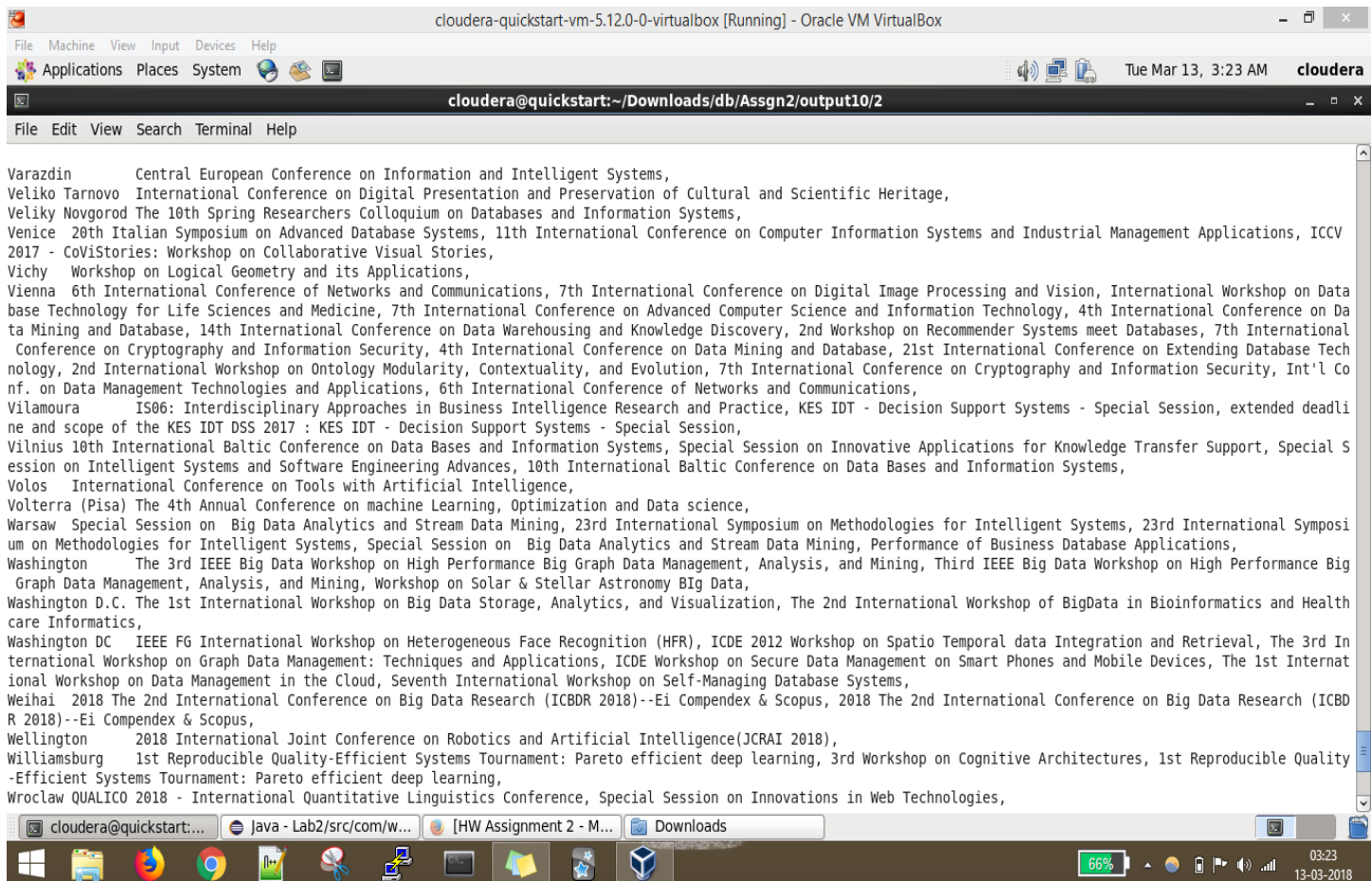
## Graph:



## **Task II:** Output the list of conferences per city.

- For this task, I included created another Mapper and Reducer class.
- I created a separate job for these Mapper and Reducer class in the previous driver file itself.
- The Mapper passes each city as the outputKey to the reducer.
- Each conference then, is passes as the outputValue to the reducer, corresponding to its city.
- The reducer, in turn, appends each conference name to its corresponding city in which it happened.

**Output:** The list of city along with the conferences that took place in those cities.



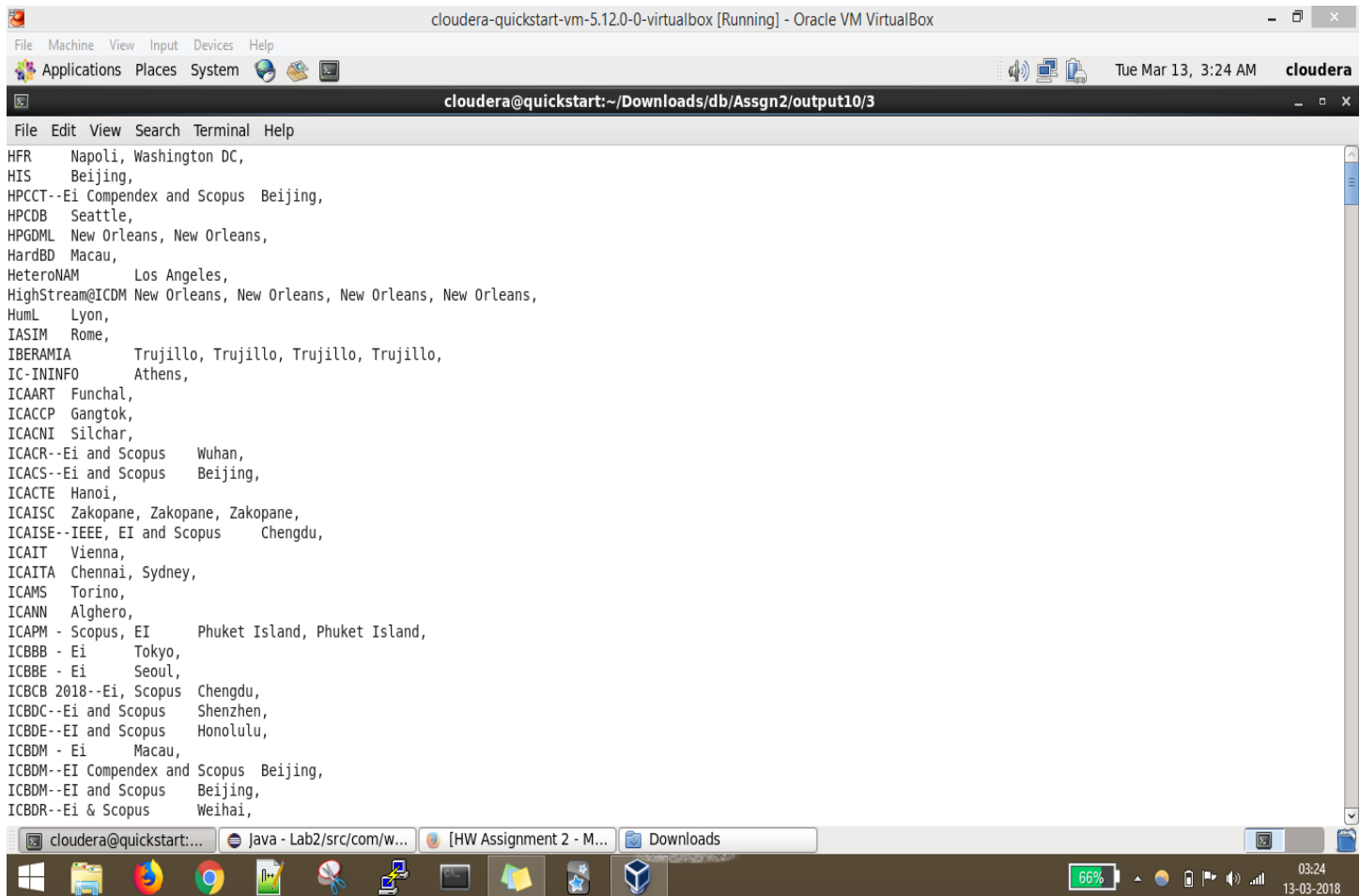
The screenshot shows a terminal window titled "cloudera-quickstart-vm-5.12.0-0-virtualbox [Running] - Oracle VM VirtualBox". The terminal output lists conferences grouped by city. The cities listed are: Varazdin, Veliko Tarnovo, Veliky Novgorod, Venice, 2017 - CoViStories, Vichy, Vienna, Vilamoura, Vilnius, Volos, Volterra (Pisa), Warsaw, Washington, Washington D.C., Washington DC, Weihai, Wellington, Williamsburg, and Wroclaw. Each city is followed by a list of conferences that took place there. For example, under "Varazdin", it lists "Central European Conference on Information and Intelligent Systems,". The terminal window also shows a taskbar at the bottom with various application icons and a system tray on the right.

## **Task III:** For each conference regardless of the year (e.g., KDD), output the list of cities.

- I split the data that was the input for the mapper with regards to tabs.
- The conference name (event) was in the column 1 of my wifi\_crawl data file.
- This I passed as the outputKey to the reducer.
- Then I passed the city name, which was column 4 of my data file, as outputValue to the reducer.
- The reducer then appended each of the conference name to its corresponding city.



**Output:** Conference with a list of cities in which they were held.



```
cloudera@quickstart:~/Downloads/db/Assgn2/output10/3
File Edit View Search Terminal Help
HFR Napoli, Washington DC,
HIS Beijing,
HPCCT--Ei Compendex and Scopus Beijing,
HPCDB Seattle,
HPGDML New Orleans, New Orleans,
HardBD Macau,
HeteroNAM Los Angeles,
HighStream@ICDM New Orleans, New Orleans, New Orleans, New Orleans,
HumL Lyon,
IASIM Rome,
IBERAMIA Trujillo, Trujillo, Trujillo, Trujillo,
IC-ININFO Athens,
ICAART Funchal,
ICACCP Gangtok,
ICACNI Silchar,
ICACR--Ei and Scopus Wuhan,
ICACS--Ei and Scopus Beijing,
ICACTE Hanoi,
ICAISC Zakopane, Zakopane, Zakopane,
ICAISE--IEEE, EI and Scopus Chengdu,
ICAIT Vienna,
ICAITA Chennai, Sydney,
ICAMS Torino,
ICANN Alghero,
ICAPM - Scopus, EI Phuket Island, Phuket Island,
ICBBB - Ei Tokyo,
ICBBE - Ei Seoul,
ICBCB 2018--Ei, Scopus Chengdu,
ICBDC--Ei and Scopus Shenzhen,
ICBDE--EI and Scopus Honolulu,
ICBDM - Ei Macau,
ICBDM--Ei Compendex and Scopus Beijing,
ICBDM--EI and Scopus Beijing,
ICBDR--Ei & Scopus Weihai,
```

**Task IV:** For each city compute and plot a time series of number of conferences per year.

- For this task, I created 2 Mappers and 2 Reducers.
- The first mapper takes the event column, concatenates it with the year column value, and pass this new string as the outputValue.
- The city column is passed to the reducer as the ouputKey.
- The first reducer appends the list of events that happened in a year corresponding to the city.
- The second mapper then concatenates city with the year and passes it on as the outputKey.
- The outputValue is an IntWritable(1), which is used to increase the total count of each event that happened in that year in that particular city.

**Output:**

Number of conferences per year that happened in a particular city.

cloudera-quickstart-vm-5.12.0-0-virtualbox [Running] - Oracle VM VirtualBox

File Machine View Input Devices Help

Applications Places System

Tue Mar 13, 11:12 PM cloudera

cloudera@quickstart: ~/Downloads/db/Assign2/finalOutput/4

File Edit View Search Terminal Help

|                                      |    |  |
|--------------------------------------|----|--|
| Burlingame2012                       | 1  |  |
| Burlingame2013                       | 1  |  |
| Busan2012                            | 1  |  |
| CANCUN2017                           | 1  |  |
| Caen2017                             | 1  |  |
| Cagliari2017                         | 1  |  |
| Cairo2017                            | 2  |  |
| Cairo2018                            | 1  |  |
| Cala Millor2018                      | 1  |  |
| Calgary2018                          | 1  |  |
| Cali2018                             | 1  |  |
| Cambridge2012                        | 1  |  |
| Cambridge2017                        | 3  |  |
| Cambridge2018                        | 1  |  |
| Cancun2017                           | 1  |  |
| Casablanca2018                       | 2  |  |
| Central University of Technology2017 | 2  |  |
| Certosa di Pontignano - Siena2018    | 1  |  |
| Ceske Budejovice2018                 | 1  |  |
| Changa2017                           | 2  |  |
| Changchun2018                        | 1  |  |
| Changsha2014                         | 1  |  |
| Chania2012                           | 3  |  |
| ChengDu2017                          | 2  |  |
| Chengdu2016                          | 4  |  |
| Chengdu2018                          | 18 |  |
| Chennai2018                          | 8  |  |
| Chiang Mai2018                       | 1  |  |
| Chicago IL2014                       | 1  |  |
| Chicago2014                          | 4  |  |
| Chicago2017                          | 1  |  |
| Chongqing2018                        | 4  |  |
| Clermont-Ferrand2011                 | 1  |  |
| Coimbra2013                          | 2  |  |

Downloads

cloudera@quickstart:~/Downloads/db/Assign2/finalOutput/4

Java - assign\_2/src/gr... Fwd: - akhilgour209@... Downloads cloudera Desktop

46% 23:48 13-03-2018

cloudera-quickstart-vm-5.12.0-0-virtualbox [Running] - Oracle VM VirtualBox

File Machine View Input Devices Help

Applications Places System

Tue Mar 13, 11:49 PM cloudera

cloudera@quickstart: ~/Downloads/db/Assign2/finalOutput/4

File Edit View Search Terminal Help

|                             |   |  |
|-----------------------------|---|--|
| Jadavpur University2017     | 1 |  |
| Jain College of Enginee2017 | 1 |  |
| Jaipur2017                  | 6 |  |
| Jeju Island2017             | 1 |  |
| Jeju Island2018             | 1 |  |
| Jeju2018                    | 1 |  |
| Jinan2018                   | 4 |  |
| Jyvaskyla2018               | 1 |  |
| KAIST2017                   | 1 |  |
| Kalamata2018                | 1 |  |
| Kalbis Institute2017        | 1 |  |
| Kanazawa2017                | 2 |  |
| Kansas City2017             | 1 |  |
| Kaohsiung2011               | 1 |  |
| Kaohsiung2012               | 1 |  |
| Kaohsiung2014               | 1 |  |
| Kazan2013                   | 1 |  |
| Khenchela2016               | 1 |  |
| Kitakyushu2017              | 1 |  |
| Koblenz2013                 | 1 |  |
| Kohala Coast2015            | 1 |  |
| Kolkata2017                 | 1 |  |
| Krakow2018                  | 1 |  |
| Kuala Lumpur2013            | 3 |  |
| Kuala Lumpur2018            | 1 |  |
| Kuantan2018                 | 2 |  |
| Kyoto Terra2013             | 1 |  |
| Kyoto2017                   | 2 |  |
| Kyoto2018                   | 4 |  |
| La Habana2013               | 1 |  |
| La Havana2018               | 2 |  |
| La Rioja2017                | 1 |  |
| Lago Di Garda2013           | 1 |  |
| Laguna Hills2018            | 2 |  |

Downloads

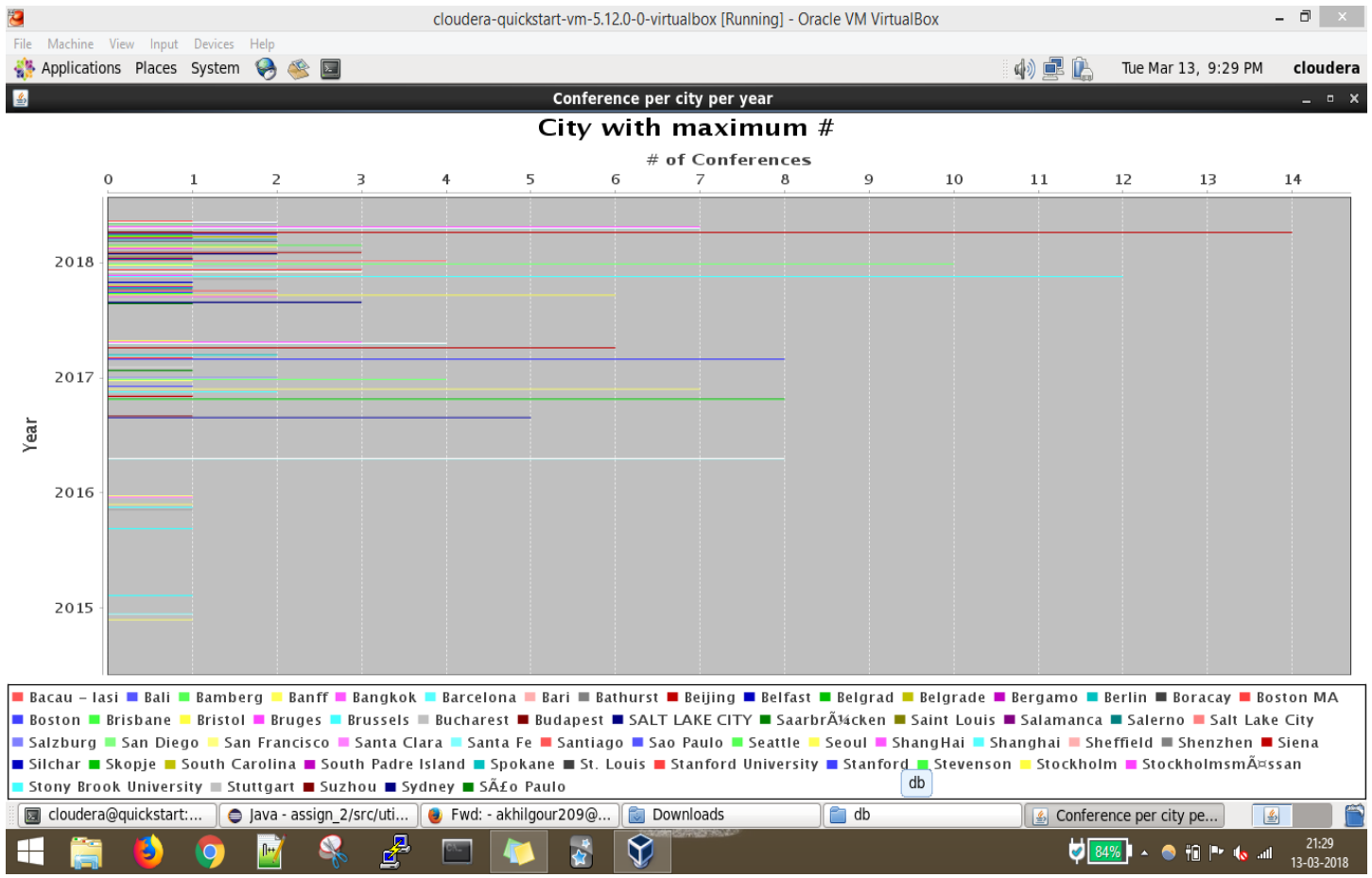
cloudera@quickstart:~/Downloads/db/Assign2/finalOutput/4

Java - assign\_2/src/gr... Fwd: - akhilgour209@... Downloads cloudera Desktop

45% 23:49 13-03-2018



## Graph:



## Lessons learnt:

- How to use multiple Mapper/Reducer.
- Plotting instances of the output to a graph.
- How to include multiple Mapper/Reducer jobs within a java application.