IERG 4330/ESTR 4316/IEMS 5730 Spring 2022
Homework 2

Release date: Feb 20, 2022

Due date: Mar 8, 2022 (Tuesday) 11:59:00 pm

We will discuss the solution soon after the deadline. No late homework will be accepted!

Every Student MUST include the following statement, together with his/her signature in the submitted homework.

I declare that the assignment submitted on Elearning system is original except for source material explicitly acknowledged, and that the same or related material has not been previously submitted for another course. I also acknowledge that I am aware of University policy and regulations on honesty in academic work, and of the disciplinary guidelines and procedures applicable to breaches of such policy and regulations, as contained in the website http://www.cuhk.edu.hk/policy/academichonesty/.

Signed (Student Eddie Christopher Fox III) Date: March 4, 2022

Name Eddie Christopher Fox III SID 1155160788

Submission notice:

- Submit your homework via the elearning system.
- All students are required to submit this assignment.

General homework policies:

A student may discuss the problems with others. However, the work a student turns in must be created COMPLETELY by oneself ALONE. A student may not share ANY written work or pictures, nor may one copy answers from any source other than one's own brain.

Each student **MUST LIST** on the homework paper the **name of every person he/she has discussed or worked with**. If the answer includes content from any other source, the student **MUST STATE THE SOURCE**. Failure to do so is cheating and will result in sanctions. Copying answers from someone else is cheating even if one lists their name(s) on the homework.

If there is information you need to solve a problem but the information is not stated in the problem, try to find the data somewhere. If you cannot find it, state what data you need, make a reasonable estimate of its value, and justify any assumptions you make. You will be graded not only on whether your answer is correct, but also on whether you have done an intelligent analysis.

IEMS 5730: Spring 2022 Assignment #2

Question 1 Section A: [Bonus] Installing Pig on Hadoop Cluster

I refer to [1] for instructions on installing Pig on the Hadoop Cluster, provided by the Assignment 2 PDF. Even though I am installing Pig on the Hadoop cluster, I prefer to use the IE DIC cluster after this to ensure there is no errors in installation that may cause bugs later, as well as to leverage the (probably) faster performance and greater available memory of the cluster in case there is memory scalability issues.

Hadoop Cluster Debugging:

First, I go to Google Cloud and start up all my nodes again (which I had shut down / stopped to avoid being billed) and modify my Putty sessions to account for their new external IP addresses.

The cluster was not functioning properly and would not show datanodes in the slave nodes when running start-dfs.sh from sbin. I tried many methods to debug this, but all failed. Ultimately, I had to delete all the virtual machines in my hadoop cluster and recreate them one by one from the machine images I made of each machine at the end of Assignment 0 (good thing you suggested doing so).

Afterwards, I had to do the following to get the cluster working again (after some debugging).

1. Modify putty sessions to account for new external IP addresses
2. Modify hosts file of every VM to reflect the new internal IP addresses
3. Delete all nodes (including secondary namenode) from known_hosts file under ssh directory and adding the ssh keys again (internal IP address changed, creating issues as the security measures thought I was being hit with a man in the middle attack). Did this on all nodes.
4. Delete subdirectories of tmp directory under /usr/local/hadoop on each node.
5. Delete all subdirectories of hadoop_data directory under /usr/local/hadoop on each node.
6. Format namenode on master node
7. Run start-dfs.sh and start-yarn.sh on master node.
8. Verify functionality through jps and the localhost web interface.

Actual Pig Installation:

With this out of the way, I was able to install Pig. The download was not under the "Pig Releases" hyperlink under [1], but I was able to find the proper source at [2], and downloaded pig-0.17.0.tar.gz by copying the link from [2] and running "wget https://dlcdn.apache.org/pig/pig-0.17.0/pig-0.17.0.tar.gz" into my home directory that already had the hadoop.tar.gz

Index of /pig/pig-0.17.0

Name	Last modified	Size	Description
Parent Directory		-	
RELEASE_NOTES.txt	2020-07-06 15:20	1.9K	
pig-0.17.0-src.tar.gz	2020-07-06 15:21	15M	
pig-0.17.0-src.tar.gz.asc	2020-07-06 15:20	488	
pig-0.17.0-src.tar.gz.md5	2020-07-06 15:20	56	
pig-0.17.0.tar.gz	2020-07-06 15:21	220M	
pig-0.17.0.tar.gz.asc	2020-07-06 15:20	488	
pig-0.17.0.tar.gz.md5	2020-07-06 15:21	52	

```
EddieCFox@master:/usr$ cd ~
EddieCFox@master:~$ ls
hadoop-2.9.2.tar.gz
EddieCFox@master:~$ wget https://dlcdn.apache.org/pig/pig-0.17.0/pig-0.17.0.tar.gz
--2022-02-25 15:39:43--  https://dlcdn.apache.org/pig/pig-0.17.0/pig-0.17.0.tar.gz
Resolving dlcdn.apache.org (dlcdn.apache.org)... 151.101.2.132, 2a04:4e42::644
Connecting to dlcdn.apache.org (dlcdn.apache.org)|151.101.2.132|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 230606579 (220M) [application/x-gzip]
Saving to: 'pig-0.17.0.tar.gz'

pig-0.17.0.tar.gz                   100%[===================================================================================================>] 219.92M  6.73MB/s    in 21s

2022-02-25 15:40:05 (10.3 MB/s) - 'pig-0.17.0.tar.gz' saved [230606579/230606579]

EddieCFox@master:~$
```

I then ran “tar -xzvf pig-0.17.0.tar.gz” to unpack it per [3].

Impossible to capture the output of this fully because there was so many files that I can’t scroll back to when I ran the command. The ending of the output is here.

```
pig-0.17.0/tutorial/scripts/script2-local.pig
pig-0.17.0/tutorial/src/org/apache/pig/tutorial/ExtractHour.java
pig-0.17.0/tutorial/src/org/apache/pig/tutorial/NGramGenerator.java
pig-0.17.0/tutorial/src/org/apache/pig/tutorial/NonURLDetector.java
pig-0.17.0/tutorial/src/org/apache/pig/tutorial/ScoreGenerator.java
pig-0.17.0/tutorial/src/org/apache/pig/tutorial/ToLower.java
pig-0.17.0/tutorial/src/org/apache/pig/tutorial/TutorialTest.java
pig-0.17.0/tutorial/src/org/apache/pig/tutorial/TutorialUtil.java
pig-0.17.0/bin/pig
pig-0.17.0/bin/pig.cmd
pig-0.17.0/bin/pig.py
EddieCFox@master:~$ ls
hadoop-2.9.2.tar.gz  pig-0.17.0  pig-0.17.0.tar.gz
EddieCFox@master:~$ cd pig-0.17.0/
EddieCFox@master:~/pig-0.17.0$ ls
CHANGES.txt  NOTICE.txt  RELEASE_NOTES.txt  build.xml  contrib  ivy      legacy  lib-src  pig-0.17.0-core-h2.jar  shims  test
LICENSE.txt  README.txt  bin                conf       docs     ivy.xml  lib     license  scripts                 src    tutorial
EddieCFox@master:~/pig-0.17.0$
```

I navigate into the pig directory and confirm the contents.

```
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
export PATH=/home/EddieCFox/pig-0.17.0/bin:$PATH
export PIG_CLASSPATH=/usr/local/hadoop/etc/hadoop
```

I add the later two lines to my .bashrc file in the home directory. The first line, exporting the PATH was advised by [1] and is used so I can call it simply by using “Pig”. The second line, about exporting the PIG_CLASSPATH is inspired by [1] and [3] which both mention this is necessary to run Pig on the Hadoop cluster instead of locally.

[1] suggests I test the Pig installation using “pig -help”, so I do.

```
EddieCFox@master:~$ pig -help

Apache Pig version 0.17.0 (r1797386)
compiled Jun 02 2017, 15:41:58

USAGE: Pig [options] [-] : Run interactively in grunt shell.
       Pig [options] -e[xecute] cmd [cmd ...] : Run cmd(s).
       Pig [options] [-f[ile]] file : Run cmds found in file.
  options include:
    -4, -log4jconf - Log4j configuration file, overrides log conf
    -b, -brief - Brief logging (no timestamps)
    -c, -check - Syntax check
    -d, -debug - Debug level, INFO is default
    -e, -execute - Commands to execute (within quotes)
    -f, -file - Path to the script to execute
    -g, -embedded - ScriptEngine classname or keyword for the ScriptEngine
    -h, -help - Display this message. You can specify topic to get help for that topic.
        properties is the only topic currently supported: -h properties.
    -i, -version - Display version information
    -l, -logfile - Path to client side log file; default is current working directory.
    -m, -param_file - Path to the parameter file
    -p, -param - Key value pair of the form param=val
    -r, -dryrun - Produces script with substituted parameters. Script is not executed.
    -t, -optimizer_off - Turn optimizations off. The following values are supported:
            ConstantCalculator - Calculate constants at compile time
            SplitFilter - Split filter conditions
            PushUpFilter - Filter as early as possible
            MergeFilter - Merge filter conditions
            PushDownForeachFlatten - Join or explode as late as possible
            LimitOptimizer - Limit as early as possible
            ColumnMapKeyPrune - Remove unused data
            AddForEach - Add ForEach to remove unneeded columns
            MergeForEach - Merge adjacent ForEach
            GroupByConstParallelSetter - Force parallel 1 for "group all" statement
            PartitionFilterOptimizer - Pushdown partition filter conditions to loader implementing LoadMetaData
            PredicatePushdownOptimizer - Pushdown filter predicates to loader implementing LoadPredicatePushDown
            All - Disable all optimizations
        All optimizations listed here are enabled by default. Optimization values are case insensitive.
    -v, -verbose - Print all error messages to screen
    -w, -warning - Turn warning logging on; also turns warning aggregation off
    -x, -exectype - Set execution mode: local|mapreduce|tez, default is mapreduce.
    -F, -stop_on_failure - Aborts execution on the first failed job; default is off
    -M, -no_multiquery - Turn multiquery optimization off; default is on
    -N, -no_fetch - Turn fetch optimization off; default is on
    -P, -propertyFile - Path to property file
    -printCmdDebug - Overrides anything else and prints the actual command used to run Pig, including
                     any environment variables that are set by the pig command.
22/02/25 15:57:46 INFO pig.Main: Pig script completed in 143 milliseconds (143 ms)
EddieCFox@master:~$
```

With this working, I now try invoke pig itself in local mode using “pig -x local” [1].

```
EddieCFox@master:~$ pig -x local
22/02/25 15:59:48 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
22/02/25 15:59:48 INFO pig.ExecTypeProvider: Picked LOCAL as the ExecType
2022-02-25 15:59:48,329 [main] INFO  org.apache.pig.Main - Apache Pig version 0.17.0 (r1797386) compiled Jun 02 2017, 15:41:58
2022-02-25 15:59:48,329 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/EddieCFox/pig_1645804788328.log
2022-02-25 15:59:48,340 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/EddieCFox/.pigbootup not found
2022-02-25 15:59:48,489 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2022-02-25 15:59:48,491 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: file:///
2022-02-25 15:59:48,628 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2022-02-25 15:59:48,641 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-eb63daca-2faf-4fc3-9f6c-3a2f5ca4e3ac
2022-02-25 15:59:48,642 [main] WARN  org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
grunt>
```

The grunt shell opens successfully. There are some messages, but no errors, merely saying that certain configurations are deprecated, likely because I configured Hadoop using Hadoop 2.X when Hadoop is now on 3.X and has different names for different configuration options.

We can see that when it is executing locally, it notes this by “Picked LOCAL as the ExecType”.

Finally, I run “pig” with no switches to confirm it can run via the Hadoop cluster.

```
EddieCFox@master:~$ pig
22/02/25 16:05:01 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
22/02/25 16:05:01 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
22/02/25 16:05:01 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2022-02-25 16:05:01,849 [main] INFO  org.apache.pig.Main - Apache Pig version 0.17.0 (r1797386) compiled Jun 02 2017, 15:41:58
2022-02-25 16:05:01,849 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/EddieCFox/pig_1645805101848.log
2022-02-25 16:05:01,860 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/EddieCFox/.pigbootup not found
2022-02-25 16:05:02,306 [main] WARN  org.apache.hadoop.util.NativeCodeLoader - Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
2022-02-25 16:05:02,323 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2022-02-25 16:05:02,323 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://master:9000
2022-02-25 16:05:02,801 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-fecc5184-e991-4bff-8335-847648374aeb
2022-02-25 16:05:02,801 [main] WARN  org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
grunt>
```

It first tries the local execution type, then tries MapReduce, then finally says “Picked MapReduce as the Exec type”, indicating that it is using the hadoop cluster for the pig commands.

One line of interest says something like “WARN.... Unable to load native-hadoop library for your platform... using builtin-java classes where applicable.” [4] notes this is just a warning and won’t impact hadoop or pig functionality. I could add a few export commands to fix it, but decide against doing so since I am using the IE DIC cluster for Pig anyway.

Question 1 Section B: Uploading datasets to HDFS and joining them

I first confirm that the IE DIC cluster works as expected.

```
[s1155160788@dicvmd10 ~]$ pig
22/02/26 09:04:56 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
22/02/26 09:04:56 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
22/02/26 09:04:56 INFO pig.ExecTypeProvider: Trying ExecType : TEZ_LOCAL
22/02/26 09:04:56 INFO pig.ExecTypeProvider: Trying ExecType : TEZ
22/02/26 09:04:56 INFO pig.ExecTypeProvider: Picked TEZ as the ExecType
2022-02-26 09:04:56,822 [main] INFO  org.apache.pig.Main - Apache Pig version 0.16.0.2.6.5.0-292 (rUnversioned directory) compiled May 11 2018, 07:56:28
2022-02-26 09:04:56,822 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/s1155160788/pig_1645837496821.log
2022-02-26 09:04:56,835 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/s1155160788/.pigbootup not found
2022-02-26 09:04:57,168 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://dicvmd2.ie.cuhk.edu.hk:8020
2022-02-26 09:04:57,631 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-6b0510c9-30cc-4ab8-9948-af3b2464c493
2022-02-26 09:04:57,854 [main] INFO  org.apache.hadoop.yarn.client.api.impl.TimelineClientImpl - Timeline service address: http://dicvmd2.ie.cuhk.edu.hk:8188/ws/v1/timeline/
2022-02-26 09:04:57,924 [main] INFO  org.apache.pig.backend.hadoop.PigATSClient - Created ATS Hook
grunt> quit
2022-02-26 09:05:21,020 [main] INFO  org.apache.pig.Main - Pig script completed in 24 seconds and 307 milliseconds (24307 ms)
```

From this we can see that it is already configured to run Tez Pig (probably for performance) by typing “pig” without needing to use “pig -x tez” and also that this is version 0.16 of pig, not version 0.17. Version 0.17 introduces Pig on Spark, but Tez should be sufficient.

I do some tidying up of my directories and put everything from Assignment 1 into an Assignment 1 directory, then create an Assignment 2 directory and create a folder for the raw datasets.

From this new directory, I run “wget http://storage.googleapis.com/books/ngrams/books/googlebooks-eng-all-1gram-20120701-a.gz” for the first part of the dataset and “wget http://storage.googleapis.com/books/ngrams/books/googlebooks-eng-all-1gram-20120701-b.gz” for the second part of the dataset, to download them into the dataset directory.

```
[s1155160788@dicvmd10 Datasets_Raw]$ wget http://storage.googleapis.com/books/ngrams/books/googlebooks-eng-all-1gram-20120701-a.gz
--2022-02-26 09:16:36--  http://storage.googleapis.com/books/ngrams/books/googlebooks-eng-all-1gram-20120701-a.gz
Resolving storage.googleapis.com (storage.googleapis.com)... 142.250.66.48, 142.250.66.80, 142.250.204.80, ...
Connecting to storage.googleapis.com (storage.googleapis.com)|142.250.66.48|:80... connected.
HTTP request sent, awaiting response... 200 OK
Length: 343487895 (328M) [binary/octet-stream]
Saving to: 'googlebooks-eng-all-1gram-20120701-a.gz'

100%[===================================================================================================================================>] 343,487,895 27.6MB/s   in 12s

2022-02-26 09:16:48 (27.9 MB/s) - 'googlebooks-eng-all-1gram-20120701-a.gz' saved [343487895/343487895]

[s1155160788@dicvmd10 Datasets_Raw]$ wget http://storage.googleapis.com/books/ngrams/books/googlebooks-eng-all-1gram-20120701-b.gz
--2022-02-26 09:17:02--  http://storage.googleapis.com/books/ngrams/books/googlebooks-eng-all-1gram-20120701-b.gz
Resolving storage.googleapis.com (storage.googleapis.com)... 142.250.66.112, 142.250.204.144, 172.217.31.16, ...
Connecting to storage.googleapis.com (storage.googleapis.com)|142.250.66.112|:80... connected.
HTTP request sent, awaiting response... 200 OK
Length: 247964820 (236M) [binary/octet-stream]
Saving to: 'googlebooks-eng-all-1gram-20120701-b.gz'

100%[===================================================================================================================================>] 247,964,820 22.3MB/s   in 11s

2022-02-26 09:17:15 (22.3 MB/s) - 'googlebooks-eng-all-1gram-20120701-b.gz' saved [247964820/247964820]

[s1155160788@dicvmd10 Datasets_Raw]$
```

After viewing the output briefly with cat and seeing gibberish, it is obvious I need to decompress the gzip file, which I do with [4] by running "gzip -d filename". This command decompresses the gzip files then deletes the gzip file, leaving only the uncompressed version.

```
[s1155160788@dicvmd10 Datasets_Raw]$ ls
googlebooks-eng-all-1gram-20120701-a.gz  googlebooks-eng-all-1gram-20120701-b.gz
[s1155160788@dicvmd10 Datasets_Raw]$ gzip -d googlebooks-eng-all-1gram-20120701-a.gz
[s1155160788@dicvmd10 Datasets_Raw]$ gzip -d googlebooks-eng-all-1gram-20120701-b.gz
[s1155160788@dicvmd10 Datasets_Raw]$ ls
googlebooks-eng-all-1gram-20120701-a  googlebooks-eng-all-1gram-20120701-b
[s1155160788@dicvmd10 Datasets_Raw]$
```

```
[s1155160788@dicvmd10 Datasets_Raw]$ hdfs dfs -copyFromLocal googlebooks-eng-all-1gram-20120701-a
[s1155160788@dicvmd10 Datasets_Raw]$ hdfs dfs -copyFromLocal googlebooks-eng-all-1gram-20120701-b
[s1155160788@dicvmd10 Datasets_Raw]$ hdfs dfs -ls
Found 4 items
-rw-r--r--   3 s1155160788 student 1801526075 2022-02-26 09:25 googlebooks-eng-all-1gram-20120701-a
-rw-r--r--   3 s1155160788 student 1268392934 2022-02-26 09:25 googlebooks-eng-all-1gram-20120701-b
-rw-r--r--   3 s1155160788 student   17103777 2022-02-09 16:28 movielens_large_updated.csv
-rw-r--r--   3 s1155160788 student     893989 2022-02-09 16:28 movielens_small.csv
[s1155160788@dicvmd10 Datasets_Raw]$
```

To write Pig scripts, I use Visual Studio Code on my laptop using a Pig Latin extension, then transfer to the server using Filezilla.

The first thing I need to do is load each dataset that I have uploaded into HDFS. To do this, I need to know their location. [6] notes I can use "hdfs dfs -ls /" to list the root directory of HDFS. Using this and drilling down one folder at a time, I determine that my datasets are located at "/user/s1155160788".

```
q1.pig
C: > Users > Eddie > Documents > School > CUHK > 2021 - 2022 Term 2 > IEMS 5730 - Big Data Systems and Information Processing > Week 7 > Assignment 2 > q1.pig
A = LOAD 'googlebooks-eng-all-1gram-20120701-a' USING PigStorage('\t') AS (bigram:chararray, year:int, match_count:int, volume_count:int);
B = LOAD 'googlebooks-eng-all-1gram-20120701-b' USING PigStorage('\t') AS (bigram:chararray, year:int, match_count:int, volume_count:int);
C = UNION A, B;
STORE C INTO '/output' USING PigStorage('\t');
```

I create these first two lines by referencing [7], which provides examples of loading tab separated tables using PigStorage and specifying the schema. I use the same names as the Assignment 2 PDF.

The 3rd line also comes from [7]. TA's clarified on the Blackboard Forum that by joining the files into one table, they meant UNION, and this is the standard form for that.

[8] notes on Page 33 and Page 36 that by default, Pig operates on your HDFS home directory for LOAD, so I only need to specify the name of the file. I refer to the storage syntax in [8] page 36 to store the result as tab delimited in an output directory.

[1] notes that the script can be executed simply by using a command like “pig [script_name]”

```
Caused by: org.apache.hadoop.ipc.RemoteException(org.apache.hadoop.security.AccessControlException): Permission denied: user=s1155160788, access=WRITE, inode="/output/_temporary/1":hdfs:hdf
s:drwxr-xr-x
        at org.apache.hadoop.hdfs.server.namenode.FSPermissionChecker.check(FSPermissionChecker.java:353)
        at org.apache.hadoop.hdfs.server.namenode.FSPermissionChecker.check(FSPermissionChecker.java:325)
        at org.apache.hadoop.hdfs.server.namenode.FSPermissionChecker.checkPermission(FSPermissionChecker.java:246)
        at org.apache.hadoop.hdfs.server.namenode.FSPermissionChecker.checkPermission(FSPermissionChecker.java:190)
        at org.apache.hadoop.hdfs.server.namenode.FSDirectory.checkPermission(FSDirectory.java:1950)
        at org.apache.hadoop.hdfs.server.namenode.FSDirectory.checkPermission(FSDirectory.java:1934)
        at org.apache.hadoop.hdfs.server.namenode.FSDirectory.checkAncestorAccess(FSDirectory.java:1917)
        at org.apache.hadoop.hdfs.server.namenode.FSDirMkdirOp.mkdirs(FSDirMkdirOp.java:71)
        at org.apache.hadoop.hdfs.server.namenode.FSNamesystem.mkdirs(FSNamesystem.java:4185)
        at org.apache.hadoop.hdfs.server.namenode.NameNodeRpcServer.mkdirs(NameNodeRpcServer.java:1109)
        at org.apache.hadoop.hdfs.protocolPB.ClientNamenodeProtocolServerSideTranslatorPB.mkdirs(ClientNamenodeProtocolServerSideTranslatorPB.java:645)
        at org.apache.hadoop.hdfs.protocol.proto.ClientNamenodeProtocolProtos$ClientNamenodeProtocol$2.callBlockingMethod(ClientNamenodeProtocolProtos.java)
        at org.apache.hadoop.ipc.ProtobufRpcEngine$Server$ProtoBufRpcInvoker.call(ProtobufRpcEngine.java:640)
        at org.apache.hadoop.ipc.RPC$Server.call(RPC.java:982)
        at org.apache.hadoop.ipc.Server$Handler$1.run(Server.java:2351)
        at org.apache.hadoop.ipc.Server$Handler$1.run(Server.java:2347)
        at java.security.AccessController.doPrivileged(Native Method)
        at javax.security.auth.Subject.doAs(Subject.java:422)
        at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1869)
        at org.apache.hadoop.ipc.Server$Handler.run(Server.java:2347)

        at org.apache.hadoop.ipc.Client.getRpcResponse(Client.java:1554)
        at org.apache.hadoop.ipc.Client.call(Client.java:1498)
        at org.apache.hadoop.ipc.Client.call(Client.java:1398)
        at org.apache.hadoop.ipc.ProtobufRpcEngine$Invoker.invoke(ProtobufRpcEngine.java:233)
        at com.sun.proxy.$Proxy12.mkdirs(Unknown Source)
        at org.apache.hadoop.hdfs.protocolPB.ClientNamenodeProtocolTranslatorPB.mkdirs(ClientNamenodeProtocolTranslatorPB.java:610)
        at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
        at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:62)
        at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
        at java.lang.reflect.Method.invoke(Method.java:498)
        at org.apache.hadoop.io.retry.RetryInvocationHandler.invokeMethod(RetryInvocationHandler.java:290)
        at org.apache.hadoop.io.retry.RetryInvocationHandler.invoke(RetryInvocationHandler.java:202)
        at org.apache.hadoop.io.retry.RetryInvocationHandler.invoke(RetryInvocationHandler.java:184)
        at com.sun.proxy.$Proxy13.mkdirs(Unknown Source)
        at org.apache.hadoop.hdfs.DFSClient.primitiveMkdir(DFSClient.java:3099)
        ... 35 more
, Vertex vertex_1644821811052_0559_1_01 [scope-30] killed/failed due to:INIT_FAILURE]
DAG did not succeed due to VERTEX_FAILURE. failedVertices:2 killedVertices:0
Details at logfile: /home/s1155160788/Assignment2/pig_1646028453060.log
2022-02-28 14:07:40,434 [main] INFO  org.apache.pig.Main - Pig script completed in 7 seconds and 477 milliseconds (7477 ms)
2022-02-28 14:07:40,434 [main] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezLauncher - Shutting down thread pool
2022-02-28 14:07:40,440 [pool-1-thread-1] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezSessionManager - Shutting down Tez session org.apache.tez.client.TezClient@61789894
2022-02-28 14:07:40 Shutting down Tez session , sessionName=PigLatin:q1.pig, applicationId=application_1644821811052_0559
2022-02-28 14:07:40,440 [pool-1-thread-1] INFO  org.apache.tez.client.TezClient - Shutting down Tez Session, sessionName=PigLatin:q1.pig, applicationId=application_1644821811052_0559
[s1155160788@dicvmd10 Assignment2]$
```

My first attempt leaves a long string of errors. I have omitted most of them. The errors seem to primarily have issue with the store command, saying I don’t have permission to create an /output directory. The most likely explanation is that while LOAD operates on the home HDFS directory, STORE needs the full HDFS path, and thus tried to store output in an output directory made from the root of the HDFS rather than from my user folder, which it did not have permission to do so.

Final Script / Code for Section B:

q1.pig

C: > Users > Eddie > Documents > School > CUHK > 2021 - 2022 Term 2 > IEMS 5730 - Big Data Systems and Information Processing > Week 7 > Assignment 2 > q1.pig

```
A = LOAD 'googlebooks-eng-all-1gram-20120701-a' USING PigStorage('\t') AS (bigram:chararray, year:int, match_count:int, volume_count:int);
B = LOAD 'googlebooks-eng-all-1gram-20120701-b' USING PigStorage('\t') AS (bigram:chararray, year:int, match_count:int, volume_count:int);
C = UNION A, B;
STORE C INTO '/user/s1155160788/output' USING PigStorage('\t');
```

Screenshot of output on next page.

```
[s1155160788@dicvmd10 Assignment2]$ pig q1.pig
22/02/28 14:11:26 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
22/02/28 14:11:26 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
22/02/28 14:11:26 INFO pig.ExecTypeProvider: Trying ExecType : TEZ_LOCAL
22/02/28 14:11:26 INFO pig.ExecTypeProvider: Trying ExecType : TEZ
22/02/28 14:11:26 INFO pig.ExecTypeProvider: Picked TEZ as the ExecType
2022-02-28 14:11:26,427 [main] INFO  org.apache.pig.Main - Apache Pig version 0.16.0.2.6.5.0-292 (rUnversioned directory) compiled May 11 2018, 07:56:28
2022-02-28 14:11:26,428 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/s1155160788/Assignment2/pig_1646028686426.log
2022-02-28 14:11:26,820 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/s1155160788/.pigbootup not found
2022-02-28 14:11:26,886 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://dicvmd2.ie.cuhk.edu.hk:8020
2022-02-28 14:11:27,212 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-q1.pig-76c82af5-fa73-4911-a5f0-35da919a0e56
2022-02-28 14:11:27,432 [main] INFO  org.apache.hadoop.yarn.client.api.impl.TimelineClientImpl - Timeline service address: http://dicvmd2.ie.cuhk.edu.hk:8188/ws/v1/timeline/
2022-02-28 14:11:27,496 [main] INFO  org.apache.pig.backend.hadoop.PigATSClient - Created ATS Hook
2022-02-28 14:11:27,954 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: UNION
2022-02-28 14:11:27,976 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2022-02-28 14:11:27,996 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelS
etter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCa
stInserter]}
2022-02-28 14:11:28,050 [main] INFO  org.apache.pig.impl.util.SpillableMemoryManager - Selected heap (PS Old Gen) of size 699400192 to monitor. collectionUsageThreshold = 489580128, usageTh
reshold = 489580128
2022-02-28 14:11:28,114 [main] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezLauncher - Tez staging directory is /tmp/s1155160788/staging and resources directory is /tmp/templ0
71818178
2022-02-28 14:11:28,140 [main] INFO  org.apache.pig.backend.hadoop.executionengine.tez.plan.TezCompiler - File concatenation threshold: 100 optimistic? false
2022-02-28 14:11:28,222 [main] INFO  org.apache.pig.builtin.PigStorage - Using PigTextInputFormat
2022-02-28 14:11:28,226 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2022-02-28 14:11:28,226 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
2022-02-28 14:11:28,255 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths (combined) to process : 14
2022-02-28 14:11:28,282 [main] INFO  org.apache.pig.builtin.PigStorage - Using PigTextInputFormat
2022-02-28 14:11:28,283 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2022-02-28 14:11:28,283 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
2022-02-28 14:11:28,286 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths (combined) to process : 10
2022-02-28 14:11:28,508 [main] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezJobCompiler - Local resource: joda-time-2.9.4.jar
2022-02-28 14:11:28,508 [main] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezJobCompiler - Local resource: antlr-runtime-3.4.jar
2022-02-28 14:11:28,508 [main] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezJobCompiler - Local resource: pig-0.16.0.2.6.5.0-292-core-h2.jar
2022-02-28 14:11:28,508 [main] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezJobCompiler - Local resource: automaton-1.11-8.jar
2022-02-28 14:11:28,611 [main] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezDagBuilder - For vertex - scope-30: parallelism=14, memory=65536, java opts=-XX:+PrintGCDetails -ve
rbose:gc -XX:+PrintGCTimeStamps -XX:+UseNUMA -XX:+UseG1GC -XX:+ResizeTLAB
2022-02-28 14:11:28,611 [main] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezDagBuilder - Processing aliases: A
2022-02-28 14:11:28,612 [main] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezDagBuilder - Detailed locations: A[1,4],A[-1,-1]
2022-02-28 14:11:28,612 [main] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezDagBuilder - Pig features in the vertex:
2022-02-28 14:11:28,664 [main] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezDagBuilder - For vertex - scope-31: parallelism=10, memory=65536, java opts=-XX:+PrintGCDetails -ve
rbose:gc -XX:+PrintGCTimeStamps -XX:+UseNUMA -XX:+UseG1GC -XX:+ResizeTLAB
2022-02-28 14:11:28,664 [main] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezDagBuilder - Processing aliases: B
2022-02-28 14:11:28,664 [main] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezDagBuilder - Detailed locations: B[2,4],B[-1,-1]
2022-02-28 14:11:28,664 [main] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezDagBuilder - Pig features in the vertex:
2022-02-28 14:11:28,693 [main] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezJobCompiler - Total estimated parallelism is 24
2022-02-28 14:11:28,733 [PigTezLauncher-0] INFO  org.apache.pig.tools.pigstats.tez.TezScriptState - Pig script settings are added to the job
2022-02-28 14:11:28,800 [PigTezLauncher-0] INFO  org.apache.tez.client.TezClient - Tez Client Version: [ component=tez-api, version=0.7.0.2.6.5.0-292, revision=2fcfb8aa6a462569e160756dfb735
14f99cbf248, SCM-URL=scm:git:https://git-wip-us.apache.org/repos/asf/tez.git, buildTime=2018-05-11T06:44:47Z ]
2022-02-28 14:11:28,893 [PigTezLauncher-0] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at dicvmd2.ie.cuhk.edu.hk/172.16.5.202:10200
```

```
2022-02-28 14:11:28,893 [PigTezLauncher-0] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at dicvmd2.ie.cuhk.edu.hk/172.16.5.202:10200
2022-02-28 14:11:28,895 [PigTezLauncher-0] INFO  org.apache.tez.client.TezClient - Session mode. Starting session.
2022-02-28 14:11:28,897 [PigTezLauncher-0] INFO  org.apache.tez.client.TezClientUtils - Using tez.lib.uris value from configuration: /hdp/apps/2.6.5.0-292/tez/tez.tar.gz
2022-02-28 14:11:28,909 [PigTezLauncher-0] INFO  org.apache.hadoop.yarn.client.RequestHedgingRMFailoverProxyProvider - Looking for the active RM in [rm1, rm2]...
2022-02-28 14:11:28,915 [PigTezLauncher-0] INFO  org.apache.hadoop.yarn.client.RequestHedgingRMFailoverProxyProvider - Found active RM [rm2]
2022-02-28 14:11:28,921 [PigTezLauncher-0] INFO  org.apache.tez.client.TezClient - Tez system stage directory hdfs://dicvmd2.ie.cuhk.edu.hk:8020/tmp/s1155160788/staging/.tez/application_164
4821811052_0560 doesn't exist and is created
2022-02-28 14:11:29,184 [PigTezLauncher-0] INFO  org.apache.hadoop.yarn.client.api.impl.YarnClientImpl - Submitted application application_1644821811052_0560
2022-02-28 14:11:29,185 [PigTezLauncher-0] INFO  org.apache.tez.client.TezClient - The url to track the Tez Session: http://dicvmd1.ie.cuhk.edu.hk:8088/proxy/application_1644821811052_0560/
2022-02-28 14:11:31,774 [PigTezLauncher-0] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezJob - Submitting DAG PigLatin:q1.pig-0_scope-0
2022-02-28 14:11:31,774 [PigTezLauncher-0] INFO  org.apache.tez.client.TezClient - Submitting dag to TezSession, sessionName=PigLatin:q1.pig, applicationId=application_1644821811052_0560, d
agName=PigLatin:q1.pig-0_scope-0, callerContext={ context=PIG, callerType=PIG_SCRIPT_ID, callerId=PIG-q1.pig-76c82af5-fa73-4911-a5f0-35da919a0e56 }
2022-02-28 14:11:32,027 [PigTezLauncher-0] INFO  org.apache.tez.client.TezClient - Submitted dag to TezSession, sessionName=PigLatin:q1.pig, applicationId=application_1644821811052_0560, da
gName=PigLatin:q1.pig-0_scope-0
2022-02-28 14:11:32,066 [PigTezLauncher-0] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at dicvmd2.ie.cuhk.edu.hk/172.16.5.202:10200
2022-02-28 14:11:32,068 [PigTezLauncher-0] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezJob - Submitted DAG PigLatin:q1.pig-0_scope-0. Application id: application_164482181105
2_0560
2022-02-28 14:11:32,069 [PigTezLauncher-0] INFO  org.apache.hadoop.yarn.client.RequestHedgingRMFailoverProxyProvider - Looking for the active RM in [rm1, rm2]...
2022-02-28 14:11:32,070 [PigTezLauncher-0] INFO  org.apache.hadoop.yarn.client.RequestHedgingRMFailoverProxyProvider - Found active RM [rm2]
2022-02-28 14:11:32,714 [main] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezLauncher - HadoopJobId: job_1644821811052_0560
2022-02-28 14:11:33,068 [Timer-0] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezJob - DAG Status: status=RUNNING, progress=TotalTasks: 24 Succeeded: 0 Running: 0 Failed: 0 Kill
ed: 0, diagnostics=, counters=null
2022-02-28 14:11:53,068 [Timer-0] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezJob - DAG Status: status=RUNNING, progress=TotalTasks: 24 Succeeded: 1 Running: 1 Failed: 0 Kill
ed: 0, diagnostics=, counters=null
2022-02-28 14:12:13,069 [Timer-0] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezJob - DAG Status: status=RUNNING, progress=TotalTasks: 24 Succeeded: 3 Running: 1 Failed: 0 Kill
ed: 0, diagnostics=, counters=null
2022-02-28 14:12:33,068 [Timer-0] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezJob - DAG Status: status=RUNNING, progress=TotalTasks: 24 Succeeded: 5 Running: 1 Failed: 0 Kill
ed: 0, diagnostics=, counters=null
2022-02-28 14:12:53,068 [Timer-0] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezJob - DAG Status: status=RUNNING, progress=TotalTasks: 24 Succeeded: 7 Running: 1 Failed: 0 Kill
ed: 0, diagnostics=, counters=null
2022-02-28 14:13:13,069 [Timer-0] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezJob - DAG Status: status=RUNNING, progress=TotalTasks: 24 Succeeded: 9 Running: 1 Failed: 0 Kill
ed: 0, diagnostics=, counters=null
2022-02-28 14:13:33,068 [Timer-0] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezJob - DAG Status: status=RUNNING, progress=TotalTasks: 24 Succeeded: 11 Running: 1 Failed: 0 Kil
led: 0, diagnostics=, counters=null
2022-02-28 14:13:53,068 [Timer-0] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezJob - DAG Status: status=RUNNING, progress=TotalTasks: 24 Succeeded: 13 Running: 1 Failed: 0 Kil
led: 0, diagnostics=, counters=null
2022-02-28 14:14:13,069 [Timer-0] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezJob - DAG Status: status=RUNNING, progress=TotalTasks: 24 Succeeded: 15 Running: 1 Failed: 0 Kil
led: 0, diagnostics=, counters=null
2022-02-28 14:14:33,068 [Timer-0] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezJob - DAG Status: status=RUNNING, progress=TotalTasks: 24 Succeeded: 16 Running: 1 Failed: 0 Kil
led: 0, diagnostics=, counters=null
2022-02-28 14:14:53,068 [Timer-0] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezJob - DAG Status: status=RUNNING, progress=TotalTasks: 24 Succeeded: 18 Running: 1 Failed: 0 Kil
led: 0, diagnostics=, counters=null
2022-02-28 14:15:13,068 [Timer-0] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezJob - DAG Status: status=RUNNING, progress=TotalTasks: 24 Succeeded: 20 Running: 1 Failed: 0 Kil
led: 0, diagnostics=, counters=null
2022-02-28 14:15:33,068 [Timer-0] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezJob - DAG Status: status=RUNNING, progress=TotalTasks: 24 Succeeded: 22 Running: 1 Failed: 0 Kil
led: 0, diagnostics=, counters=null
2022-02-28 14:15:45,526 [PigTezLauncher-0] INFO  org.apache.tez.common.counters.Limits - Counter limits initialized with parameters:  GROUP_NAME_MAX=256, MAX_GROUPS=3000, COUNTER_NAME_MAX=6
4, MAX_COUNTERS=10000
2022-02-28 14:15:45,529 [PigTezLauncher-0] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezJob - DAG Status: status=SUCCEEDED, progress=TotalTasks: 24 Succeeded: 24 Running: 0 Fa
iled: 0 Killed: 0, diagnostics=, counters=Counters: 29
```

```
iled: 0 Killed: 0, diagnostics=, counters=Counters: 29
        org.apache.tez.common.counters.DAGCounter
                NUM_SUCCEEDED_TASKS=24
                TOTAL_LAUNCHED_TASKS=24
                DATA_LOCAL_TASKS=9
                RACK_LOCAL_TASKS=15
                AM_CPU_MILLISECONDS=5740
                AM_GC_TIME_MILLIS=0
        File System Counters
                HDFS_BYTES_READ=3072802593
                HDFS_BYTES_WRITTEN=3069919009
                HDFS_READ_OPS=96
                HDFS_WRITE_OPS=48
                HDFS_OP_CREATE=24
                HDFS_OP_GET_FILE_STATUS=72
                HDFS_OP_OPEN=24
                HDFS_OP_RENAME=24
        org.apache.tez.common.counters.TaskCounter
                GC_TIME_MILLIS=1450
                CPU_MILLISECONDS=264010
                PHYSICAL_MEMORY_BYTES=51539607552
                VIRTUAL_MEMORY_BYTES=1399799275520
                COMMITTED_HEAP_BYTES=51539607552
                INPUT_RECORDS_PROCESSED=148170422
                INPUT_SPLIT_LENGTH_BYTES=3069919009
                OUTPUT_RECORDS=148170422
        MultiStoreCounters
                Output records in _0_output=148170422
        TaskCounter_scope_30_INPUT_scope_0
                INPUT_RECORDS_PROCESSED=86618505
                INPUT_SPLIT_LENGTH_BYTES=1801526075
        TaskCounter_scope_30_OUTPUT_scope_29
                OUTPUT_RECORDS=86618505
        TaskCounter_scope_31_INPUT_scope_14
                INPUT_RECORDS_PROCESSED=61551917
                INPUT_SPLIT_LENGTH_BYTES=1268392934
        TaskCounter_scope_31_OUTPUT_scope_29
                OUTPUT_RECORDS=61551917
2022-02-28 14:15:45,744 [main] INFO  org.apache.pig.tools.pigstats.tez.TezPigScriptStats - Script Statistics:

       HadoopVersion: 2.7.3.2.6.5.0-292
          PigVersion: 0.16.0.2.6.5.0-292
          TezVersion: 0.7.0.2.6.5.0-292
              UserId: s1155160788
            FileName: q1.pig
           StartedAt: 2022-02-28 14:11:28
          FinishedAt: 2022-02-28 14:15:45
            Features: UNION

Success!
```

```
Success!


DAG 0:
                                 Name: PigLatin:q1.pig-0_scope-0
                        ApplicationId: job_1644821811052_0560
                   TotalLaunchedTasks: 24
                        FileBytesRead: 0
                     FileBytesWritten: 0
                         HdfsBytesRead: 3072802593
                      HdfsBytesWritten: 3069919009
     SpillableMemoryManager spill count: 0
              Bags proactively spilled: 0
           Records proactively spilled: 0

DAG Plan:
Tez vertex scope-30     ->      Tez vertex group scope-36,
Tez vertex scope-31     ->      Tez vertex group scope-36,
Tez vertex group scope-36

Vertex Stats:
VertexId Parallelism TotalTasks   InputRecords   ReduceInputRecords  OutputRecords  FileBytesRead FileBytesWritten  HdfsBytesRead HdfsBytesWritten Alias        Feature Outputs
scope-30          14         14       86618505                    0       86618505              0               0     1803230011       1801526075 A            /user/s1155160788/output,
scope-31          10         10       61551917                    0       61551917              0               0     1269572582       1268392934 B            /user/s1155160788/output,

Input(s):
Successfully read 86618505 records (1803230011 bytes) from: "hdfs://dicvmd2.ie.cuhk.edu.hk:8020/user/s1155160788/googlebooks-eng-all-1gram-20120701-a"
Successfully read 61551917 records (1269572582 bytes) from: "hdfs://dicvmd2.ie.cuhk.edu.hk:8020/user/s1155160788/googlebooks-eng-all-1gram-20120701-b"

Output(s):
Successfully stored 148170422 records (3069919009 bytes) in: "/user/s1155160788/output"

2022-02-28 14:15:45,756 [main] INFO  org.apache.pig.Main - Pig script completed in 4 minutes, 19 seconds and 436 milliseconds (259436 ms)
2022-02-28 14:15:45,756 [main] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezLauncher - Shutting down thread pool
2022-02-28 14:15:45,767 [pool-1-thread-1] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezSessionManager - Shutting down Tez session org.apache.tez.client.TezClient@6b9f3cbb
2022-02-28 14:15:45 Shutting down Tez session , sessionName=PigLatin:q1.pig, applicationId=application_1644821811052_0560
2022-02-28 14:15:45,767 [pool-1-thread-1] INFO  org.apache.tez.client.TezClient - Shutting down Tez Session, sessionName=PigLatin:q1.pig, applicationId=application_1644821811052_0560
[s1155160788@dicvmd10 Assignment2]$
```

We get a successful output, and now have to process / validate it.

```
[s1155160788@dicvmd10 Assignment2]$ hdfs dfs -ls
Found 5 items
-rw-r--r--   3 s1155160788 student 1801526075 2022-02-26 09:25 googlebooks-eng-all-1gram-20120701-a
-rw-r--r--   3 s1155160788 student 1268392934 2022-02-26 09:25 googlebooks-eng-all-1gram-20120701-b
-rw-r--r--   3 s1155160788 student   17103777 2022-02-09 16:28 movielens_large_updated.csv
-rw-r--r--   3 s1155160788 student     893989 2022-02-09 16:28 movielens_small.csv
drwxr-xr-x   - s1155160788 student          0 2022-02-28 14:15 output
[s1155160788@dicvmd10 Assignment2]$ hdfs dfs -ls output
Found 25 items
-rw-r--r--   3 s1155160788 student          0 2022-02-28 14:15 output/_SUCCESS
-rw-r--r--   3 s1155160788 student  134217745 2022-02-28 14:13 output/part-v000-o000-r-00000
-rw-r--r--   3 s1155160788 student  134217712 2022-02-28 14:13 output/part-v000-o000-r-00001
-rw-r--r--   3 s1155160788 student  134217735 2022-02-28 14:13 output/part-v000-o000-r-00002
-rw-r--r--   3 s1155160788 student  134217731 2022-02-28 14:14 output/part-v000-o000-r-00003
-rw-r--r--   3 s1155160788 student  134217720 2022-02-28 14:11 output/part-v000-o000-r-00004
-rw-r--r--   3 s1155160788 student  134217744 2022-02-28 14:14 output/part-v000-o000-r-00005
-rw-r--r--   3 s1155160788 student  134217725 2022-02-28 14:12 output/part-v000-o000-r-00006
-rw-r--r--   3 s1155160788 student  134217740 2022-02-28 14:14 output/part-v000-o000-r-00007
-rw-r--r--   3 s1155160788 student  134217710 2022-02-28 14:12 output/part-v000-o000-r-00008
-rw-r--r--   3 s1155160788 student   60433372 2022-02-28 14:13 output/part-v000-o000-r-00009
-rw-r--r--   3 s1155160788 student  134217740 2022-02-28 14:11 output/part-v001-o000-r-00000
-rw-r--r--   3 s1155160788 student  134217728 2022-02-28 14:13 output/part-v001-o000-r-00001
-rw-r--r--   3 s1155160788 student  134217729 2022-02-28 14:14 output/part-v001-o000-r-00002
-rw-r--r--   3 s1155160788 student  134217726 2022-02-28 14:14 output/part-v001-o000-r-00003
-rw-r--r--   3 s1155160788 student  134217736 2022-02-28 14:12 output/part-v001-o000-r-00004
-rw-r--r--   3 s1155160788 student  134217720 2022-02-28 14:12 output/part-v001-o000-r-00005
-rw-r--r--   3 s1155160788 student  134217725 2022-02-28 14:12 output/part-v001-o000-r-00006
-rw-r--r--   3 s1155160788 student  134217735 2022-02-28 14:12 output/part-v001-o000-r-00007
-rw-r--r--   3 s1155160788 student  134217735 2022-02-28 14:14 output/part-v001-o000-r-00008
-rw-r--r--   3 s1155160788 student  134217726 2022-02-28 14:15 output/part-v001-o000-r-00009
-rw-r--r--   3 s1155160788 student  134217726 2022-02-28 14:15 output/part-v001-o000-r-00010
-rw-r--r--   3 s1155160788 student  134217716 2022-02-28 14:15 output/part-v001-o000-r-00011
-rw-r--r--   3 s1155160788 student  134217728 2022-02-28 14:15 output/part-v001-o000-r-00012
-rw-r--r--   3 s1155160788 student   56695605 2022-02-28 14:15 output/part-v001-o000-r-00013
[s1155160788@dicvmd10 Assignment2]$
```

I copy the output from HDFS to local then use cat to combine the output into one final.

Using filezilla, I add up the sizes of the original datasets and ensure the size of the union output is the same as the original. 1,801,526,075 bytes (Dataset A) + 1,268,392,934 (Dataset B) = 3,069,919,009 bytes, which is exactly what we get from the union.

Question 1 Section C: Calculate average number of occurrences per year for each unique bigram

I first perform some clean up and take the unified version of the bigram datasets and upload it to the HDFS with the name ‘united_books’, then delete the output directory with the 14 original parts. This will simplify loading the stored union output in the next script.

Since the bytes of the union are exactly the same as the sum of the set, and based on what I’ve read about UNION, we can assume that the totals weren’t summed, and that the tables were just concatenated together.

Attempt #1:

We should not assume that the bigrams in each dataset are unique. It is possible that 1: Both datasets have some bigrams in common and 2: Both datasets have some years in common for the same bigram.

Our script needs to first group all the entries that share the same bigram. For each bigram, we then need to either group all items from the same year or determine the number of unique years for that bigram. Afterwards, we can calculate the average by summing up the match_count and dividing by the number of unique year entries under the bigram.

I got a number of syntax errors and was debugging for awhile when trying to calculate this, so I decided to run the pig script at an intermediate stage after grouping by bigram, determining the total number of unique years for the bigram, and calculating the sum of the match_count for that bigram.

C: > Users > Eddie > Documents > School > CUHK > 2021 - 2022 Term 2 > IEMS 5730 - Big Data Systems and Information Processing > Week 7 > Assignment 2 > q1c.pig

```
C = LOAD 'united_books' USING PigStorage('\t') AS (bigram:chararray, year:int, match_count:int, volume_count:int);
Bigram_Grouped = GROUP C BY bigram;
Bigram_Summed = FOREACH Bigram_Grouped {
    distinct_years = DISTINCT C.year;
    GENERATE group as bigram, COUNT(distinct_years), SUM(C.match_count);
};
STORE Bigram_Summed INTO '/user/s1155160788/output' USING PigStorage('\t');
```

The generation and storing of this intermediate output took 14.5 minutes. I realized it would be costly to wait every time I was evaluating the input, and learned about the illustrate and describe commands to help debugging in Pig in [9].

I learned more about groups in [10], and the idea on how to get the unique number of years for each bigram came from [11]. Basically, there are two statements in the 4 each. The first one uses DISTINCT to remove all items with duplicate years. This doesn't actually remove them, just creates the subset and loads it into distinct_years. It doesn't matter which of the duplicates are removed because it will leave 1 tuple for each year, and we are just using a COUNT of the distinct_years variable to get the total number of years the bigram appears. Afterwards, sum will sum the match_count.

Screenshot on next page.

174316	buoyline	42	124
174317	bupposed	51	55
174318	bur'glar	33	59
174319	bur_VERB	353	47965
174320	buradaki	40	161
174321	buraka_X	32	167
174322	buranjis	54	2439
174323	burb_ADJ	43	170
174324	burbank_	22	40
174325	burberry	110	1584
174326	burbidge	37	107
174327	burbujas	73	884

buoyline_NOUN	1868	4	2
buoyline_NOUN	1875	4	2
buoyline_NOUN	1876	1	1
buoyline_NOUN	1880	1	1
buoyline_NOUN	1897	1	1
buoyline_NOUN	1898	4	4
buoyline_NOUN	1905	6	3
buoyline_NOUN	1934	1	1
buoyline_NOUN	1939	1	1
buoyline_NOUN	1945	2	2
buoyline_NOUN	1955	2	1
buoyline_NOUN	1962	1	1
buoyline_NOUN	1964	12	2
buoyline_NOUN	1965	2	1
buoyline_NOUN	1966	2	1
buoyline_NOUN	1967	5	3
buoyline_NOUN	1971	5	2
buoyline_NOUN	1974	3	2
buoyline_NOUN	1976	1	1
buoyline_NOUN	1978	1	1
buoyline_NOUN	1981	2	2
buoyline_NOUN	1982	6	2
buoyline_NOUN	1985	1	1
buoyline_NOUN	1987	13	2
buoyline_NOUN	1989	6	4
buoyline_NOUN	1993	2	2
buoyline_NOUN	1994	1	1
buoyline_NOUN	1997	2	1
buoyline_NOUN	1998	1	1
buoyline_NOUN	1999	1	1
buoyline_NOUN	2002	1	1
buoyline_NOUN	2005	1	1
buoyline_NOUN	2006	1	1
buoyline_NOUN	2007	1	1
buoyline_NOUN	2008	6	4

buoyline	1868	4	2
buoyline	1875	4	2
buoyline	1876	1	1
buoyline	1880	1	1
buoyline	1897	1	1
buoyline	1898	4	4
buoyline	1905	6	3
buoyline	1918	1	1
buoyline	1934	1	1
buoyline	1939	1	1
buoyline	1945	2	2
buoyline	1953	1	1
buoyline	1955	2	1
buoyline	1962	1	1
buoyline	1964	12	2
buoyline	1965	2	1
buoyline	1966	2	1
buoyline	1967	5	3
buoyline	1971	5	2
buoyline	1974	4	2
buoyline	1976	1	1
buoyline	1978	1	1
buoyline	1981	2	2
buoyline	1982	7	2
buoyline	1984	2	2
buoyline	1985	2	1
buoyline	1986	1	1
buoyline	1987	18	4
buoyline	1989	6	4
buoyline	1992	1	1
buoyline	1993	2	2
buoyline	1994	3	2
buoyline	1995	1	1
buoyline	1997	2	1
buoyline	1998	1	1
buoyline	1999	1	1
buoyline	2000	1	1
buoyline	2002	1	1
buoyline	2005	1	1
buoyline	2006	3	1
buoyline	2007	1	1
buoyline	2008	6	4

Here, I take two examples to validate the intermediate input. The first screenshot has my intermediate output, the bigram, the number of unique years it appears in, and the number of times it appears. I take buoyline and buoyline_NOUN as examples and confirm manually that buoyline appears 42 times and has a sum count of 124, and that buoyline_NOUN appears 35 times and has a sum count of 104.

From the data so far, it doesn't seem that the datasets share any bigrams, and thus do not share any years either. It thus seems unnecessary for me to have coded the program such that I count the number of unique years per bigram.

```
], TaskAttempt 3 failed, info=[Error: Failure while running task:org.apache.pig.backend.executionengine.ExecException: ERROR 0: Scalar has more than one row in the output. 1st : (A!,293,375
51), 2nd :(A2,256,1336353) (common cause: "JOIN" then "FOREACH ... GENERATE foo.bar" should be "foo::bar" )
        at org.apache.pig.backend.hadoop.executionengine.tez.plan.operator.POValueOutputTez.getNextTuple(POValueOutputTez.java:167)
        at org.apache.pig.backend.hadoop.executionengine.tez.runtime.PigProcessor.runPipeline(PigProcessor.java:376)
        at org.apache.pig.backend.hadoop.executionengine.tez.runtime.PigProcessor.run(PigProcessor.java:241)
        at org.apache.tez.runtime.LogicalIOProcessorRuntimeTask.run(LogicalIOProcessorRuntimeTask.java:347)
        at org.apache.tez.runtime.task.TezTaskRunner$TaskRunnerCallable$1.run(TezTaskRunner.java:194)
        at org.apache.tez.runtime.task.TezTaskRunner$TaskRunnerCallable$1.run(TezTaskRunner.java:185)
        at java.security.AccessController.doPrivileged(Native Method)
        at javax.security.auth.Subject.doAs(Subject.java:422)
        at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1869)
        at org.apache.tez.runtime.task.TezTaskRunner$TaskRunnerCallable.callInternal(TezTaskRunner.java:185)
        at org.apache.tez.runtime.task.TezTaskRunner$TaskRunnerCallable.callInternal(TezTaskRunner.java:181)
        at org.apache.tez.common.CallableWithNdc.call(CallableWithNdc.java:36)
        at java.util.concurrent.FutureTask.run(FutureTask.java:266)
        at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1142)
        at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:617)
        at java.lang.Thread.run(Thread.java:745)
]], Vertex did not succeed due to OWN_TASK_FAILURE, failedTasks:1 killedTasks:16, Vertex vertex_1644821811052_0576_1_02 [scope-55] killed/failed due to:OWN_TASK_FAILURE]
Vertex killed, vertexName=scope-57, vertexId=vertex_1644821811052_0576_1_03, diagnostics=[Vertex received Kill while in RUNNING state., Vertex did not succeed due to OTHER_VERTEX_FAILURE, f
ailedTasks:0 killedTasks:17, Vertex vertex_1644821811052_0576_1_03 [scope-57] killed/failed due to:OTHER_VERTEX_FAILURE]
DAG did not succeed due to VERTEX_FAILURE. failedVertices:1 killedVertices:1
Details at logfile: /home/s1155160788/Assignment2/pig_1646049308266.log
2022-02-28 20:09:51,826 [main] INFO  org.apache.pig.Main - Pig script completed in 14 minutes, 43 seconds and 669 milliseconds (883669 ms)
2022-02-28 20:09:51,827 [main] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezLauncher - Shutting down thread pool
2022-02-28 20:09:51,834 [pool-1-thread-1] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezSessionManager - Shutting down Tez session org.apache.tez.client.TezClient@36154b8a
2022-02-28 20:09:51 Shutting down Tez session , sessionName=PigLatin:q1c.pig, applicationId=application_1644821811052_0576
2022-02-28 20:09:51,835 [pool-1-thread-1] INFO  org.apache.tez.client.TezClient - Shutting down Tez Session, sessionName=PigLatin:q1c.pig, applicationId=application_1644821811052_0576
[s1155160788@dicvmd10 Assignment2]$
```

I started getting errors halfway through with about 14 minutes of calculation. I asked the WeChat group and received responses from Qi Jiadun, Liu Yaohui, and someone else with a Chinese name that they believe there is no shared bigrams between the two datasets. With this in mind, I can greatly simplify the code.

Final Part C Script / Code:

```
D = LOAD 'united_books' USING PigStorage('\t') AS (bigram:chararray, year:int, match_count:float, volume_count:int);
Bigram_Grouped = GROUP D BY bigram;
Bigram_Averaged = FOREACH Bigram_Grouped GENERATE group as bigram, AVG(D.match_count) as average;
STORE Bigram_Averaged INTO '/user/s1155160788/output' USING PigStorage('\t');
```

After debugging some minor syntactical errors, I generate this code by referencing [12] slide 31 for the proper notation of the AVG aggregation function.

```
[s1155160788@dicvmd10 Assignment2]$ pig q1c.pig
22/02/28 20:21:46 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
22/02/28 20:21:46 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
22/02/28 20:21:46 INFO pig.ExecTypeProvider: Trying ExecType : TEZ_LOCAL
22/02/28 20:21:46 INFO pig.ExecTypeProvider: Trying ExecType : TEZ
22/02/28 20:21:46 INFO pig.ExecTypeProvider: Picked TEZ as the ExecType
2022-02-28 20:21:46,596 [main] INFO  org.apache.pig.Main - Apache Pig version 0.16.0.2.6.5.0-292 (rUnversioned directory) compiled May 11 2018, 07:56:28
2022-02-28 20:21:46,596 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/s1155160788/Assignment2/pig_1646050906595.log
2022-02-28 20:21:46,981 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/s1155160788/.pigbootup not found
2022-02-28 20:21:47,048 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://dicvmd2.ie.cuhk.edu.hk:8020
2022-02-28 20:21:47,365 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-q1c.pig-6e6e863c-ab12-43c5-8db4-0fb17f788781
2022-02-28 20:21:47,581 [main] INFO  org.apache.hadoop.yarn.client.api.impl.TimelineClientImpl - Timeline service address: http://dicvmd2.ie.cuhk.edu.hk:8188/ws/v1/timeline/
2022-02-28 20:21:47,648 [main] INFO  org.apache.pig.backend.hadoop.PigATSClient - Created ATS Hook
2022-02-28 20:21:48,097 [main] ERROR org.apache.pig.tools.grunt.Grunt - ERROR 1025:
<file q1c.pig, line 3, column 71> Invalid field projection. Projected field [match_count] does not exist in schema: group:chararray,D:bag{:tuple(bigram:chararray,year:int,match_count:float,
volume_count:int)}.
Details at logfile: /home/s1155160788/Assignment2/pig_1646050906595.log
2022-02-28 20:21:48,107 [main] INFO  org.apache.pig.Main - Pig script completed in 1 second and 617 milliseconds (1617 ms)
[s1155160788@dicvmd10 Assignment2]$ chmod +x q1c.pig
[s1155160788@dicvmd10 Assignment2]$ pig q1c.pig
22/02/28 20:23:54 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
22/02/28 20:23:54 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
22/02/28 20:23:54 INFO pig.ExecTypeProvider: Trying ExecType : TEZ_LOCAL
22/02/28 20:23:54 INFO pig.ExecTypeProvider: Trying ExecType : TEZ
22/02/28 20:23:54 INFO pig.ExecTypeProvider: Picked TEZ as the ExecType
2022-02-28 20:23:54,349 [main] INFO  org.apache.pig.Main - Apache Pig version 0.16.0.2.6.5.0-292 (rUnversioned directory) compiled May 11 2018, 07:56:28
2022-02-28 20:23:54,349 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/s1155160788/Assignment2/pig_1646051034348.log
2022-02-28 20:23:54,743 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/s1155160788/.pigbootup not found
2022-02-28 20:23:54,813 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://dicvmd2.ie.cuhk.edu.hk:8020
2022-02-28 20:23:55,139 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-q1c.pig-bda297cf-5887-45e9-9f03-de07c495d818
2022-02-28 20:23:55,358 [main] INFO  org.apache.hadoop.yarn.client.api.impl.TimelineClientImpl - Timeline service address: http://dicvmd2.ie.cuhk.edu.hk:8188/ws/v1/timeline/
2022-02-28 20:23:55,425 [main] INFO  org.apache.pig.backend.hadoop.PigATSClient - Created ATS Hook
2022-02-28 20:23:55,870 [main] ERROR org.apache.pig.tools.grunt.Grunt - ERROR 1025:
<file q1c.pig, line 3, column 80> Invalid field projection. Projected field [match_count] does not exist in schema: group:chararray,D:bag{:tuple(bigram:chararray,year:int,match_count:float,
volume_count:int)}.
Details at logfile: /home/s1155160788/Assignment2/pig_1646051034348.log
2022-02-28 20:23:55,880 [main] INFO  org.apache.pig.Main - Pig script completed in 1 second and 638 milliseconds (1638 ms)
[s1155160788@dicvmd10 Assignment2]$ chmod +x q1c.pig
[s1155160788@dicvmd10 Assignment2]$ pig q1c.pig
22/02/28 20:26:03 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
22/02/28 20:26:03 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
22/02/28 20:26:03 INFO pig.ExecTypeProvider: Trying ExecType : TEZ_LOCAL
22/02/28 20:26:03 INFO pig.ExecTypeProvider: Trying ExecType : TEZ
22/02/28 20:26:03 INFO pig.ExecTypeProvider: Picked TEZ as the ExecType
2022-02-28 20:26:03,106 [main] INFO  org.apache.pig.Main - Apache Pig version 0.16.0.2.6.5.0-292 (rUnversioned directory) compiled May 11 2018, 07:56:28
2022-02-28 20:26:03,106 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/s1155160788/Assignment2/pig_1646051163105.log
2022-02-28 20:26:03,500 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/s1155160788/.pigbootup not found
2022-02-28 20:26:03,566 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://dicvmd2.ie.cuhk.edu.hk:8020
2022-02-28 20:26:03,882 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-q1c.pig-7f5e208c-2a16-4099-ac9f-6a66d3f05f3b
2022-02-28 20:26:04,104 [main] INFO  org.apache.hadoop.yarn.client.api.impl.TimelineClientImpl - Timeline service address: http://dicvmd2.ie.cuhk.edu.hk:8188/ws/v1/timeline/
```

```
2022-02-28 20:26:04,104 [main] INFO  org.apache.hadoop.yarn.client.api.impl.TimelineClientImpl - Timeline service address: http://dicvmd2.ie.cuhk.edu.hk:8188/ws/v1/timeline/
2022-02-28 20:26:04,175 [main] INFO  org.apache.pig.backend.hadoop.PigATSClient - Created ATS Hook
2022-02-28 20:26:04,649 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: GROUP_BY
2022-02-28 20:26:04,670 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2022-02-28 20:26:04,696 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelS
etter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCa
stInserter]}
2022-02-28 20:26:04,734 [main] INFO  org.apache.pig.impl.util.SpillableMemoryManager - Selected heap (PS Old Gen) of size 699400192 to monitor. collectionUsageThreshold = 489580128, usageTh
reshold = 489580128
2022-02-28 20:26:04,788 [main] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezLauncher - Tez staging directory is /tmp/s1155160788/staging and resources directory is /tmp/temp21
14518043
2022-02-28 20:26:04,811 [main] INFO  org.apache.pig.backend.hadoop.executionengine.tez.plan.TezCompiler - File concatenation threshold: 100 optimistic? false
2022-02-28 20:26:04,830 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.CombinerOptimizerUtil - Choosing to move algebraic foreach to combiner
2022-02-28 20:26:04,880 [main] INFO  org.apache.pig.builtin.PigStorage - Using PigTextInputFormat
2022-02-28 20:26:04,883 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2022-02-28 20:26:04,883 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
2022-02-28 20:26:04,913 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths (combined) to process : 23
2022-02-28 20:26:05,138 [main] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezJobCompiler - Local resource: joda-time-2.9.4.jar
2022-02-28 20:26:05,138 [main] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezJobCompiler - Local resource: antlr-runtime-3.4.jar
2022-02-28 20:26:05,138 [main] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezJobCompiler - Local resource: pig-0.16.0.2.6.5.0-292-core-h2.jar
2022-02-28 20:26:05,138 [main] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezJobCompiler - Local resource: automaton-1.11-8.jar
2022-02-28 20:26:05,242 [main] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezDagBuilder - For vertex - scope-26: parallelism=23, memory=65536, java opts=-XX:+PrintGCDetails -ve
rbose:gc -XX:+PrintGCTimeStamps -XX:+UseNUMA -XX:+UseG1GC -XX:+ResizeTLAB
2022-02-28 20:26:05,242 [main] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezDagBuilder - Processing aliases: Bigram_Averaged,Bigram_Grouped,D
2022-02-28 20:26:05,242 [main] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezDagBuilder - Detailed locations: D[1,4],D[-1,-1],Bigram_Averaged[3,18],Bigram_Grouped[2,17]
2022-02-28 20:26:05,242 [main] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezDagBuilder - Pig features in the vertex:
2022-02-28 20:26:05,288 [main] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezDagBuilder - Set auto parallelism for vertex scope-27
2022-02-28 20:26:05,288 [main] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezDagBuilder - For vertex - scope-27: parallelism=17, memory=65536, java opts=-XX:+PrintGCDetails -ve
rbose:gc -XX:+PrintGCTimeStamps -XX:+UseNUMA -XX:+UseG1GC -XX:+ResizeTLAB
2022-02-28 20:26:05,288 [main] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezDagBuilder - Processing aliases: Bigram_Averaged
2022-02-28 20:26:05,288 [main] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezDagBuilder - Detailed locations: Bigram_Averaged[3,18]
2022-02-28 20:26:05,288 [main] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezDagBuilder - Pig features in the vertex: GROUP_BY
2022-02-28 20:26:05,332 [main] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezJobCompiler - Total estimated parallelism is 40
2022-02-28 20:26:05,371 [PigTezLauncher-0] INFO  org.apache.pig.tools.pigstats.tez.TezScriptState - Pig script settings are added to the job
2022-02-28 20:26:05,439 [PigTezLauncher-0] INFO  org.apache.tez.client.TezClient - Tez Client Version: [ component=tez-api, version=0.7.0.2.6.5.0-292, revision=2fcfb8aa6a462569e160756dfb735
14f99cbf248, SCM-URL=scm:git:https://git-wip-us.apache.org/repos/asf/tez.git, buildTime=2018-05-11T06:44:47Z ]
2022-02-28 20:26:05,532 [PigTezLauncher-0] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at dicvmd2.ie.cuhk.edu.hk/172.16.5.202:10200
2022-02-28 20:26:05,535 [PigTezLauncher-0] INFO  org.apache.tez.client.TezClient - Session mode. Starting session.
2022-02-28 20:26:05,537 [PigTezLauncher-0] INFO  org.apache.tez.client.TezClientUtils - Using tez.lib.uris value from configuration: /hdp/apps/2.6.5.0-292/tez/tez.tar.gz
2022-02-28 20:26:05,549 [PigTezLauncher-0] INFO  org.apache.hadoop.yarn.client.RequestHedgingRMFailoverProxyProvider - Looking for the active RM in [rm1, rm2]...
2022-02-28 20:26:05,555 [PigTezLauncher-0] INFO  org.apache.hadoop.yarn.client.RequestHedgingRMFailoverProxyProvider - Found active RM [rm2]
2022-02-28 20:26:05,560 [PigTezLauncher-0] INFO  org.apache.tez.client.TezClient - Tez system stage directory hdfs://dicvmd2.ie.cuhk.edu.hk:8020/tmp/s1155160788/staging/.tez/application_164
4821811052_0578 doesn't exist and is created
2022-02-28 20:26:05,623 [PigTezLauncher-0] INFO  org.apache.hadoop.yarn.client.api.impl.YarnClientImpl - Submitted application application_1644821811052_0578
2022-02-28 20:26:05,625 [PigTezLauncher-0] INFO  org.apache.tez.client.TezClient - The url to track the Tez Session: http://dicvmd1.ie.cuhk.edu.hk:8088/proxy/application_1644821811052_0578/
2022-02-28 20:26:08,202 [PigTezLauncher-0] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezJob - Submitting DAG PigLatin:q1c.pig-0_scope-0
2022-02-28 20:26:08,202 [PigTezLauncher-0] INFO  org.apache.tez.client.TezClient - Submitting dag to TezSession, sessionName=PigLatin:q1c.pig, applicationId=application_1644821811052_0578,
dagName=PigLatin:q1c.pig-0_scope-0, callerContext={ context=PIG, callerType=PIG_SCRIPT_ID, callerId=PIG-q1c.pig-7f5e208c-2a16-4099-ac9f-6a66d3f05f3b }
2022-02-28 20:26:08,457 [PigTezLauncher-0] INFO  org.apache.tez.client.TezClient - Submitted dag to TezSession, sessionName=PigLatin:q1c.pig, applicationId=application_1644821811052_0578, d
agName=PigLatin:q1c.pig-0_scope-0
```

```
agName=PigLatin:q1c.pig-0_scope-0
2022-02-28 20:26:08,506 [PigTezLauncher-0] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at dicvmd2.ie.cuhk.edu.hk/172.16.5.202:10200
2022-02-28 20:26:08,508 [PigTezLauncher-0] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezJob - Submitted DAG PigLatin:q1c.pig-0_scope-0. Application id: application_16448218110
52_0578
2022-02-28 20:26:08,509 [PigTezLauncher-0] INFO  org.apache.hadoop.yarn.client.RequestHedgingRMFailoverProxyProvider - Looking for the active RM in [rm1, rm2]...
2022-02-28 20:26:08,510 [PigTezLauncher-0] INFO  org.apache.hadoop.yarn.client.RequestHedgingRMFailoverProxyProvider - Found active RM [rm2]
2022-02-28 20:26:09,350 [main] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezLauncher - HadoopJobId: job_1644821811052_0578
2022-02-28 20:26:09,508 [Timer-0] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezJob - DAG Status: status=RUNNING, progress=TotalTasks: 40 Succeeded: 0 Running: 0 Failed: 0 Kill
ed: 0, diagnostics=, counters=null
2022-02-28 20:26:29,508 [Timer-0] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezJob - DAG Status: status=RUNNING, progress=TotalTasks: 40 Succeeded: 0 Running: 1 Failed: 0 Kill
ed: 0, diagnostics=, counters=null
2022-02-28 20:26:49,508 [Timer-0] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezJob - DAG Status: status=RUNNING, progress=TotalTasks: 40 Succeeded: 1 Running: 1 Failed: 0 Kill
ed: 0, diagnostics=, counters=null
2022-02-28 20:27:09,508 [Timer-0] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezJob - DAG Status: status=RUNNING, progress=TotalTasks: 40 Succeeded: 2 Running: 1 Failed: 0 Kill
ed: 0, diagnostics=, counters=null
2022-02-28 20:27:29,508 [Timer-0] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezJob - DAG Status: status=RUNNING, progress=TotalTasks: 40 Succeeded: 4 Running: 1 Failed: 0 Kill
ed: 0, diagnostics=, counters=null
2022-02-28 20:27:49,508 [Timer-0] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezJob - DAG Status: status=RUNNING, progress=TotalTasks: 40 Succeeded: 5 Running: 1 Failed: 0 Kill
ed: 0, diagnostics=, counters=null
2022-02-28 20:28:09,509 [Timer-0] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezJob - DAG Status: status=RUNNING, progress=TotalTasks: 29 Succeeded: 6 Running: 1 Failed: 0 Kill
ed: 0, diagnostics=, counters=null
2022-02-28 20:28:29,508 [Timer-0] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezJob - DAG Status: status=RUNNING, progress=TotalTasks: 29 Succeeded: 7 Running: 1 Failed: 0 Kill
ed: 0, diagnostics=, counters=null
2022-02-28 20:28:49,508 [Timer-0] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezJob - DAG Status: status=RUNNING, progress=TotalTasks: 29 Succeeded: 8 Running: 1 Failed: 0 Kill
ed: 0, diagnostics=, counters=null
2022-02-28 20:29:09,508 [Timer-0] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezJob - DAG Status: status=RUNNING, progress=TotalTasks: 29 Succeeded: 9 Running: 1 Failed: 0 Kill
ed: 0, diagnostics=, counters=null
2022-02-28 20:29:29,509 [Timer-0] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezJob - DAG Status: status=RUNNING, progress=TotalTasks: 29 Succeeded: 10 Running: 1 Failed: 0 Kil
led: 0, diagnostics=, counters=null
2022-02-28 20:29:49,509 [Timer-0] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezJob - DAG Status: status=RUNNING, progress=TotalTasks: 29 Succeeded: 11 Running: 1 Failed: 0 Kil
led: 0, diagnostics=, counters=null
2022-02-28 20:30:09,509 [Timer-0] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezJob - DAG Status: status=RUNNING, progress=TotalTasks: 29 Succeeded: 12 Running: 1 Failed: 0 Kil
led: 0, diagnostics=, counters=null
2022-02-28 20:30:29,509 [Timer-0] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezJob - DAG Status: status=RUNNING, progress=TotalTasks: 29 Succeeded: 13 Running: 1 Failed: 0 Kil
led: 0, diagnostics=, counters=null
2022-02-28 20:31:09,509 [Timer-0] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezJob - DAG Status: status=RUNNING, progress=TotalTasks: 29 Succeeded: 14 Running: 1 Failed: 0 Kil
led: 0, diagnostics=, counters=null
2022-02-28 20:31:29,509 [Timer-0] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezJob - DAG Status: status=RUNNING, progress=TotalTasks: 29 Succeeded: 15 Running: 1 Failed: 0 Kil
led: 0, diagnostics=, counters=null
2022-02-28 20:31:49,509 [Timer-0] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezJob - DAG Status: status=RUNNING, progress=TotalTasks: 29 Succeeded: 16 Running: 1 Failed: 0 Kil
led: 0, diagnostics=, counters=null
2022-02-28 20:32:09,510 [Timer-0] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezJob - DAG Status: status=RUNNING, progress=TotalTasks: 29 Succeeded: 17 Running: 1 Failed: 0 Kil
led: 0, diagnostics=, counters=null
2022-02-28 20:32:29,510 [Timer-0] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezJob - DAG Status: status=RUNNING, progress=TotalTasks: 29 Succeeded: 18 Running: 1 Failed: 0 Kil
led: 0, diagnostics=, counters=null
2022-02-28 20:32:49,510 [Timer-0] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezJob - DAG Status: status=RUNNING, progress=TotalTasks: 29 Succeeded: 19 Running: 1 Failed: 0 Kil
led: 0, diagnostics=, counters=null
2022-02-28 20:33:09,510 [Timer-0] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezJob - DAG Status: status=RUNNING, progress=TotalTasks: 29 Succeeded: 20 Running: 1 Failed: 0 Kil
led: 0, diagnostics=, counters=null
2022-02-28 20:33:29,511 [Timer-0] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezJob - DAG Status: status=RUNNING, progress=TotalTasks: 29 Succeeded: 21 Running: 1 Failed: 0 Kil
```

```
led: 0, diagnostics=, counters=null
2022-02-28 20:33:29,511 [Timer-0] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezJob - DAG Status: status=RUNNING, progress=TotalTasks: 29 Succeeded: 21 Running: 1 Failed: 0 Kil
led: 0, diagnostics=, counters=null
2022-02-28 20:33:49,511 [Timer-0] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezJob - DAG Status: status=RUNNING, progress=TotalTasks: 29 Succeeded: 22 Running: 1 Failed: 0 Kil
led: 0, diagnostics=, counters=null
2022-02-28 20:34:09,511 [Timer-0] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezJob - DAG Status: status=RUNNING, progress=TotalTasks: 29 Succeeded: 28 Running: 1 Failed: 0 Kil
led: 0, diagnostics=, counters=null
2022-02-28 20:34:09,991 [PigTezLauncher-0] INFO  org.apache.tez.common.counters.Limits - Counter limits initialized with parameters:  GROUP_NAME_MAX=256, MAX_GROUPS=3000, COUNTER_NAME_MAX=6
4, MAX_COUNTERS=10000
2022-02-28 20:34:09,994 [PigTezLauncher-0] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezJob - DAG Status: status=SUCCEEDED, progress=TotalTasks: 29 Succeeded: 29 Running: 0 Fa
iled: 0 Killed: 0, diagnostics=, counters=Counters: 102
        org.apache.tez.common.counters.DAGCounter
                NUM_SUCCEEDED_TASKS=29
                TOTAL_LAUNCHED_TASKS=29
                DATA_LOCAL_TASKS=11
                RACK_LOCAL_TASKS=12
                AM_CPU_MILLISECONDS=7060
                AM_GC_TIME_MILLIS=0
        File System Counters
                FILE_BYTES_READ=50617332
                FILE_BYTES_WRITTEN=37629450
                HDFS_BYTES_READ=3072802593
                HDFS_BYTES_WRITTEN=68316032
                HDFS_READ_OPS=41
                HDFS_WRITE_OPS=12
                HDFS_OP_CREATE=6
                HDFS_OP_GET_FILE_STATUS=18
                HDFS_OP_OPEN=23
                HDFS_OP_RENAME=6
        org.apache.tez.common.counters.TaskCounter
                REDUCE_INPUT_GROUPS=2458721
                REDUCE_INPUT_RECORDS=2458743
                COMBINE_INPUT_RECORDS=0
                SPILLED_RECORDS=4917486
                NUM_SHUFFLED_INPUTS=391
                NUM_SKIPPED_INPUTS=0
                NUM_FAILED_SHUFFLE_INPUTS=0
                MERGED_MAP_OUTPUTS=391
                GC_TIME_MILLIS=828
                CPU_MILLISECONDS=1205140
                PHYSICAL_MEMORY_BYTES=1182097473536
                VIRTUAL_MEMORY_BYTES=1694166802432
                COMMITTED_HEAP_BYTES=1182097473536
                INPUT_RECORDS_PROCESSED=148170422
                INPUT_SPLIT_LENGTH_BYTES=3069919009
                OUTPUT_RECORDS=150629143
                OUTPUT_BYTES=4023990818
                OUTPUT_BYTES_WITH_OVERHEAD=75404318
                OUTPUT_BYTES_PHYSICAL=37619882
                ADDITIONAL_SPILLS_BYTES_WRITTEN=0
```

Continued on next page due to sizing.

```
                ADDITIONAL_SPILLS_BYTES_WRITTEN=0
                ADDITIONAL_SPILLS_BYTES_READ=37619882
                ADDITIONAL_SPILL_COUNT=0
                SHUFFLE_CHUNK_COUNT=23
                SHUFFLE_BYTES=37619882
                SHUFFLE_BYTES_DECOMPRESSED=75404318
                SHUFFLE_BYTES_TO_MEM=0
                SHUFFLE_BYTES_TO_DISK=0
                SHUFFLE_BYTES_DISK_DIRECT=37619882
                NUM_MEM_TO_DISK_MERGES=0
                NUM_DISK_TO_DISK_MERGES=0
                SHUFFLE_PHASE_TIME=106
                MERGE_PHASE_TIME=269
                FIRST_EVENT_RECEIVED=16
                LAST_EVENT_RECEIVED=72
        MultiStoreCounters
                Output records in _0_output=2458721
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        Shuffle Errors_scope_27_INPUT_scope_26
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        TaskCounter_scope_26_INPUT_scope_0
                INPUT_RECORDS_PROCESSED=148170422
                INPUT_SPLIT_LENGTH_BYTES=3069919009
        TaskCounter_scope_26_OUTPUT_scope_27
                ADDITIONAL_SPILLS_BYTES_READ=0
                ADDITIONAL_SPILLS_BYTES_WRITTEN=0
                ADDITIONAL_SPILL_COUNT=0
                OUTPUT_BYTES=4023990818
                OUTPUT_BYTES_PHYSICAL=37619882
                OUTPUT_BYTES_WITH_OVERHEAD=75404318
                OUTPUT_RECORDS=148170422
                SHUFFLE_CHUNK_COUNT=23
                SPILLED_RECORDS=2458743
        TaskCounter_scope_27_INPUT_scope_26
                ADDITIONAL_SPILLS_BYTES_READ=37619882
                ADDITIONAL_SPILLS_BYTES_WRITTEN=0
                COMBINE_INPUT_RECORDS=0
                FIRST_EVENT_RECEIVED=16
                LAST_EVENT_RECEIVED=72
```

```
                LAST_EVENT_RECEIVED=72
                MERGED_MAP_OUTPUTS=391
                MERGE_PHASE_TIME=269
                NUM_DISK_TO_DISK_MERGES=0
                NUM_FAILED_SHUFFLE_INPUTS=0
                NUM_MEM_TO_DISK_MERGES=0
                NUM_SHUFFLED_INPUTS=391
                NUM_SKIPPED_INPUTS=0
                REDUCE_INPUT_GROUPS=2458721
                REDUCE_INPUT_RECORDS=2458743
                SHUFFLE_BYTES=37619882
                SHUFFLE_BYTES_DECOMPRESSED=75404318
                SHUFFLE_BYTES_DISK_DIRECT=37619882
                SHUFFLE_BYTES_TO_DISK=0
                SHUFFLE_BYTES_TO_MEM=0
                SHUFFLE_PHASE_TIME=106
                SPILLED_RECORDS=2458743
        TaskCounter_scope_27_OUTPUT_scope_25
                OUTPUT_RECORDS=2458721
        org.apache.hadoop.mapreduce.TaskCounter
                COMBINE_INPUT_RECORDS=2458743
                COMBINE_OUTPUT_RECORDS=148170422
        org.apache.hadoop.mapreduce.TaskCounter_scope_26_OUTPUT_scope_27
                COMBINE_INPUT_RECORDS=2458743
                COMBINE_OUTPUT_RECORDS=148170422
        org.apache.hadoop.mapreduce.TaskCounter_scope_27_INPUT_scope_26
                COMBINE_INPUT_RECORDS=0
                COMBINE_OUTPUT_RECORDS=0
2022-02-28 20:34:10,403 [main] INFO  org.apache.pig.tools.pigstats.tez.TezPigScriptStats - Script Statistics:

       HadoopVersion: 2.7.3.2.6.5.0-292
          PigVersion: 0.16.0.2.6.5.0-292
          TezVersion: 0.7.0.2.6.5.0-292
              UserId: s1155160788
            FileName: q1c.pig
           StartedAt: 2022-02-28 20:26:04
          FinishedAt: 2022-02-28 20:34:10
            Features: GROUP_BY

Success!


DAG 0:
                                   Name: PigLatin:q1c.pig-0_scope-0
                          ApplicationId: job_1644821811052_0578
                      TotalLaunchedTasks: 29
                           FileBytesRead: 50617332
                        FileBytesWritten: 37629450
                           HdfsBytesRead: 3072802593
                        HdfsBytesWritten: 68316032
```

```
DAG 0:
                                   Name: PigLatin:q1c.pig-0_scope-0
                          ApplicationId: job_1644821811052_0578
                      TotalLaunchedTasks: 29
                           FileBytesRead: 50617332
                        FileBytesWritten: 37629450
                           HdfsBytesRead: 3072802593
                        HdfsBytesWritten: 68316032
      SpillableMemoryManager spill count: 0
                 Bags proactively spilled: 0
            Records proactively spilled: 0

DAG Plan:
Tez vertex scope-26     ->      Tez vertex scope-27,
Tez vertex scope-27

Vertex Stats:
VertexId Parallelism TotalTasks   InputRecords   ReduceInputRecords  OutputRecords  FileBytesRead FileBytesWritten  HdfsBytesRead HdfsBytesWritten Alias        Feature Outputs
scope-26          23         23      148170422                    0      148170422           9568         37629450     3072802593                0 Bigram_Averaged,Bigram_Grouped,D
scope-27          17          6              0              2458743        2458721       50607764                0              0         68316032 Bigram_Averaged       GROUP_BY        /user
/s1155160788/output,

Input(s):
Successfully read 148170422 records (3072802593 bytes) from: "hdfs://dicvmd2.ie.cuhk.edu.hk:8020/user/s1155160788/united_books"

Output(s):
Successfully stored 2458721 records (68316032 bytes) in: "/user/s1155160788/output"

2022-02-28 20:34:10,415 [main] INFO  org.apache.pig.Main - Pig script completed in 8 minutes, 7 seconds and 416 milliseconds (487416 ms)
2022-02-28 20:34:10,415 [main] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezLauncher - Shutting down thread pool
2022-02-28 20:34:10,424 [pool-1-thread-1] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezSessionManager - Shutting down Tez session org.apache.tez.client.TezClient@486e7200
2022-02-28 20:34:10 Shutting down Tez session , sessionName=PigLatin:q1c.pig, applicationId=application_1644821811052_0578
2022-02-28 20:34:10,425 [pool-1-thread-1] INFO  org.apache.tez.client.TezClient - Shutting down Tez Session, sessionName=PigLatin:q1c.pig, applicationId=application_1644821811052_0578
[s1155160788@dicvmd10 Assignment2]$
```

As I am starting to get conscious of how verbose the Pig output is, I will omit most of the output in the future unless required by the assignment or if it is notable.

```
buoyline     2.9523809523809526
buprofen     3.5714285714285716
bur'glar     1.7878787878787878
bur_VERB     135.87818696883852
bura_ADJ     5.948453608247423
bura_DET     4.159574468085107
bura_NUM     2.156862745098039
burabura     4.357142857142857
buraco_X     7.859154929577465
burada_X     21.033333333333335
buradaki     4.025
buraka_X     5.21875
buraku_.     2.375
burdan_X     2.2941176470588234
burden.1     11.254545454545454
burden.2     7.074324324324325
```

A plausible result is produced. I verify this output by comparing the calculated average from this output to the calculated one the intermediate output of Attempt #1.

```
buovancy_NOUN    2.506172839506173
buoyance_NOUN    24.161290322580644
buoyancv_NOUN    1.7777777777777777
buoyancy_VERB    28.611510791366907
buoyanoy_NOUN    1.8305084745762712
buoyline_NOUN    2.9714285714285715
```

*Untitled - Notepad

File Edit Format View Help

```
Buoyline: Appears 42 times, sum count of 124. Average: 2.952
Buoyline_NOUN: Appears 35 times, has a sum count of 104. Average: 2.97
```

As we can see, the average from the produced output matches the manual calculation, and is thus verified.

With this, we can consider Section C complete.

Part D: Generate Top 20 Bigrams with highest average number of occurrences

Final Script / Code for Part D on next page.

q1d.pig

C: > Users > Eddie > Documents > School > CUHK > 2021 - 2022 Term 2 > IEMS 5730 - Big Data Systems and Informatio

```
D = LOAD 'PartC' USING PigStorage('\t') AS (bigram:chararray, average:float);
D_Ordered = ORDER D BY average DESC;
Top20 = LIMIT D_Ordered 20;
STORE Top20 INTO '/user/s1155160788/output' USING PigStorage('\t');
```

I refer to [7] for the ORDER BY and LIMIT notation.

```
            Features: ORDER_BY,LIMIT

Success!


DAG 0:
                                  Name: PigLatin:q1d.pig-0_scope-0
                         ApplicationId: job_1644821811052_0582
                    TotalLaunchedTasks: 9
                         FileBytesRead: 28904405
                      FileBytesWritten: 28746668
                         HdfsBytesRead: 136632064
                      HdfsBytesWritten: 318
      SpillableMemoryManager spill count: 0
                Bags proactively spilled: 0
            Records proactively spilled: 0

DAG Plan:
Tez vertex scope-11     ->      Tez vertex scope-26,
Tez vertex scope-26     ->      Tez vertex scope-36,
Tez vertex scope-36     ->      Tez vertex scope-37,
Tez vertex scope-37     ->      Tez vertex scope-48,
Tez vertex scope-48

Vertex Stats:
VertexId Parallelism TotalTasks   InputRecords   ReduceInputRecords  OutputRecords  FileBytesRead FileBytesWritten  HdfsBytesRead HdfsBytesWritten Alias        Feature Outputs
scope-11          1          1        2458721                    0            100             32            1280       68316032                0 D    SAMPLER
scope-26          1          1              0                  100              1           1280             217              0                0
scope-36          1          1        2458722                    0        2458721            104        28743369       68316032                0 D,D_Ordered
scope-37         -1          4              0                  110             80       28901187            1802              0                0      ORDER_BY
scope-48          1          1              0                   20             20           1802               0              0              318      LIMIT   /user/s1155160788/output,

Input(s):
Successfully read 2458721 records (68316032 bytes) from: "hdfs://dicvmd2.ie.cuhk.edu.hk:8020/user/s1155160788/PartC"
Successfully read 2458721 records (68316032 bytes) from: "hdfs://dicvmd2.ie.cuhk.edu.hk:8020/user/s1155160788/PartC"

Output(s):
Successfully stored 20 records (318 bytes) in: "/user/s1155160788/output"

2022-02-28 21:12:16,519 [main] INFO  org.apache.pig.Main - Pig script completed in 24 seconds and 463 milliseconds (24463 ms)
2022-02-28 21:12:16,519 [main] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezLauncher - Shutting down thread pool
2022-02-28 21:12:16,524 [pool-1-thread-1] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezSessionManager - Shutting down Tez session org.apache.tez.client.TezClient@405fc6eb
2022-02-28 21:12:16 Shutting down Tez session , sessionName=PigLatin:q1d.pig, applicationId=application_1644821811052_0582
2022-02-28 21:12:16,525 [pool-1-thread-1] INFO  org.apache.tez.client.TezClient - Shutting down Tez Session, sessionName=PigLatin:q1d.pig, applicationId=application_1644821811052_0582
[s1155160788@dicvmd10 Assignment2]$
```

PartD

C: > Users > Eddie > Documents > Schoo

```
and 2.5932078E7
and_CONJ    2.5906234E7
a   1.6665891E7
a_DET   1.6645121E7
as  6179734.0
be  5629591.5
be_VERB 5621156.0
as_ADP  5360444.0
by  5294067.0
by_ADP  5272952.0
are 4298564.5
are_VERB    4298561.5
at  3676050.2
at_ADP  3670625.8
an  2979272.8
an_DET  2977978.0
but 2471102.5
but_CONJ    2468978.0
all 2189962.8
all_DET 2161257.2
```

The output from Part D looks reasonable. The results are all extremely common words like and, a, be, and as.

With this, I consider Part D complete.

Question 1 Final Submission: Script that does Part B-D

I am not sure if I need to make a script that does all of Part B – D in one shot, operating with the raw dataset files until the task is completed. To be on the safe side, I am making one. Obviously, Part A is not included because that was installing Pig, however please note I did do the Part A bonus at the beginning of this assignment.

q1.pig

C: > Users > Eddie > Documents > School > CUHK > 2021 - 2022 Term 2 > IEMS 5730 - Big Data Systems and Information Processing > Week 7 > Assignment 2 > q1.pig

```
1  A = LOAD 'googlebooks-eng-all-1gram-20120701-a' USING PigStorage('\t') AS (bigram:chararray, year:int, match_count:int, volume_count:int);
2  B = LOAD 'googlebooks-eng-all-1gram-20120701-b' USING PigStorage('\t') AS (bigram:chararray, year:int, match_count:int, volume_count:int);
3  C = UNION A, B;
4  Bigram_Grouped = GROUP C BY bigram;
5  Bigram_Averaged = FOREACH Bigram_Grouped GENERATE group as bigram, AVG(C.$2) as average;
6  Bigram_Ordered = ORDER Bigram_Averaged BY average DESC;
7  Top20 = LIMIT Bigram_Ordered 20;
8  STORE Top20 INTO '/user/s1155160788/output' USING PigStorage('\t');
```

I take out the store commands until the end and modify the variable names to make this work, otherwise copying and pasting from my Part B-D scripts. One note is in line 5 I change AVG(C.match_count) to AVG(C.$2) because I do not believe that in this combined script, C, being the UNION of A and B, has a defined schema, unlike the individual part script where I load it in with a defined schema.

```
Success!


DAG 0:
                                  Name: PigLatin:q1.pig-0_scope-0
                        ApplicationId: job_1644821811052_0586
                   TotalLaunchedTasks: 64
                        FileBytesRead: 131730584
                     FileBytesWritten: 112699322
                        HdfsBytesRead: 3072802593
                     HdfsBytesWritten: 466
     SpillableMemoryManager spill count: 0
               Bags proactively spilled: 0
            Records proactively spilled: 0

DAG Plan:
Tez vertex scope-43     ->      Tez vertex group scope-104,
Tez vertex scope-44     ->      Tez vertex group scope-104,
Tez vertex group scope-104      ->      Tez vertex scope-49,
Tez vertex scope-49     ->      Tez vertex scope-58,Tez vertex scope-68,
Tez vertex scope-58     ->      Tez vertex scope-68,
Tez vertex scope-68     ->      Tez vertex scope-70,
Tez vertex scope-70     ->      Tez vertex scope-81,
Tez vertex scope-81

Vertex Stats:
VertexId Parallelism TotalTasks   InputRecords   ReduceInputRecords  OutputRecords  FileBytesRead FileBytesWritten  HdfsBytesRead HdfsBytesWritten Alias        Feature Outputs
scope-43          14         14       86618505                    0       86618505           5824         22114536     1803230011                0 A,Bigram_Averaged,Bigram_Grouped
scope-44          10         10       61551917                    0       61551917           4160         15309835     1269572582                0 B,Bigram_Averaged,Bigram_Grouped
scope-49          17         17              0              2458743        2460421       52608645         40674381              0                0 Bigram_Averaged,Bigram_Ordered       GROUP
_BY,SAMPLER
scope-58           1          1              0                 1700              1          24519            226              0                0
scope-68          17         17        2458722                    0        2458721       40651856         34598400              0                0 Bigram_Ordered
scope-70          -1          4              0                  154             80       38433636             1944              0                0      ORDER_BY
scope-81           1          1              0                   20             20           1944                0              0              466      LIMIT   /user/s1155160788/output,

Input(s):
Successfully read 86618505 records (1803230011 bytes) from: "hdfs://dicvmd2.ie.cuhk.edu.hk:8020/user/s1155160788/googlebooks-eng-all-1gram-20120701-a"
Successfully read 61551917 records (1269572582 bytes) from: "hdfs://dicvmd2.ie.cuhk.edu.hk:8020/user/s1155160788/googlebooks-eng-all-1gram-20120701-b"

Output(s):
Successfully stored 20 records (466 bytes) in: "/user/s1155160788/output"

2022-02-28 21:35:41,781 [main] INFO  org.apache.pig.Main - Pig script completed in 8 minutes, 50 seconds and 595 milliseconds (530595 ms)
2022-02-28 21:35:41,782 [main] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezLauncher - Shutting down thread pool
2022-02-28 21:35:41,791 [pool-1-thread-1] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezSessionManager - Shutting down Tez session org.apache.tez.client.TezClient@36b84746
2022-02-28 21:35:41 Shutting down Tez session , sessionName=PigLatin:q1.pig, applicationId=application_1644821811052_0586
2022-02-28 21:35:41,791 [pool-1-thread-1] INFO  org.apache.tez.client.TezClient - Shutting down Tez Session, sessionName=PigLatin:q1.pig, applicationId=application_1644821811052_0586
[s1155160788@dicvmd10 Assignment2]$
```

The Pig TEZ script executes successfully in 8 minutes and 50.595 seconds, and the output is identical.

Note: The output below uses scientific notation. So 2.593207744E7 is 2.593207744 * 10^7 = 25,932,077.44

q1

C: > Users > Eddie > Documents > School > CUHK >

```
and 2.593207744E7
and_CONJ    2.5906234451764707E7
a   1.6665890811764706E7
a_DET   1.6645121127058823E7
as  6179734.075294117
be  5629591.52
be_VERB 5621156.232941177
as_ADP  5360443.872941176
by  5294067.04
by_ADP  5272951.997647059
are 4298564.341176471
are_VERB    4298561.303529412
at  3676050.1529411767
at_ADP  3670625.785882353
an  2979272.7411764706
an_DET  2977977.8870588234
but 2471102.4964705883
but_CONJ    2468978.0564705883
all 2189962.722352941
all_DET 2161257.294117647
```

Question 2 Section A: Hive Installation on Local Hadoop Cluster

I refer to [13] for this process. I download from version 2.3.8 of Hive from https://archive.apache.org/dist/hive/hive-2.3.8/ . Even though version 2.3.9 is available and noted as stable, I go for 2.3.8 for consistency with the Assignment 2 PDF.

I download it to the server using wget.

```
EddieCFox@master:~$ wget https://archive.apache.org/dist/hive/hive-2.3.8/apache-hive-2.3.8-bin.tar.gz
--2022-02-28 15:02:14--  https://archive.apache.org/dist/hive/hive-2.3.8/apache-hive-2.3.8-bin.tar.gz
Resolving archive.apache.org (archive.apache.org)... 138.201.131.134, 2a01:4f8:172:2ec5::2
Connecting to archive.apache.org (archive.apache.org)|138.201.131.134|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 286137131 (273M) [application/x-gzip]
Saving to: 'apache-hive-2.3.8-bin.tar.gz'

apache-hive-2.3.8-bin.tar.gz                    100%[===================================================================================================>] 272.88M  10.8MB/s    in 27s

2022-02-28 15:02:42 (10.1 MB/s) - 'apache-hive-2.3.8-bin.tar.gz' saved [286137131/286137131]

EddieCFox@master:~$ ls
apache-hive-2.3.8-bin.tar.gz  hadoop-2.9.2.tar.gz  pig-0.17.0  pig-0.17.0.tar.gz
EddieCFox@master:~$ ls -a
.  ..  .bash_history  .bash_logout  .bashrc  .cache  .gnupg  .pig_history  .profile  .ssh  .viminfo  apache-hive-2.3.8-bin.tar.gz  hadoop-2.9.2.tar.gz  pig-0.17.0  pig-0.17.0.tar.gz
EddieCFox@master:~$ pwd
/home/EddieCFox
EddieCFox@master:~$
```

```
EddieCFox@master:~$ tar -xzvf apache-hive-2.3.8-bin.tar.gz
apache-hive-2.3.8-bin/LICENSE
apache-hive-2.3.8-bin/RELEASE_NOTES.txt
apache-hive-2.3.8-bin/NOTICE
apache-hive-2.3.8-bin/binary-package-licenses/com.thoughtworks.paranamer-LICENSE
apache-hive-2.3.8-bin/binary-package-licenses/org.codehaus.janino-LICENSE
apache-hive-2.3.8-bin/binary-package-licenses/org.jamon.jamon-runtime-LICENSE
apache-hive-2.3.8-bin/binary-package-licenses/org.mozilla.rhino-LICENSE
apache-hive-2.3.8-bin/binary-package-licenses/org.jruby-LICENSE
apache-hive-2.3.8-bin/binary-package-licenses/jline-LICENSE
apache-hive-2.3.8-bin/binary-package-licenses/org.antlr-LICENSE
apache-hive-2.3.8-bin/binary-package-licenses/org.abego.treelayout.core-LICENSE
apache-hive-2.3.8-bin/binary-package-licenses/sqlline-LICENSE
apache-hive-2.3.8-bin/binary-package-licenses/org.slf4j-LICENSE
apache-hive-2.3.8-bin/binary-package-licenses/org.antlr.antlr4-LICENSE
apache-hive-2.3.8-bin/binary-package-licenses/com.google.protobuf-LICENSE
apache-hive-2.3.8-bin/binary-package-licenses/com.sun.jersey-LICENSE
apache-hive-2.3.8-bin/binary-package-licenses/javolution-LICENSE
apache-hive-2.3.8-bin/binary-package-licenses/NOTICE
apache-hive-2.3.8-bin/binary-package-licenses/asm-LICENSE
apache-hive-2.3.8-bin/binary-package-licenses/com.ibm.icu.icu4j-LICENSE
apache-hive-2.3.8-bin/binary-package-licenses/org.antlr.stringtemplate-LICENSE
apache-hive-2.3.8-bin/binary-package-licenses/javax.transaction.transaction-api-LICENSE
apache-hive-2.3.8-bin/examples/files/hive_626_bar.txt
apache-hive-2.3.8-bin/examples/files/docurl.txt
apache-hive-2.3.8-bin/examples/files/tjoin2.txt
apache-hive-2.3.8-bin/examples/files/person+age.txt
apache-hive-2.3.8-bin/examples/files/avro_charvarchar.txt
apache-hive-2.3.8-bin/examples/files/parquet_array_null_element.txt
apache-hive-2.3.8-bin/examples/files/smbbucket_3.rc
apache-hive-2.3.8-bin/examples/files/tiny_a.txt
apache-hive-2.3.8-bin/examples/files/parquet_type_promotion.txt
apache-hive-2.3.8-bin/examples/files/over10k.gz
apache-hive-2.3.8-bin/examples/files/map_table.txt
apache-hive-2.3.8-bin/examples/files/table1.avsc
apache-hive-2.3.8-bin/examples/files/in9.txt
apache-hive-2.3.8-bin/examples/files/tjoin1.txt
apache-hive-2.3.8-bin/examples/files/sortdp.txt
apache-hive-2.3.8-bin/examples/files/tsformat.json
apache-hive-2.3.8-bin/examples/files/sample-queryplan-in-history.txt
apache-hive-2.3.8-bin/examples/files/tiny_b.txt
apache-hive-2.3.8-bin/examples/files/non_ascii_tbl.txt
apache-hive-2.3.8-bin/examples/files/in8.txt
apache-hive-2.3.8-bin/examples/files/dec_old.avro
apache-hive-2.3.8-bin/examples/files/fact-data.txt
apache-hive-2.3.8-bin/examples/files/customer_address.txt
apache-hive-2.3.8-bin/examples/files/create_nested_type.txt
apache-hive-2.3.8-bin/examples/files/kv8.txt
apache-hive-2.3.8-bin/examples/files/part.rc
apache-hive-2.3.8-bin/examples/files/flights_tiny.txt.1
```

Full output of this command omitted.

```
EddieCFox@master:~$ ls
apache-hive-2.3.8-bin  apache-hive-2.3.8-bin.tar.gz  hadoop-2.9.2.tar.gz  pig-0.17.0  pig-0.17.0.tar.gz
EddieCFox@master:~$ cd apache-hive-2.3.8-bin/
EddieCFox@master:~/apache-hive-2.3.8-bin$ export HIVE_HOME={{pwd}}
EddieCFox@master:~/apache-hive-2.3.8-bin$ export PATH=$HIVE_HOME/bin:$PATH
EddieCFox@master:~/apache-hive-2.3.8-bin$
```

Afterwards I follow the remaining installation instructions in [13] under "Installing Hive from a Stable Release."

```
EddieCFox@master:/$ cd /usr/local/hadoop
EddieCFox@master:/usr/local/hadoop$ sbin/start-dfs.sh
Starting namenodes on [master]
master: starting namenode, logging to /usr/local/hadoop/logs/hadoop-EddieCFox-namenode-master.out
slave2: starting datanode, logging to /usr/local/hadoop/logs/hadoop-EddieCFox-datanode-slave2.out
slave1: starting datanode, logging to /usr/local/hadoop/logs/hadoop-EddieCFox-datanode-slave1.out
slave3: starting datanode, logging to /usr/local/hadoop/logs/hadoop-EddieCFox-datanode-slave3.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /usr/local/hadoop/logs/hadoop-EddieCFox-secondarynamenode-master.out
EddieCFox@master:/usr/local/hadoop$ sbin/start-yarn.sh
starting yarn daemons
starting resourcemanager, logging to /usr/local/hadoop/logs/yarn-EddieCFox-resourcemanager-master.out
slave2: starting nodemanager, logging to /usr/local/hadoop/logs/yarn-EddieCFox-nodemanager-slave2.out
slave1: starting nodemanager, logging to /usr/local/hadoop/logs/yarn-EddieCFox-nodemanager-slave1.out
slave3: starting nodemanager, logging to /usr/local/hadoop/logs/yarn-EddieCFox-nodemanager-slave3.out
EddieCFox@master:/usr/local/hadoop$ jps
32593 NameNode
833 Jps
564 ResourceManager
392 SecondaryNameNode
EddieCFox@master:/usr/local/hadoop$
```

```
EddieCFox@slave1:~$ jps
28824 NodeManager
28601 DataNode
28922 Jps
EddieCFox@slave1:~$
```

I start up HDFS and Yarn again and confirm with jps on master and slave nodes that everything is functioning normally.

```
EddieCFox@master:/usr/local/hadoop$ bin/hdfs dfs -mkdir /tmp
```

```
EddieCFox@master:/usr/local/hadoop$ bin/hdfs dfs -ls /
Found 2 items
drwxr-xr-x   - EddieCFox supergroup          0 2022-02-28 15:13 /tmp
drwxr-xr-x   - EddieCFox supergroup          0 2022-02-25 15:00 /user
EddieCFox@master:/usr/local/hadoop$ bin/hdfs dfs -mkdir /user/hive
EddieCFox@master:/usr/local/hadoop$ bin/hdfs dfs -mkdir /user/hive/warehouse
EddieCFox@master:/usr/local/hadoop$ bin/hdfs dfs -chmod g+w /tmp
EddieCFox@master:/usr/local/hadoop$ bin/hdfs dfs -chmod g+w /user/hive/warehouse
```

```
EddieCFox@master:/usr/local/hadoop$ bin/hdfs dfs -ls /user
Found 2 items
drwxr-xr-x   - EddieCFox supergroup          0 2022-02-25 15:00 /user/EddieCFox
drwxr-xr-x   - EddieCFox supergroup          0 2022-02-28 15:15 /user/hive
EddieCFox@master:/usr/local/hadoop$ bin/hdfs dfs -ls /user/hive
Found 1 items
drwxrwxr-x   - EddieCFox supergroup          0 2022-02-28 15:15 /user/hive/warehouse
EddieCFox@master:/usr/local/hadoop$
```

I follow the instructions in [13] under “Running Hive”, and also mentioned in the Assignment 2 PDF, to set up the HDFS accordingly.

```
EddieCFox@master:/usr/local/hadoop$ $HIVE_HOME/bin/hive
-bash: {{pwd}}/bin/hive: No such file or directory
EddieCFox@master:/usr/local/hadoop$ echo $HIVE_HOME
{{pwd}}
EddieCFox@master:/usr/local/hadoop$
```

I realize that when the instructions said to set $HIVE_HOME as {{pwd}} they meant to actually take the contents of pwd when in the installation directory and put that instead. I ran “export HIVE_HOME=/home/EddieCFox/apache-hive-2.3.8-bin” and “export PATH=$HIVE_HOME/bin:$PATH” again.

```
EddieCFox@master:~/apache-hive-2.3.8-bin$ hive
Cannot find hadoop installation: $HADOOP_HOME or $HADOOP_PREFIX must be set or hadoop must be in the path
EddieCFox@master:~/apache-hive-2.3.8-bin$ echo $HADOOP_HOME

EddieCFox@master:~/apache-hive-2.3.8-bin$
```

```
EddieCFox@master:~/apache-hive-2.3.8-bin$ export HADOOP_HOME=/usr/local/hadoop
EddieCFox@master:~/apache-hive-2.3.8-bin$ hive
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/EddieCFox/apache-hive-2.3.8-bin/lib/log4j-slf4j-impl-2.6.2.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hadoop/share/hadoop/common/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]

Logging initialized using configuration in jar:file:/home/EddieCFox/apache-hive-2.3.8-bin/lib/hive-common-2.3.8.jar!/hive-log4j2.properties Async: true
Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
hive>
```

I configure HADOOP_HOME according to [14] and then run hive, launching it successfully.

```
hive> ! hive --version;
Hive 2.3.8
Git git://chaos-mbp.lan/Users/chao/git/hive -r f1e87137034e4ecbe39a859d4ef44319800016d7
Compiled by chao on Thu Jan 7 11:36:26 PST 2021
From source with checksum 2992381e2a287352c65262bf40d3f932
hive> quit;
EddieCFox@master:~/apache-hive-2.3.8-bin$
```

Using [15], I check the Hive version. With this, I consider it to be installed.

Question 2 Section B: (Question 1) on Hive

```
[s1155160788@dicvmd10 ~]$ hive
log4j:WARN No such property [maxFileSize] in org.apache.log4j.DailyRollingFileAppender.

Logging initialized using configuration in file:/etc/hive/2.6.5.0-292/0/hive-log4j.properties
hive> ! hive --version;
Hive 1.2.1000.2.6.5.0-292
Subversion git://ctr-e138-1518143905142-227581-01-000005.hwx.site/grid/0/jenkins/workspace/HDP-parallel-centos7/SOURCES/hive -r 6cbb7b58d08f594c2545e8a177e4f6b7ed4a9f77
Compiled by jenkins on Fri May 11 07:53:58 UTC 2018
From source with checksum cb7deec5db32ca9a06dc93b73dc02d1f
```

Back on the IE DIC cluster, we see the version is either 2.6.5.0-292 or version 1.2.1000.2.6.5.0-292.

```
[s1155160788@dicvmd10 ~]$ hdfs dfs -mkdir /tmp
mkdir: `/tmp': File exists
[s1155160788@dicvmd10 ~]$ hdfs dfs -mkdir /user
mkdir: `/user': File exists
[s1155160788@dicvmd10 ~]$ hdfs dfs -mkdir /user/hive
mkdir: `/user/hive': File exists
[s1155160788@dicvmd10 ~]$ hdfs dfs -mkdir /user/hive/warehouse
mkdir: Permission denied: user=s1155160788, access=WRITE, inode="/user/hive/warehouse":hive:h
dfs:drwxr-xr-x
```

I try to run the same hdfs operations as before but all the folders already exist besides the hive warehouse, which I do not have permission to create. This makes sense since these directory commands are operating off of the root instead of my user partition, so I try to run them relative to my home directory in HDFS instead.

```
[s1155160788@dicvmd10 ~]$ hdfs dfs -mkdir /user/s1155160788/tmp
[s1155160788@dicvmd10 ~]$ hdfs dfs -mkdir /user/s1155160788/user
[s1155160788@dicvmd10 ~]$ hdfs dfs -mkdir /user/s1155160788/user/hive
```

```
[s1155160788@dicvmd10 ~]$ hdfs dfs -mkdir /user/s1155160788/user/hive/warehouse
[s1155160788@dicvmd10 ~]$ hdfs dfs -ls
Found 7 items
drwxr-xr-x   - s1155160788 student          0 2022-02-28 23:42 .hiveJars
-rw-r--r--   3 s1155160788 student 1801526075 2022-02-26 09:25 googlebooks-eng-all-1gram-20120701-a
-rw-r--r--   3 s1155160788 student 1268392934 2022-02-26 09:25 googlebooks-eng-all-1gram-20120701-b
-rw-r--r--   3 s1155160788 student   17103777 2022-02-09 16:28 movielens_large_updated.csv
-rw-r--r--   3 s1155160788 student     893989 2022-02-09 16:28 movielens_small.csv
drwxr-xr-x   - s1155160788 student          0 2022-03-01 09:43 tmp
drwxr-xr-x   - s1155160788 student          0 2022-03-01 09:44 user
[s1155160788@dicvmd10 ~]$
```

```
[s1155160788@dicvmd10 ~]$ hdfs dfs -chmod g+w tmp
[s1155160788@dicvmd10 ~]$ hdfs dfs -chmod g+w /user/s1155160788/user/hive/warehouse
[s1155160788@dicvmd10 ~]$
```

I need to specify the absolute path for the later command because even if I can access user by name, to reach the hive and warehouse folders, user/hive and user/hive/warehouse don't work.

On the IE DIC, $HIVE_HOME is empty, so I need to fill it. To find this, I refer to [16] which advises going into hive shell and running "!env;"

```
HIVE_HOME=/usr/hdp/2.6.5.0-292/hive
```

```
[s1155160788@dicvmd10 hive]$ cd /usr/hdp/2.6.5.0-292/hive
[s1155160788@dicvmd10 hive]$ ls
bin  conf  doc  hive.tar.gz  jdbc  lib  log  man  metastore  scripts
[s1155160788@dicvmd10 hive]$ cd bin
[s1155160788@dicvmd10 bin]$ ls
beeline         ext   hive            hive.distro  hiveserver2.distro  metatool
beeline.distro  hcat  hive-config.sh  hiveserver2  init-hive-dfs.sh    schematool
[s1155160788@dicvmd10 bin]$
```

I navigate to the HIVE_HOME mentioned in the env, go into bin, and see schematool.

I then modify my .bashrc to add HIVE_HOME and the path, then run source on .bashrc to reload it.

```
export HIVE_HOME=/usr/hdp/2.6.5.0-292/hive
export PATH=$HIVE_HOME/bin:$PATH
~
```

```
[s1155160788@dicvmd10 ~]$ schematool
HiveSchemaTool:Parsing failed.  Reason: Missing required option: [-help print this message, -
initSchemaTo Schema initialization to a version, -upgradeSchemaFrom Schema upgrade from a ver
sion, -initSchema Schema initialization, -upgradeSchema Schema upgrade, -info Show config and
 schema details]
usage: schemaTool
 -dbOpts <databaseOpts>            Backend DB specific options
 -dbType <databaseType>            Metastore database type
 -dryRun                           list SQL scripts (no execute)
 -help                             print this message
 -info                             Show config and schema details
 -initSchema                       Schema initialization
 -initSchemaTo <initTo>            Schema initialization to a version
 -passWord <password>              Override config file password
 -upgradeSchema                    Schema upgrade
 -upgradeSchemaFrom <upgradeFrom>  Schema upgrade from a version
 -userName <user>                  Override config file user name
 -verbose                          only print SQL statements
[s1155160788@dicvmd10 ~]$
```

Running schematool now provides output.

```
<property>
  <name>hive.metastore.warehouse.dir</name>
  <value>/apps/hive/warehouse</value>
</property>
```

Searching in the hive-site.xml in the conf folder of HIVE_HOME shows the metastore is in /apps/hive/warehouse, yet when I search for /apps, it says there is no such folder. It is probably hidden, so I decide to proceed.

[13] notes to run “schematool -dbType derby -initSchema”

```
[s1155160788@dicvmd10 conf]$ schematool -dbType derby -initSchema --verbose
Metastore connection URL:        jdbc:mysql://dicvmd3.ie.cuhk.edu.hk/hive?createDatabaseIfNotExist=true
Metastore Connection Driver :    com.mysql.jdbc.Driver
Metastore connection User:       hive
org.apache.hadoop.hive.metastore.HiveMetaException: Failed to load driver
Underlying cause: java.lang.ClassNotFoundException : com.mysql.jdbc.Driver
org.apache.hadoop.hive.metastore.HiveMetaException: Failed to load driver
        at org.apache.hive.beeline.HiveSchemaHelper.getConnectionToMetastore(HiveSchemaHelper.java:82)
        at org.apache.hive.beeline.HiveSchemaTool.getConnectionToMetastore(HiveSchemaTool.java:121)
        at org.apache.hive.beeline.HiveSchemaTool.testConnectionToMetastore(HiveSchemaTool.java:169)
        at org.apache.hive.beeline.HiveSchemaTool.doInit(HiveSchemaTool.java:272)
        at org.apache.hive.beeline.HiveSchemaTool.doInit(HiveSchemaTool.java:258)
        at org.apache.hive.beeline.HiveSchemaTool.main(HiveSchemaTool.java:528)
        at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
        at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:62)
        at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
        at java.lang.reflect.Method.invoke(Method.java:498)
        at org.apache.hadoop.util.RunJar.run(RunJar.java:233)
        at org.apache.hadoop.util.RunJar.main(RunJar.java:148)
Caused by: java.lang.ClassNotFoundException: com.mysql.jdbc.Driver
        at java.net.URLClassLoader.findClass(URLClassLoader.java:381)
        at java.lang.ClassLoader.loadClass(ClassLoader.java:424)
        at sun.misc.Launcher$AppClassLoader.loadClass(Launcher.java:331)
        at java.lang.ClassLoader.loadClass(ClassLoader.java:357)
        at java.lang.Class.forName0(Native Method)
        at java.lang.Class.forName(Class.java:264)
        at org.apache.hive.beeline.HiveSchemaHelper.getConnectionToMetastore(HiveSchemaHelper.java:73)
        ... 11 more
*** schemaTool failed ***
[s1155160788@dicvmd10 conf]$
```

Schematool fails. The most likely reason for this is because whatever needs to be initialized by schematool is already done by the IE DIC administrators, and there is likely permission issues preventing me from being able to do this, so I just proceed on.

When programming this, I refer to [17], [12] (the hive slides) and [18]. I am creating a script is Visual Studio Code after downloading the Hive SQL extension for interpretation.

Assignment 2 suggests testing on a small subset of the dataset until the commands / scripts work as desired.

I first open up the raw datasets A and B with PilotEdit and identify a good cutting off point. Using cat and head commands, I create a file called A_Test with the first 605 lines of the A dataset. I choose 605 specifically to ensure the final bigram is not split, which would lead to partial results and inaccuracy in calculations. For B, I make a file called B_Test with the first 641 lines.

```
[s1155160788@dicvmd10 ~]$ hdfs dfs -mkdir A_Test
[s1155160788@dicvmd10 ~]$ hdfs dfs -mkdir B_Test
```

```
[s1155160788@dicvmd10 Datasets_Test]$ hdfs dfs -copyFromLocal A_Test A_Test
[s1155160788@dicvmd10 Datasets_Test]$ hdfs dfs -copyFromLocal B_Test B_Test
```

I copy A_Test and B_Test files from local into respective A_Test and B_Test directories after making them in HDFS.

Attempt 1:

q1.sql q1.pig

C: > Users > Eddie > Documents > School > CUHK > 2021 - 2022 Term 2 > IEMS 5730 - Big Data Systems and Information

```sql
create external table A_Table (
            bigram STRING,
            year INT,
            match_count INT,
            volume_count INT)
row format serde 'org.apache.hadoop.hive.serde2.OpenCSVSerde' with serdeproperties (
    "separatorChar" = "\t",
    "quoteChar" = "\'")
stored as textfile
location '/user/s1155160788/A_Test/';

create external table B_Table (
            bigram STRING,
            year INT,
            match_count INT,
            volume_count INT)
row format serde 'org.apache.hadoop.hive.serde2.OpenCSVSerde' with serdeproperties (
    "separatorChar" = "\t",
    "quoteChar" = "\'")
stored as textfile
location '/user/s1155160788/B_Test/';

LOAD DATA INPATH '/user/s1155160788/A_Test/A_Test' OVERWRITE INTO TABLE A_Table;
LOAD DATA INPATH '/user/s1155160788/B_Test/B_Test' OVERWRITE INTO TABLE B_Table;
```

I first use [17] and [12] slide 80 to write the equivalent of the first 2 lines of my Pig script where I load each dataset (the test subset) into the respective tables after creating them, then use “hive -f q1.sql” to run the script [18].

```
[s1155160788@dicvmd10 Assignment2]$ hive -f q1.sql
log4j:WARN No such property [maxFileSize] in org.apache.log4j.DailyRollingFileAppender.

Logging initialized using configuration in file:/etc/hive/2.6.5.0-292/0/hive-log4j.properties
OK
Time taken: 1.087 seconds
OK
Time taken: 0.211 seconds
Loading data to table default.a_table
Failed with exception Unable to move source hdfs://dicvmd2.ie.cuhk.edu.hk:8020/user/s1155160788/A_Test/A_Test to destination hdfs://dicvmd2.ie.cuhk.edu.hk:8020/user/s1155160788/A_Test/A_Tes
t
FAILED: Execution Error, return code 1 from org.apache.hadoop.hive.ql.exec.MoveTask
```

Upon trying my script, I would the b_table would have the data loaded, but mysteriously not the a_table.

Correct / final attempt:

I had a feeling it might have to do with permissions, I recalled that the conf said the metastore was at /apps/hive/warehouse, and I found it through ls searching in the hdfs instead of the local directory.

```
[s1155160788@dicvmd10 Datasets_Test]$ hdfs dfs -ls /
Found 11 items
drwxrwxrwx   - yarn   hadoop          0 2022-02-28 21:49 /app-logs
drwxr-xr-x   - hdfs   hdfs            0 2021-10-23 18:07 /apps
drwxr-xr-x   - yarn   hadoop          0 2021-10-23 18:06 /ats
drwxr-xr-x   - hdfs   hdfs            0 2021-10-23 18:06 /hdp
drwxr-xr-x   - hdfs   hdfs            0 2021-10-28 18:41 /home
drwx------   - livy   hdfs            0 2021-10-23 18:06 /livy2-recovery
drwxr-xr-x   - mapred hdfs            0 2021-10-23 18:06 /mapred
drwxrwxrwx   - mapred hadoop          0 2021-10-23 18:06 /mr-history
drwxrwxrwx   - spark  hadoop          0 2022-03-01 13:51 /spark2-history
drwxrwxrwx   - hdfs   hdfs            0 2022-03-01 12:23 /tmp
drwxr-xr-x   - hdfs   hdfs            0 2022-02-20 01:24 /user
[s1155160788@dicvmd10 Datasets_Test]$ hdfs dfs -ls /apps
Found 2 items
drwxr-xr-x   - hdfs hdfs          0 2021-10-23 18:06 /apps/hbase
drwxr-xr-x   - hdfs hdfs          0 2021-10-23 18:07 /apps/hive
[s1155160788@dicvmd10 Datasets_Test]$ hdfs dfs -ls /apps/hive
Found 1 items
drwxrwxrwx   - hive hadoop          0 2022-02-26 20:41 /apps/hive/warehouse
[s1155160788@dicvmd10 Datasets_Test]$
```

Permissions appear to be configured correctly.

Continued on next page.

```
hive> DESCRIBE a_table;
OK
bigram                  string                  from deserializer
year                    string                  from deserializer
match_count             string                  from deserializer
volume_count            string                  from deserializer
Time taken: 0.329 seconds, Fetched: 4 row(s)
hive> DESCRIBE b_table;
OK
bigram                  string                  from deserializer
year                    string                  from deserializer
match_count             string                  from deserializer
volume_count            string                  from deserializer
```

I decide to move the files out of the A_Test and B_Test directories and into the home directory.

One issue I noticed is that describing the schema of the table, every item is of type string. [19] says this is a feature of OpenCSVSerde which our tutorial used. Another issue was that even though Table B was loading successfully, I noticed that about half of the rows were NULL. Looking at the test input, the lines with NULL were lines with single quote marks (like ‘). This was likely also an issue with the OpenCSVSerde as it has a quote character field that I had previously just copied with whatever the tutorial said.

Instead I found how to tab delimit it without serde in [20]. I also change from external tables to regular tables, in case it helps with the permissions.

```sql
CREATE TABLE A_Table (
            bigram STRING,
            year INT,
            match_count INT,
            volume_count INT)
ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t'
stored as textfile;

CREATE TABLE B_Table (
            bigram STRING,
            year INT,
            match_count INT,
            volume_count INT)
row format DELIMITED FIELDS TERMINATED BY '\t'
stored as textfile;

LOAD DATA INPATH '/user/s1155160788/A_Test' OVERWRITE INTO TABLE A_Table;
LOAD DATA INPATH '/user/s1155160788/B_Test' OVERWRITE INTO TABLE B_Table;
```

```
[s1155160788@dicvmd10 Datasets_Test]$ ls
A_Test  B_Test  q1.sql
[s1155160788@dicvmd10 Datasets_Test]$ hive -f q1.sql
log4j:WARN No such property [maxFileSize] in org.apache.log4j.DailyRollingFileAppender.

Logging initialized using configuration in file:/etc/hive/2.6.5.0-292/0/hive-log4j.properties
OK
Time taken: 1.172 seconds
OK
Time taken: 0.216 seconds
Loading data to table default.a_table
chgrp: changing ownership of 'hdfs://dicvmd2.ie.cuhk.edu.hk:8020/apps/hive/warehouse/a_table/A_Test': User null does not belong to hadoop
Table default.a_table stats: [numFiles=1, numRows=0, totalSize=12008, rawDataSize=0]
OK
Time taken: 0.754 seconds
Loading data to table default.b_table
chgrp: changing ownership of 'hdfs://dicvmd2.ie.cuhk.edu.hk:8020/apps/hive/warehouse/b_table/B_Test': User null does not belong to hadoop
Table default.b_table stats: [numFiles=1, numRows=0, totalSize=12746, rawDataSize=0]
OK
Time taken: 0.75 seconds
[s1155160788@dicvmd10 Datasets_Test]$
```

```
A.L.I.V.E.      2004    3       2
A.L.I.V.E.      2008    2       2
Time taken: 0.214 seconds, Fetched: 605 row(s)
hive> select * from b_table;
OK
B'enard 1974    1       1
B'enard 1982    1       1
B'enard 1993    1       1
B'enard 1997    1       1
B'enard 2001    2       1
B'enard 2003    3       2
```

```
hive> describe A_Table;
OK
bigram                  string
year                    int
match_count             int
volume_count            int
Time taken: 0.615 seconds, Fetched: 4 row(s)
hive> describe B_Table;
OK
bigram                  string
year                    int
match_count             int
volume_count            int
Time taken: 0.33 seconds, Fetched: 4 row(s)
hive>
```

Both tables show up upon selecting everything, including those with quote characters. The data types for the table schemas are also now correct.

q1.sql × q1.pig

C: > Users > Eddie > Documents > School > CUHK > 2021 - 2022 Term 2 > IEMS 5730 - Big Data Systems and Information Processing > Week 7 > Assignment 2 > q1.sql

```sql
CREATE TABLE A_Table (
            bigram STRING,
            year INT,
            match_count INT,
            volume_count INT)
ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t'
stored as textfile;

CREATE TABLE B_Table (
            bigram STRING,
            year INT,
            match_count INT,
            volume_count INT)
row format DELIMITED FIELDS TERMINATED BY '\t'
stored as textfile;

LOAD DATA INPATH '/user/s1155160788/A_Test' OVERWRITE INTO TABLE A_Table;
LOAD DATA INPATH '/user/s1155160788/B_Test' OVERWRITE INTO TABLE B_Table;

CREATE TABLE Averages_Test (
            bigram STRING,
            average INT)
row format DELIMITED FIELDS TERMINATED BY '\t'
stored as textfile;

INSERT OVERWRITE TABLE Averages_Test SELECT bigram, avg(match_count) as Average FROM(
    SELECT bigram, match_count FROM A_Table
    UNION ALL
    SELECT bigram, match_count FROM B_Table
) United_Test
GROUP BY bigram
ORDER BY Average DESC
LIMIT 20;
```

This is the final version of my test code. First, I create another table to hold the final output with just bigram and average. I decide that given the expected values, it is ok to just have the averages as INT and truncate the decimals. Unlike something going from 0 to 1 like similarity, here, there

is not much difference between an average of 346 and 345 appearances. [21] provides me with the syntax for inserting results into the table, and I refer to [22] to learn how to write the UNION statements. I am able to select just the two columns, bigram and match_count to unite, which should be faster than having to process all 4 columns. [23] informs me that UNION ALL is more applicable because UNION will remove duplicates rows. Finally, I refer to [24] to confirm my understanding of order and limit.

Screenshots on next page:

```
[s1155160788@dicvmd10 Datasets_Test]$ hive -f q1.sql
log4j:WARN No such property [maxFileSize] in org.apache.log4j.DailyRollingFileAppender.

Logging initialized using configuration in file:/etc/hive/2.6.5.0-292/0/hive-log4j.properties
OK
Time taken: 1.17 seconds
OK
Time taken: 0.214 seconds
Loading data to table default.a_table
chgrp: changing ownership of 'hdfs://dicvmd2.ie.cuhk.edu.hk:8020/apps/hive/warehouse/a_table/A_Test': User null does not belong to hadoop
Table default.a_table stats: [numFiles=1, numRows=0, totalSize=12008, rawDataSize=0]
OK
Time taken: 0.754 seconds
Loading data to table default.b_table
chgrp: changing ownership of 'hdfs://dicvmd2.ie.cuhk.edu.hk:8020/apps/hive/warehouse/b_table/B_Test': User null does not belong to hadoop
Table default.b_table stats: [numFiles=1, numRows=0, totalSize=12746, rawDataSize=0]
OK
Time taken: 0.751 seconds
OK
Time taken: 0.218 seconds
Query ID = s1155160788_20220301173720_dec7d8c2-b914-4d55-9e66-4314fa7ed359
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1644821811052_0640)

--------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      1          1        0        0       0       0
Map 5 ..........   SUCCEEDED      1          1        0        0       0       0
Reducer 3 ......   SUCCEEDED      1          1        0        0       0       0
Reducer 4 ......   SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 04/04  [==========================>>] 100%  ELAPSED TIME: 8.51 s
--------------------------------------------------------------------------------
Loading data to table default.averages_test
Table default.averages_test stats: [numFiles=1, numRows=20, totalSize=254, rawDataSize=234]
OK
Time taken: 10.558 seconds
[s1155160788@dicvmd10 Datasets_Test]$ hive
log4j:WARN No such property [maxFileSize] in org.apache.log4j.DailyRollingFileAppender.

Logging initialized using configuration in file:/etc/hive/2.6.5.0-292/0/hive-log4j.properties
hive> show tables
    > ;
OK
a
a_table
ab
averages_test
```

Screenshots continued on next page.

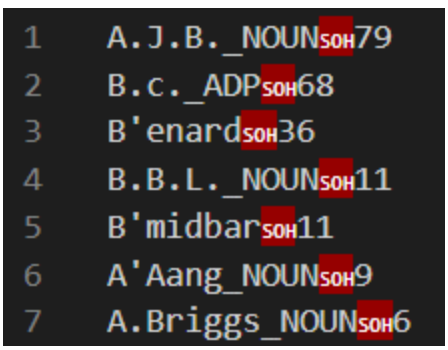
```
hive> show tables
    > ;
OK
a
a_table
ab
averages_test
b
b_table
data_big
employee
files
googlebooks
googledata
movies
res
salary
word_counts
wordavg
Time taken: 1.081 seconds, Fetched: 16 row(s)
hive> select * from averages_test
    > ;
OK
A.J.B._NOUN     79
B.c._ADP        68
B'enard 36
B.B.L._NOUN     11
B'midbar        11
A'Aang_NOUN     9
A.Briggs_NOUN   6
B.89_NOUN       6
A.D.A.A.        5
A.E.U._DET      5
A.K.K.  5
A.C.M.S._NOUN   5
A.5.3_DET       5
B.Allison_NOUN  4
A.L.I.V.E.      4
B.N.Z._NOUN     4
B.Pt._NOUN      4
A.IR._NOUN      3
A.165   3
B.5.b   3
Time taken: 0.216 seconds, Fetched: 20 row(s)
hive>
```

I get reasonable looking output and verify the results by comparing the output of the table (after a select * from) query and the previous unabridged results of averages from the Pig script. We can see that B'enard has an average of 36 appearances in both.

With this, I change all the relevant names to the appropriate ones for the actual dataset. I realize that I also didn't store the results of the table anywhere, which is required by the assignment.

```
INSERT OVERWRITE directory '/user/s1155160788/output' SELECT * from Averages_Test
```

I try this command above after seeing in [25] that I should use something similar to this to store data.



While the values appear correct, there is some weird delimiter instead of tab. [26] says that as of version 1.2, it is possible to specify a delimiter for non-local directory overwrites, which I do below.

Screenshots on next page.

```
[s1155160788@dicvmd10 Datasets_Test]$ hive -f q1.sql
log4j:WARN No such property [maxFileSize] in org.apache.log4j.DailyRollingFileAppender.

Logging initialized using configuration in file:/etc/hive/2.6.5.0-292/0/hive-log4j.properties
OK
Time taken: 1.18 seconds
OK
Time taken: 0.22 seconds
Loading data to table default.a_table
chgrp: changing ownership of 'hdfs://dicvmd2.ie.cuhk.edu.hk:8020/apps/hive/warehouse/a_table/A_Test': User null does not belong to hadoop
Table default.a_table stats: [numFiles=1, numRows=0, totalSize=12008, rawDataSize=0]
OK
Time taken: 0.754 seconds
Loading data to table default.b_table
chgrp: changing ownership of 'hdfs://dicvmd2.ie.cuhk.edu.hk:8020/apps/hive/warehouse/b_table/B_Test': User null does not belong to hadoop
Table default.b_table stats: [numFiles=1, numRows=0, totalSize=12746, rawDataSize=0]
OK
Time taken: 0.753 seconds
OK
Time taken: 0.216 seconds
Query ID = s1155160788_20220301190954_c5019300-7f32-460b-91bf-f6063311ab34
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1644821811052_0649)

--------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      1          1        0        0       0       0
Map 5 ..........   SUCCEEDED      1          1        0        0       0       0
Reducer 3 ......   SUCCEEDED      1          1        0        0       0       0
Reducer 4 ......   SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 04/04  [==========================>>] 100%  ELAPSED TIME: 7.56 s
--------------------------------------------------------------------------------
Loading data to table default.averages_test
Table default.averages_test stats: [numFiles=1, numRows=20, totalSize=254, rawDataSize=234]
OK
Time taken: 9.608 seconds
Query ID = s1155160788_20220301191004_028aa818-62d5-4531-89d4-1eba8326bdd2
Total jobs = 1
Launching Job 1 out of 1
--------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 01/01  [==========================>>] 100%  ELAPSED TIME: 0.17 s
--------------------------------------------------------------------------------
Moving data to directory /user/s1155160788/output
```

```
Moving data to directory /user/s1155160788/output
OK
Time taken: 0.497 seconds
[s1155160788@dicvmd10 Datasets_Test]$
```

```
A.J.B._NOUN 79
B.c._ADP    68
B'enard 36
B.B.L._NOUN 11
B'midbar    11
A'Aang_NOUN 9
A.Briggs_NOUN   6
B.89_NOUN   6
A.D.A.A.    5
A.E.U._DET  5
A.K.K.  5
A.C.M.S._NOUN   5
A.5.3_DET   5
B.Allison_NOUN  4
A.L.I.V.E.  4
B.N.Z._NOUN 4
B.Pt._NOUN  4
A.IR._NOUN  3
A.165   3
B.5.b   3
```

The data finally appears formatted correctly in the stored hdfs output.

Question 2 Final Submission: Full Script + Output for Full Dataset in Hive

q1.sql q1.pig

C: > Users > Eddie > Documents > School > CUHK > 2021 - 2022 Term 2 > IEMS 5730 - Big Data Systems and Information Processing > Week 7 > Assignment 2 > q1.sql

```sql
CREATE TABLE A_Books (
            bigram STRING,
            year INT,
            match_count INT,
            volume_count INT)
ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t'
stored as textfile;

CREATE TABLE B_Books (
            bigram STRING,
            year INT,
            match_count INT,
            volume_count INT)
row format DELIMITED FIELDS TERMINATED BY '\t'
stored as textfile;

LOAD DATA INPATH '/user/s1155160788/googlebooks-eng-all-1gram-20120701-a' OVERWRITE INTO TABLE A_Books;
LOAD DATA INPATH '/user/s1155160788/googlebooks-eng-all-1gram-20120701-b' OVERWRITE INTO TABLE B_Books;

CREATE TABLE Averages (
            bigram STRING,
            average INT)
row format DELIMITED FIELDS TERMINATED BY '\t'
stored as textfile;

INSERT OVERWRITE TABLE Averages SELECT bigram, avg(match_count) as Average FROM(
    SELECT bigram, match_count FROM A_Books
    UNION ALL
    SELECT bigram, match_count FROM B_Books
) United
GROUP BY bigram
ORDER BY Average DESC
LIMIT 20;

INSERT OVERWRITE directory '/user/s1155160788/output'
row format DELIMITED FIELDS TERMINATED BY '\t'
SELECT * from Averages
```

Screenshot on next page.

```
[s1155160788@dicvmd10 Datasets_Raw]$ hive -f q1.sql\
> quit;
log4j:WARN No such property [maxFileSize] in org.apache.log4j.DailyRollingFileAppender.

Logging initialized using configuration in file:/etc/hive/2.6.5.0-292/0/hive-log4j.properties
hive -f q1.Could not open input file for reading. (File file:/home/s1155160788/Assignment2/Datasets_Raw/q1.sqlquit does not exist)
sq[s1155160788@dicvmd10 Datasets_Raw]$ hive -f q1.sql
log4j:WARN No such property [maxFileSize] in org.apache.log4j.DailyRollingFileAppender.

Logging initialized using configuration in file:/etc/hive/2.6.5.0-292/0/hive-log4j.properties
OK
Time taken: 1.166 seconds
OK
Time taken: 0.215 seconds
Loading data to table default.a_books
chgrp: changing ownership of 'hdfs://dicvmd2.ie.cuhk.edu.hk:8020/apps/hive/warehouse/a_books/googlebooks-eng-all-1gram-20120701-a': User null does not belong to hadoop
Table default.a_books stats: [numFiles=1, numRows=0, totalSize=1801526075, rawDataSize=0]
OK
Time taken: 0.753 seconds
Loading data to table default.b_books
chgrp: changing ownership of 'hdfs://dicvmd2.ie.cuhk.edu.hk:8020/apps/hive/warehouse/b_books/googlebooks-eng-all-1gram-20120701-b': User null does not belong to hadoop
Table default.b_books stats: [numFiles=1, numRows=0, totalSize=1268392934, rawDataSize=0]
OK
Time taken: 0.752 seconds
OK
Time taken: 0.216 seconds
Query ID = s1155160788_20220301192020_266c0d62-9f0f-42aa-96fe-87ef2ae4a564
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1644821811052_0652)

--------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      3          3        0        0       0       0
Map 5 ..........   SUCCEEDED      3          3        0        0       0       0
Reducer 3 ......   SUCCEEDED     12         12        0        0       0       0
Reducer 4 ......   SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 04/04  [==========================>>] 100%  ELAPSED TIME: 87.09 s
--------------------------------------------------------------------------------
Loading data to table default.averages
Table default.averages stats: [numFiles=1, numRows=20, totalSize=274, rawDataSize=254]
OK
Time taken: 89.147 seconds
Query ID = s1155160788_20220301192150_dc8fffd1-2498-42df-9561-1a2402dfebe4
Total jobs = 1
Launching Job 1 out of 1
--------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
```

```
Launching Job 1 out of 1
--------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 01/01  [==========================>>] 100%  ELAPSED TIME: 0.15 s
--------------------------------------------------------------------------------
Moving data to directory /user/s1155160788/output
OK
Time taken: 0.479 seconds
[s1155160788@dicvmd10 Datasets_Raw]$
```

```
q2
C: > Users > Eddie > Documents > Sch
and 25932077
and_CONJ	25906234
a	16665890
a_DET	16645121
as	6179734
be	5629591
be_VERB	5621156
as_ADP	5360443
by	5294067
by_ADP	5272951
are	4298564
are_VERB	4298561
at	3676050
at_ADP	3670625
an	2979272
an_DET	2977977
but	2471102
but_CONJ	2468978
all	2189962
all_DET	2161257
```

I verify the results by running a split screen to the Pig output from question 1 and it is virtually identical, only different on the numbers due to int vs float formatting.

```
q2
C: > Users > Eddie > Documents > School > CUHK > 2021 - 2022 Term 2 > IEMS 5730 - Big Data Systems and Information Prool
and 25932077
and_CONJ	25906234
a	16665890
a_DET	16645121
as	6179734
be	5629591
be_VERB	5621156
as_ADP	5360443
by	5294067
by_ADP	5272951
are	4298564
are_VERB	4298561
at	3676050
at_ADP	3670625
an	2979272
an_DET	2977977
but	2471102
but_CONJ	2468978
all	2189962
all_DET	2161257
```

```
q1
C: > Users > Eddie > Documents > School > CUHK > 2021 - 2022 Term 2 > IEMS 5730 - Big Data
and 2.593207744E7
and_CONJ	2.5906234451764707E7
a	1.6665890811764706E7
a_DET	1.6645121127058823E7
as	6179734.075294117
be	5629591.52
be_VERB	5621156.232941177
as_ADP	5360443.872941176
by	5294067.04
by_ADP	5272951.997647059
are	4298564.341176471
are_VERB	4298561.303529412
at	3676050.1529411767
at_ADP	3670625.785882353
an	2979272.7411764706
an_DET	2977977.8870588234
but	2471102.4964705883
but_CONJ	2468978.0564705883
all	2189962.722352941
all_DET	2161257.294117647
```

Time comparison:

Pig: 530.595 seconds

Hive: 179.962 seconds

Explanation:

Pig directly told me how long it was. I determined the Hive time by adding up all the time statements in the output.

Both Pig and Hive seem to be operating using Tez (at least according to the IE cluster website), so the most obvious explanation for the difference in time is that the Hive script only selects the bigram and match_count columns for the union dataset instead of all 4 columns, which would halve the data that needs to be processed.

It may also be faster because I decided to just use INT's for the Hive implementation instead of needing to do float division.

Hive might also leverage SQL to have greater optimization in the selection and union queries as well.

Question 3: Similar Users Detection in the MovieLens Dataset Using Pig

Question 3 Section A: Output the number of movies each pair of users have watched

Using the Hint for Question 3A in [26], I create mostly the same code and test it with the small dataset.

q3.pig q1.pig

C: > Users > Eddie > Documents > School > CUHK > 2021 - 2022 Term 2 > IEMS 5730 - Big Data Systems and Information Processing > Week 7 > Assignment 2 > q3.pig

```
movielens = LOAD 'movielens_small.csv' USING PigStorage(',') AS (userId:int, movieId:int);
movielens_grpd = GROUP movielens BY movieId;
movielens_grpd_dbl = FOREACH movielens_grpd GENERATE group, movielens.userId AS userId1, movielens.userId AS userId2;
cowatch = FOREACH movielens_grpd_dbl GENERATE FLATTEN(userId1) as userId1, FLATTEN(userId2) as userId2;
cowatch_filtered = FILTER cowatch BY userId1 < userId2;
STORE cowatch_filtered INTO '/user/s1155160788/output' USING PigStorage(',');
```

```
269,270
269,364
269,280
269,323
269,434
269,367
269,587
269,453
269,279
269,322
269,541
269,560
269,339
269,476
269,314
269,412
269,291
269,372
269,341
269,411
269,561
269,414
269,599
```

The code from the hint accomplishes the job of the first MapReduce Job from Part A, as well as the mapper from the second MapReduce job. We go from the raw data to a list of pairs of users that have watched the same movie. Duplicates have been removed by filtering such that the first user in the pair has to be less than the last user in the pair.

At first I thought that the output was sorted by userID1 such that all pairs with the same first user are listed next to each other, however through inspecting of the output, I determined that there will be a number of pairs with the same key at any given time, but they are not all together. It shouldn't matter anyway.

The expected output from the test_movielens.txt data would be something like:

(1,2) 2

(1,3) 1

(1,4) 2

(2,3) 2

(2,4) 2

(3,4) 1

```
cowatch_grouped = GROUP cowatch_filtered BY (userId1, userId2);
cowatch_pairs = FOREACH cowatch_grouped GENERATE group, COUNT(cowatch_filtered.userId1);
```

I add these two lines after some trial and error in the Pig grunt shell, as well as guidance from [27]. We need to group by both user ID's so they remain a pair. The grouping will put every occurrence of the pair into a bag. Then for each grouping of the pairs, we generate group (the pair itself, just once) as well as the count of the first user in the pair. The count of first users in the pair will be equal to the count of the number of copies of each pair.

```
((1,2),2)
((1,3),1)
((1,4),2)
((2,3),2)
((2,4),2)
((3,4),1)
```

Dumping after adding these lines gives us exactly the predicted test dataset output.

```
cowatch_pairs = FOREACH cowatch_grouped GENERATE group, COUNT(cowatch_filtered.userId1) AS group_count;
cowatch_ordered = ORDER cowatch_pairs BY group_count DESC;
Top10 = LIMIT cowatch_ordered 10;
```

I modify the first line here to mention the COUNT as group_count, then add lines to order by count in descending order and limit the output to 10 results based on my Pig code from Question 1.

```
((1,2),2)
((1,4),2)
((2,3),2)
((2,4),2)
((1,3),1)
((3,4),1)
```

The dump looks correct, so I modify the script to store the final results in a comma separated manner.

```
STORE Top10 INTO '/user/s1155160788/output' USING PigStorage(',');
```

```
(1,2),2
(1,4),2
(2,3),2
(2,4),2
(1,3),1
(3,4),1
```

The output from the test looks correct, so I run it on the small and large datasets respectively.

Question 3 Section A Final Submission: Script + Output

Note: I am screenshotting the script for the small dataset. The script for the large dataset is identical but line 1 has a different file name instead.

q3.pig

C: > Users > Eddie > Documents > School > CUHK > 2021 - 2022 Term 2 > IEMS 5730 - Big Data Systems and Information Processing > Week 7 > Assignment 2 > q3.pig

```
1  movielens = LOAD 'movielens_small.csv' USING PigStorage(',') AS (userId:int, movieId:int);
2  movielens_grpd = GROUP movielens BY movieId;
3  movielens_grpd_dbl = FOREACH movielens_grpd GENERATE group, movielens.userId AS userId1, movielens.userId AS userId2;
4  cowatch = FOREACH movielens_grpd_dbl GENERATE FLATTEN(userId1) as userId1, FLATTEN(userId2) as userId2;
5  cowatch_filtered = FILTER cowatch BY userId1 < userId2;
6  cowatch_grouped = GROUP cowatch_filtered BY (userId1, userId2);
7  cowatch_pairs = FOREACH cowatch_grouped GENERATE group, COUNT(cowatch_filtered.userId1) AS group_count;
8  cowatch_ordered = ORDER cowatch_pairs BY group_count DESC;
9  Top10 = LIMIT cowatch_ordered 10;
10 STORE Top10 INTO '/user/s1155160788/output' USING PigStorage(',');
```

Part A Small Output

C: > Users > Eddie > Docume

```
(414,599),1338
(414,474),1077
(68,414),950
(414,448),914
(274,414),856
(474,599),837
(68,599),790
(448,599),790
(274,599),783
(288,414),723
```

Output format is (User1,User2),Number of movies both have watched

I did not modify the format of the output at all from the Pig default because it is sufficiently readable. The same is true for the large dataset output.

The large dataset script completes in 36 minutes and 36 seconds, which is notably faster than Part A in MapReduce. Probably because of query optimizations and Tez performance advantage.

Output format: (User1,User2),Number of movies both have watched

Part A Large Output

C: > Users > Eddie > Documents >

```
(59269,67385),3631
(14463,59269),2631
(59269,68259),2417
(67385,68259),2180
(19635,59269),2148
(27468,59269),2132
(14463,67385),2115
(27468,67385),2002
(19635,67385),1894
(59269,63134),1872
```

The solution for the small dataset is different from the one mentioned in [26], but I will defend my answers. The output for the large dataset here is the same as both the large dataset solution in [26] and the large dataset in my Assignment 1. It would be virtually impossible for the large dataset to have the same answers in 3 different instances (including the official solution) if the logic was incorrect. So the most likely cause of this difference in the small dataset is that the solution in [26] for the small dataset is incorrect, because I have the same output for my small dataset as in the Assignment 1, which received a score of 100, and was proven correct in terms of the large dataset.

Question 3 Section B: Top-K Most Similar Users to Each User

First, we need to get the individual counts for each user.

For the test dataset, the results should be:

User, Count

1,3

2,3

3,2

4,2

```
individual_count_grouped = GROUP movielens BY userId;
individual_count = FOREACH individual_count_grouped GENERATE FLATTEN(group) AS userId, COUNT(movielens.movieId) AS ind_count;
```

After some trial and error in the grunt shell, I create this code which groups by User ID, then generates the user id as well as the count of movies.

```
(1,3)
(2,3)
(3,2)
(4,2)
```

A dump of the output on the test data is as I expected.

```
(592,94)
(593,103)
(594,232)
(595,20)
(596,411)
(597,443)
(598,21)
(599,2478)
(600,763)
(601,101)
(602,135)
(603,943)
(604,100)
(605,221)
(606,1115)
(607,187)
(608,831)
(609,37)
(610,1302)
```

```
555   599 2478
556   6   314
557   60  22
558   600 763
559   601 101
560   602 135
561   603 943
562   604 100
563   605 221
564   606 1115
565   607 187
566   608 831
567   609 37
568   61  39
569   610 1302
570   62  366
```

I also validate the data by taking the dump in Grunt after applying it to the small movielens dataset instead of the test one, and compare it to my MapReduce calculated individual counts for the same dataset in Assignment 1. The values are the same in every tuple.

I now need to join the data from the cowitch pairs to the individual counts.

```
cowatch_flattened = FOREACH cowatch_pairs GENERATE FLATTEN(group) AS (userId1, userId2), group_count;
cowatch_joined_beginner = JOIN cowatch_flattened BY userId1, individual_count BY userId;
cowatch_joined_intermediate = JOIN cowatch_joined_beginner BY userId2, individual_count BY userId;
```

First, I flatten the group, which is the pair of users that have watched a movie together, into their respective user ID's, which I label userId1 and userId2. The 3rd column is the number of movies they have watched together. I join it to the individual count database where the first user ID of cowatch is equal to the user id in the dataset, then do the same with the second userID of cowatch.

```
(1,2,2,1,3,2,3)
(1,3,1,1,3,3,2)
(2,3,2,2,3,3,2)
(1,4,2,1,3,4,2)
(2,4,2,2,3,4,2)
(3,4,1,3,2,4,2)
```

The output is almost as I expected, except the userID column of the individual count tables are joined. We want to remove those in the next step so the final output is something like:

1,2,2,3,3

1,3,1,3,2

2,3,2,3,2

1,4,2,3,2

2,4,2,3,2

3,4,1,2,2

Where the columns are 1:

1: User Id 1

2: User Id 2

3: Number of movies they have watched.

4: Number of movies user 1 has watched.

5: Number of movies user 2 has watched.

```
cowatch_joined = FOREACH cowatch_joined_intermediate GENERATE $0 AS userId1, $1 AS userId2, $2 AS group_count, $4 AS user1Count, $6 AS user2Count;
```

```
(1,2,2,3,3)
(1,3,1,3,2)
(2,3,2,3,2)
(1,4,2,3,2)
(2,4,2,3,2)
(3,4,1,2,2)
```

After adding the line above, the output is as expected. We now have all the elements needed to calculate the similarity between two users.

The similarity calculation should give the following result:

1,2, 0.5

1,3, 0.25

2,3, 0.66

1,4, 0.66

2,4, 0.66

3,4, 0.33

```
cowatch_similarity = FOREACH cowatch_joined GENERATE userId1, userId2, ((float)(group_count)/((float)user1Count+(float)user2Count-(float)group_count)) AS similarity:float
```

```
2022-03-04 09:57:01,441 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_FLOAT 1 time(s).
```

Here, I perform the similarity calculation on each row, and omit all variables but the users and the similarity. At first, I received warnings about implicit cast to floats. Even though I was getting the desired results from just casting (group_count) to float, [29] cautions that it could lead to errors later with joins, order by, etc, and that Pig isn't the best at implicit conversion. So I modified the script to explicitly cast every variable besides the two users to float, and also specified that the output of the final column, similarity, should be a float as well.

After this, I get the desired results with no warnings.

```
(1,2,0.5)
(1,3,0.25)
(2,3,0.6666667)
(1,4,0.6666667)
(2,4,0.6666667)
(3,4,0.33333334)
```

```
92500   1,457,0.06818182
```

I further validated the output by comparing multiple rows of the current output (after applying it to the small dataset) to the output I had in Assignment 1 when calculating the similarity of the small dataset, and the results were identical every time.

```
cowatch_similarity_first = FOREACH cowatch_joined GENERATE userId1, userId2, ((float)(group_count)/((float)user1Count+(float)user2Count-(float)group_count)) AS similarity
cowatch_similarity_second = FOREACH cowatch_joined GENERATE userId2, userId1, ((float)(group_count)/((float)user1Count+(float)user2Count-(float)group_count)) AS similarit
similarity_union = UNION cowatch_similarity_first, cowatch_similarity_second;
```

I revise my similarity calculation line to try something similar to my initial attempts at Part B in Assignment 1 before I decided to just use the unix shell sorting function. If I were to just group the initial similarity calculation by the first user, it would leave out results whenever the user is larger than the second user ID, because those calculations are grouped under the first user. To remedy this, I double the similarity output, doing it again but reversing the user id to ensure that each similarity calculation is included under each user grouping.

Screenshot on next page.

```
(1,2,0.5)
(2,1,0.5)
(1,3,0.25)
(3,1,0.25)
(2,3,0.6666667)
(3,2,0.6666667)
(1,4,0.6666667)
(4,1,0.6666667)
(2,4,0.6666667)
(4,2,0.6666667)
(3,4,0.33333334)
(4,3,0.33333334)
```

```
similarity_ordered = ORDER similarity_union BY similarity DESC;
similarity_grouped = GROUP similarity_ordered BY userId1;
```

I add these lines to order by similarity in descending order, then group the output by the user.

```
(1,{(1,4,0.6666667),(1,2,0.5),(1,3,0.25)})
(2,{(2,3,0.6666667),(2,4,0.6666667),(2,1,0.5)})
(3,{(3,2,0.6666667),(3,4,0.33333334),(3,1,0.25)})
(4,{(4,1,0.6666667),(4,2,0.6666667),(4,3,0.33333334)})
```

We now have each group of users with their similarities in descending order. All users will have all their similarities included because we created the union of the similarity calculation with both users being the key.

The final output should be something like:

1, (4,2,3)

2: (3,4,1)

3: (2,4,1)

4: (1,2,3)

After a lot of trial and error, I come up with the final command to take the top 3 users for each user in terms of similarity, then generate the user followed by a group with all the top 3 users. I was inspired by [30] which notes how ORDER and LIMIT should be used in the FOREACH to work properly with groups.

```
similarity_union = UNION cowatch_similarity_first, cowatch_similarity_second;
similarity_grouped = GROUP similarity_union BY userId1;
top3 = FOREACH similarity_grouped {
    ordered = ORDER similarity_union BY similarity DESC;
    tops = LIMIT ordered 3;
    GENERATE group, tops.userId2;
    }
```

I removed similarity_ordered and moved it to the for loop for each group. The reason being that by ordering everything initially, the output results were right for the group, but when I checked it with the small datasets, the top 3 users were incorrect. I believe this is because when there is small amounts of data, it only needs one partition, and the order can be maintained while it is grouped by key, but when the dataset becomes larger than a trivial size, the order of pairs is not guaranteed, only the fact that it is grouped by key. Therefore, we need to do the ordering after the group is created.

```
(1,{(4),(2),(3)})
(2,{(3),(4),(1)})
(3,{(2),(4),(1)})
(4,{(1),(2),(3)})
```

When running it on the small dataset, I can validate the results by confirming that the output from the Pig script (on the left) exactly matches my output in the final submission of Part B in Assignment 1 for the small dataset.

```
576,{(180),(158),(16)}
577,{(202),(469),(572)}
578,{(506),(12),(83)}
579,{(264),(591),(144)}
580,{(239),(425),(480)}
581,{(300),(471),(497)}
582,{(247),(123),(148)}
583,{(143),(12),(564)}
584,{(174),(321),(94)}
585,{(553),(17),(317)}
586,{(475),(62),(596)}
587,{(156),(57),(469)}
588,{(126),(56),(379)}
589,{(235),(38),(444)}
590,{(68),(480),(288)}
```

Line number (tab) User: User1, User2, User3, in orc

top10SimilaritySmall

C: > Users > Eddie > Documents >

```
088: 581, 378, 247
188: 554, 412, 587
288: 590, 480, 68
388: 539, 49, 598
488: 263, 167, 137
588: 126, 56, 379
```

Question 3 Section B Final Submission: Full Script + Output

The output format of the Pig script is User,{(Top1),(Top2),(Top3)}, and outputs for all users in the database.

For my final submission, I take the output of the Pig Script for the relevant users that share the student ID and modify it into the format of User: Top1, Top2, Top3.

```
q3b.pig
C: > Users > Eddie > Documents > School > CUHK > 2021 - 2022 Term 2 > IEMS 5730 - Big Data Systems and Information Processing > Week 7 > Assignment 2 > q3b.pig
1  movielens = LOAD 'movielens_small.csv' USING PigStorage(',') AS (userId:int, movieId:int);
2  movielens_grpd = GROUP movielens BY movieId;
3  movielens_grpd_dbl = FOREACH movielens_grpd GENERATE group, movielens.userId AS userId1, movielens.userId AS userId2;
4  cowatch = FOREACH movielens_grpd_dbl GENERATE FLATTEN(userId1) as userId1, FLATTEN(userId2) as userId2;
5  cowatch_filtered = FILTER cowatch BY userId1 < userId2;
6  cowatch_grouped = GROUP cowatch_filtered BY (userId1, userId2);
7  cowatch_pairs = FOREACH cowatch_grouped GENERATE group, COUNT(cowatch_filtered.userId1) AS group_count;
8  individual_count_grouped = GROUP movielens BY userId;
9  individual_count = FOREACH individual_count_grouped GENERATE FLATTEN(group) AS userId, COUNT(movielens.movieId) AS ind_count;
10 cowatch_flattened = FOREACH cowatch_pairs GENERATE FLATTEN(group) AS (userId1, userId2), group_count;
11 cowatch_joined_beginner = JOIN cowatch_flattened BY userId1, individual_count BY userId;
12 cowatch_joined_intermediate = JOIN cowatch_joined_beginner BY userId2, individual_count BY userId;
13 cowatch_joined = FOREACH cowatch_joined_intermediate GENERATE $0 AS userId1, $1 AS userId2, $2 AS group_count, $4 AS user1Count, $6 AS user2Count;
14 cowatch_similarity_first = FOREACH cowatch_joined GENERATE userId1, userId2, ((float)(group_count)/((float)user1Count+(float)user2Count-(float)group_count)) AS similarity
15 cowatch_similarity_second = FOREACH cowatch_joined GENERATE userId2, userId1, ((float)(group_count)/((float)user1Count+(float)user2Count-(float)group_count)) AS similarit
16 similarity_union = UNION cowatch_similarity_first, cowatch_similarity_second;
17 similarity_grouped = GROUP similarity_union BY userId1;
18 top3 = FOREACH similarity_grouped {
19     ordered = ORDER similarity_union BY similarity DESC;
20     tops = LIMIT ordered 3;
21     GENERATE group, tops.userId2;
22     }
23 STORE top3 INTO '/user/s1155160788/output' USING PigStorage(',');
```

The end of lines 14 and 15 is “AS similarity:float”

```
Input(s):
Successfully read 1569491 records (17103777 bytes) from: "hdfs://dicvmd2.ie.cuhk.edu.hk:8020/user/s1155160788/movielens_large_updated.csv"

Output(s):
Successfully stored 53184 records (1669050 bytes) in: "/user/s1155160788/output"

2022-03-04 16:09:33,771 [main] INFO  org.apache.pig.Main - Pig script completed in 51 minutes, 8 seconds and 896 milliseconds (3068896 ms)
2022-03-04 16:09:33,775 [main] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezLauncher - Shutting down thread pool
2022-03-04 16:09:33,783 [pool-1-thread-1] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezSessionManager - Shutting down Tez session org.apache.tez.client.TezClient@167c5832
2022-03-04 16:09:33 Shutting down Tez session , sessionName=PigLatin:q3b.pig, applicationId=application_1644821811052_0996
2022-03-04 16:09:33,783 [pool-1-thread-1] INFO  org.apache.tez.client.TezClient - Shutting down Tez Session, sessionName=PigLatin:q3b.pig, applicationId=application_1644821811052_0996
[s1155160788@dicvmd10 Assignment2]$
```

For the small database, I need to output the top 3 similar users for 88, 188, 288, 388, 488, and 588. My student ID is 1155160788, so the last 2 digits is 88.

Small Dataset

```
q3b small
C: > Users > Eddie > Documents > S
1   088: 581, 378, 247
2   188: 554, 412, 587
3   288: 590, 480, 68
4   388: 539, 598, 49
5   488: 263, 167, 137
6   588: 126, 56, 379
```

Compared to the MapReduce output from Assignment 1, it is exactly identical except for user 388, it switches the order of 598 and 49. This is likely because they were tied in similarity, and it chose randomly or opportunistically which one to output first.

Large dataset on next page.

Large dataset:

```
q3b large
C: > Users > Eddie > Documents > School > CUHK > 2021 - 2022 Term 2 > IEN
788: 53822, 30449, 25907
10790 (10788 does not exist): 5152, 37500, 44521
20788: 52182, 61056, 23991
30788: 29471, 43384, 13358
40788: 19514, 24218, 25984
50789 (50788 does not exist): 65987, 41871, 17793
60788: 56437, 10077, 50594
70789 (70788 does not exist): 15418, 54351, 924
```

Here, there is more difference compared to the output of Assignment 1, but I still believe it is correct. This dataset seems to have a greater incidence of tied similarities than the smaller dataset, so within the top 3, there is some arbitrariness in terms of ordering and what is selected if there is tied similarity. Luckily, in every user, at least one of the users is shared with my output from Assignment 1 to give me some confidence in the correctness.

References

[1] https://pig.apache.org/docs/r0.17.0/start.html#Pig+Setup

[2] https://dlcdn.apache.org/pig/pig-0.17.0/

[3] https://dlcdn.apache.org/pig/pig-0.17.0/RELEASE_NOTES.txt

[4] https://stackoverflow.com/questions/19943766/hadoop-unable-to-load-native-hadoop-library-for-your-platform-warning

[5] https://linuxize.com/post/how-to-unzip-gz-file/

[6] https://stackoverflow.com/questions/36378568/how-to-find-hadoop-hdfs-directory-on-my-system

[7] https://pig.apache.org/docs/r0.17.0/basic.html

[8] A. Gates, *Programming Pig, 2nd Edition*. O'Reilly Media, Inc., 2016.

[9] https://timepasstechies.com/pig-tutorial-5-debugging-pig-diagnostic-operators-like-describe-dump-explain-illustrate/

[10] https://squarecog.wordpress.com/2010/05/11/group-operator-in-apache-pig/

[11] https://stackoverflow.com/questions/9247296/select-count-distinct-using-pig-latin

[12] https://mobitec.ie.cuhk.edu.hk/iems5730/static_files/slides/PigHiveIERG4330IEMS5730Spring2022.pdf

[13] https://cwiki.apache.org/confluence/display/Hive/GettingStarted

[14] https://stackoverflow.com/questions/30700970/hadoop-setting-the-hadoop-home-correctly-to-bin-hadoop-it-gives-command-not-foun

[15] https://stackoverflow.com/questions/11302305/how-to-know-hive-and-hadoop-versions-from-command-prompt

[16] https://stackoverflow.com/questions/39188260/how-to-get-the-default-hive-home-in-cloudera-quickstartvm-5-7-0/39188345

[17] https://mobitec.ie.cuhk.edu.hk/iems5730/tutorials/03_pig_hive/

[18] http://hortonworks.com/wp-content/uploads/2016/05/Hortonworks.CheatSheet.SQLtoHive.pdf

[19] https://stackoverflow.com/questions/50001124/why-does-all-columns-get-created-as-string-when-i-use-opencsvserde-in-hive

[20] https://lists.apache.org/thread/zt8b05mj7xb5l2jvnbng4qvrqgrkgcxj

[21] https://sparkbyexamples.com/apache-hive/hive-insert-into-vs-insert-overwrite/

[22] https://cwiki.apache.org/confluence/display/Hive/LanguageManual+Union

[23] https://stackoverflow.com/questions/49925/what-is-the-difference-between-union-and-union-all

[24] https://www.guru99.com/hive-queries-implementation.html

[25] https://www.edureka.co/community/6052/hive-to-hdfs-transfer

[26] https://mobitec.ie.cuhk.edu.hk/ierg4330/tutorials/05_hw1_mr/

[27] https://stackoverflow.com/questions/25012396/counting-elements-for-each-group-using-pig

[28] https://www.hdfstutorial.com/blog/pig-join-example-2/

[29] https://stackoverflow.com/questions/38949958/getting-an-pig-errorencountered-warning-implicit-cast-to-long-2-times

[30] https://stackoverflow.com/questions/13631909/hadoop-pig-return-top-5-rows