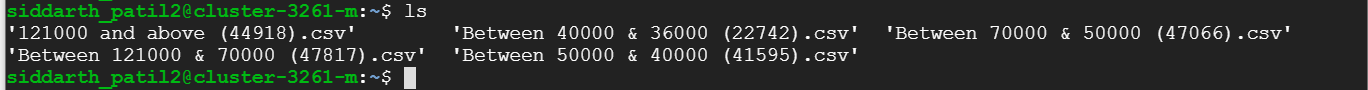
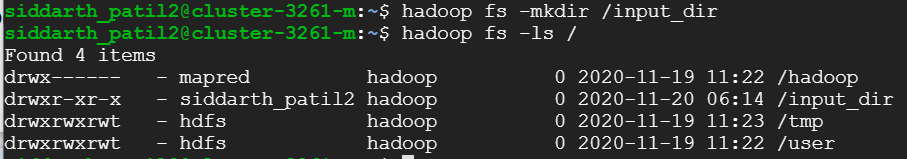
1. Uploading the data to Local in GCP using. (Naming convention isn’t good, to be tackled later)

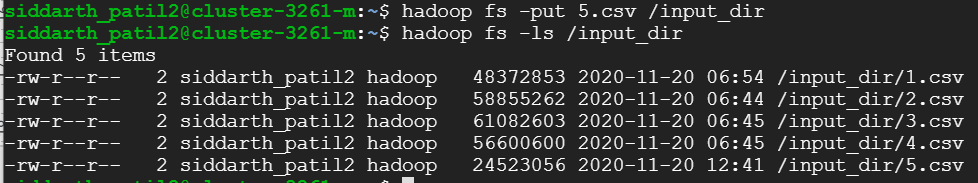


1. To tackle the naming convention, used **“mv”** command to rename all the files. Sample is shown below.



1. Moving the files to HDFS in the newly made directory “/input\_dir” and verifying.





1. To go into **pig** mode (grunt) used the command **pig.**
2. As the dataset has multiple line inputs, and also contains “,” in the row values I made use of the below commands to tackle these.

**Source:** <https://pig.apache.org/docs/r0.17.0/api/org/apache/pig/piggybank/storage/CSVExcelStorage.html>

**Commands:**

file1 = LOAD 'hdfs://cluster-cb25-m/input\_dir/1.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',', 'YES\_MULTILINE', 'UNIX', 'SKIP\_INPUT\_HEADER') AS (Id:int, PostTypeId:int, AcceptedAnswerId:int, ParentId:int, CreationDate:chararray, DeletionDate:chararray, Score:int, ViewCount:int, Body:chararray, OwnerUserId:int, OwnerDisplayName:chararray, LastEditorUserId:int, LastEditorDisplayName:chararray, LastEditDate:chararray, LastActivityDate:chararray, Title:chararray, Tags:chararray, AnswerCount:int, CommentCount:int, FavoriteCount:int, ClosedDate:chararray, CommunityOwnedDate:chararray, ContentLicense:chararray );

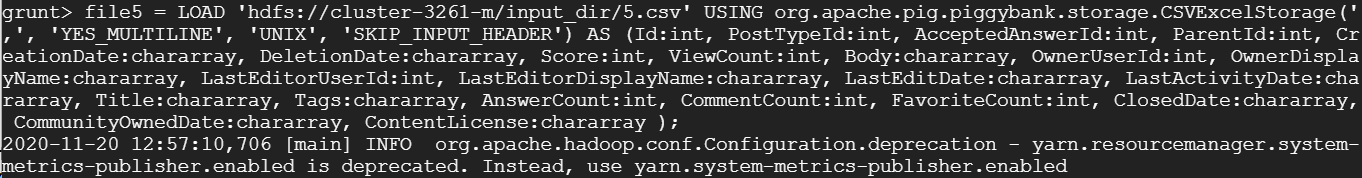
file2 = LOAD 'hdfs://cluster-cb25-m/input\_dir/2.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',', 'YES\_MULTILINE', 'UNIX', 'SKIP\_INPUT\_HEADER') AS (Id:int, PostTypeId:int, AcceptedAnswerId:int, ParentId:int, CreationDate:chararray, DeletionDate:chararray, Score:int, ViewCount:int, Body:chararray, OwnerUserId:int, OwnerDisplayName:chararray, LastEditorUserId:int, LastEditorDisplayName:chararray, LastEditDate:chararray, LastActivityDate:chararray, Title:chararray, Tags:chararray, AnswerCount:int, CommentCount:int, FavoriteCount:int, ClosedDate:chararray, CommunityOwnedDate:chararray, ContentLicense:chararray );

file3 = LOAD 'hdfs://cluster-cb25-m/input\_dir/3.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',', 'YES\_MULTILINE', 'UNIX', 'SKIP\_INPUT\_HEADER') AS (Id:int, PostTypeId:int, AcceptedAnswerId:int, ParentId:int, CreationDate:chararray, DeletionDate:chararray, Score:int, ViewCount:int, Body:chararray, OwnerUserId:int, OwnerDisplayName:chararray, LastEditorUserId:int, LastEditorDisplayName:chararray, LastEditDate:chararray, LastActivityDate:chararray, Title:chararray, Tags:chararray, AnswerCount:int, CommentCount:int, FavoriteCount:int, ClosedDate:chararray, CommunityOwnedDate:chararray, ContentLicense:chararray );

file4 = LOAD 'hdfs://cluster-cb25-m/input\_dir/4.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',', 'YES\_MULTILINE', 'UNIX', 'SKIP\_INPUT\_HEADER') AS (Id:int, PostTypeId:int, AcceptedAnswerId:int, ParentId:int, CreationDate:chararray, DeletionDate:chararray, Score:int, ViewCount:int, Body:chararray, OwnerUserId:int, OwnerDisplayName:chararray, LastEditorUserId:int, LastEditorDisplayName:chararray, LastEditDate:chararray, LastActivityDate:chararray, Title:chararray, Tags:chararray, AnswerCount:int, CommentCount:int, FavoriteCount:int, ClosedDate:chararray, CommunityOwnedDate:chararray, ContentLicense:chararray );

file5 = LOAD 'hdfs://cluster-cb25-m/input\_dir/5.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',', 'YES\_MULTILINE', 'UNIX', 'SKIP\_INPUT\_HEADER') AS (Id:int, PostTypeId:int, AcceptedAnswerId:int, ParentId:int, CreationDate:chararray, DeletionDate:chararray, Score:int, ViewCount:int, Body:chararray, OwnerUserId:int, OwnerDisplayName:chararray, LastEditorUserId:int, LastEditorDisplayName:chararray, LastEditDate:chararray, LastActivityDate:chararray, Title:chararray, Tags:chararray, AnswerCount:int, CommentCount:int, FavoriteCount:int, ClosedDate:chararray, CommunityOwnedDate:chararray, ContentLicense:chararray );

**Screenshot:**



1. As the files are loaded in different relations, **UNION** command can be used to join them.

**Source:** <https://www.tutorialspoint.com/apache_pig/apache_pig_union_operator.htm>

**Code:** combined\_file = UNION file1, file2, file3, file4, file5;

**Screenshot:** 

1. Transforming the data as per out requirements. Only kept **eight** columns and omitted the rest.

**Column kept:** Title, Body, Tag, Score, Id, ViewCount, OwnerUserId and OwnerDisplayName.

**Source:** <https://stackoverflow.com/questions/39944522/pig-script-to-replace-any-sort-of-line-break-from-a-column>

**Code:**

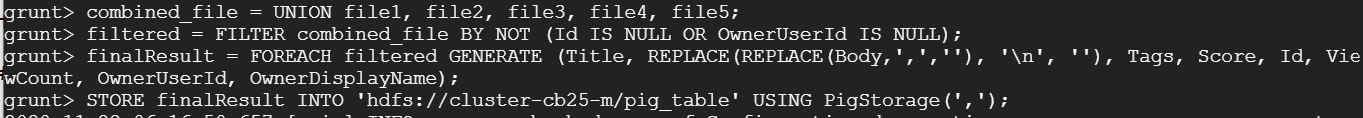
combined\_file = UNION file1, file2, file3, file4, file5;

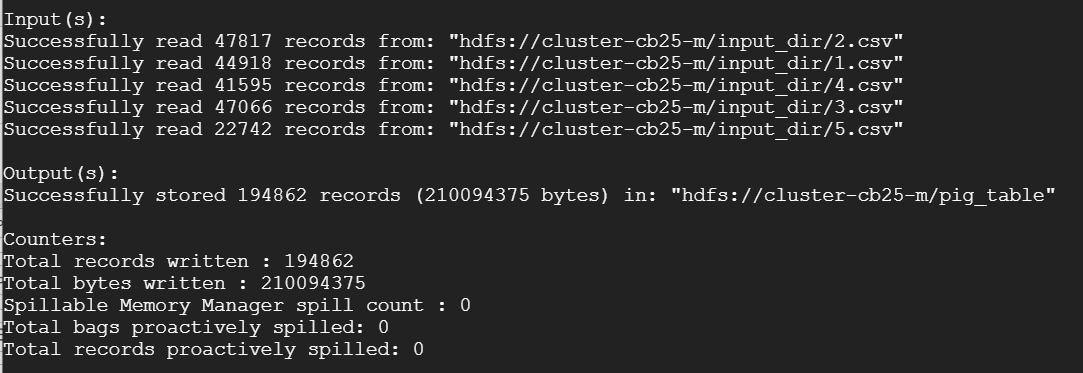
filtered = FILTER combined\_file BY NOT (Id IS NULL OR OwnerUserId IS NULL);

finalResult = FOREACH filtered GENERATE (Title, REPLACE(REPLACE(Body,',',''), '\n', ''), Tags, Score, Id, ViewCount, OwnerUserId, OwnerDisplayName);

STORE finalResult INTO 'hdfs://cluster-cb25-m/pig\_table' USING PigStorage(',');

**Screenshot:**





1. Checking if the file was stored.

**Code:** Hadoop fs -ls /pig\_table

**Screenshot:**

