Task 1: HADOOP, HDFS, PIG and HIVE

1. Import RStudio Log Files from one week in February 2019 into HDFS

Step-1 Download one week Logs from http://cran-logs.rstudio.com/

Step-2 Create a new folder to store downloaded logs files. mkdir RStudioLogs

Step-3 Verify files in RStudioLogs folder ls

Step-3 Unzip all *.zip files gunzip *.csv.gz

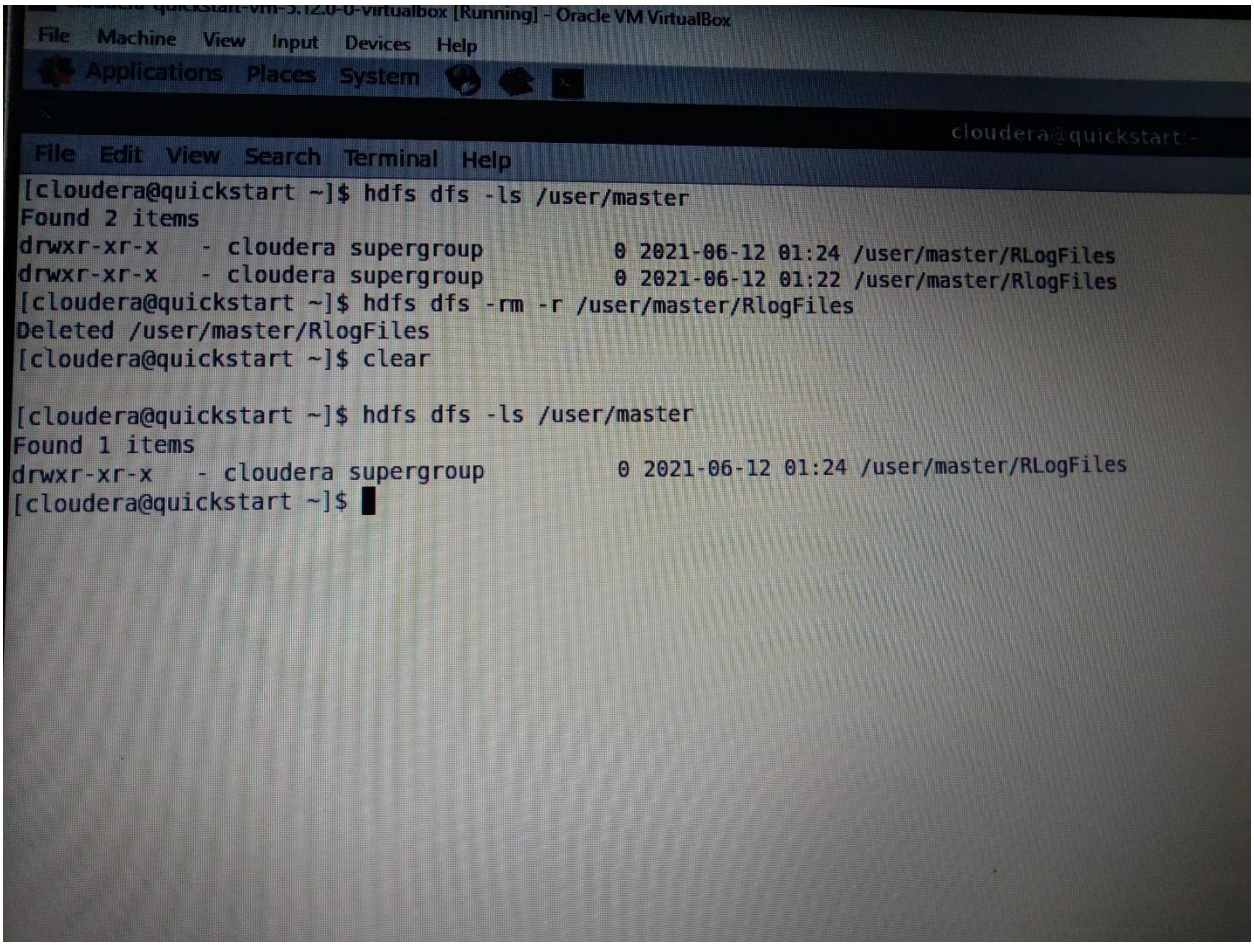
Step-4 Put on ?HDFS and verify

hdfs dfs -mkdir /user/master

hdfs dfs -mkdir /user/master/RLogFiles

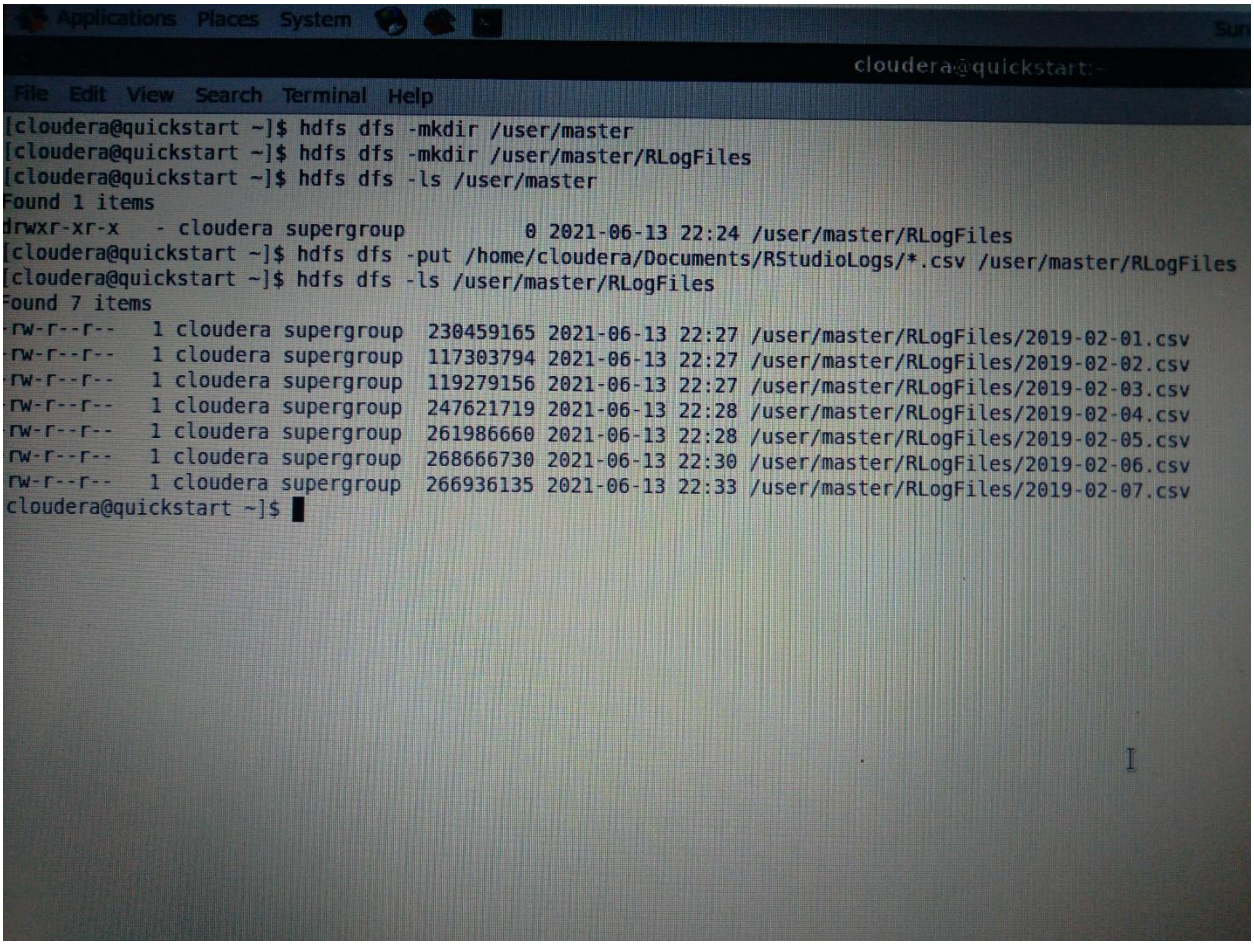
hdfs dfs -put *.csv /user/master/RLogFiles

hdfs dfs -ls /user/master/RlogFiles



Loaded files to RLogFiles : hdfs dfs –put /home/cloudera/Documents/RStudioLogs/*.csv /user/master/RLogFiles

Hdfs dfs –ls /user/master/RLogFiles

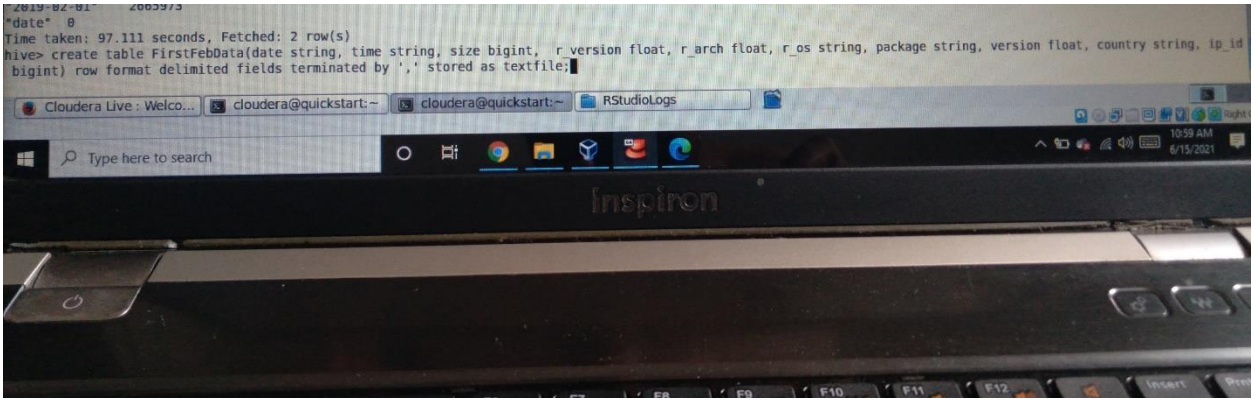


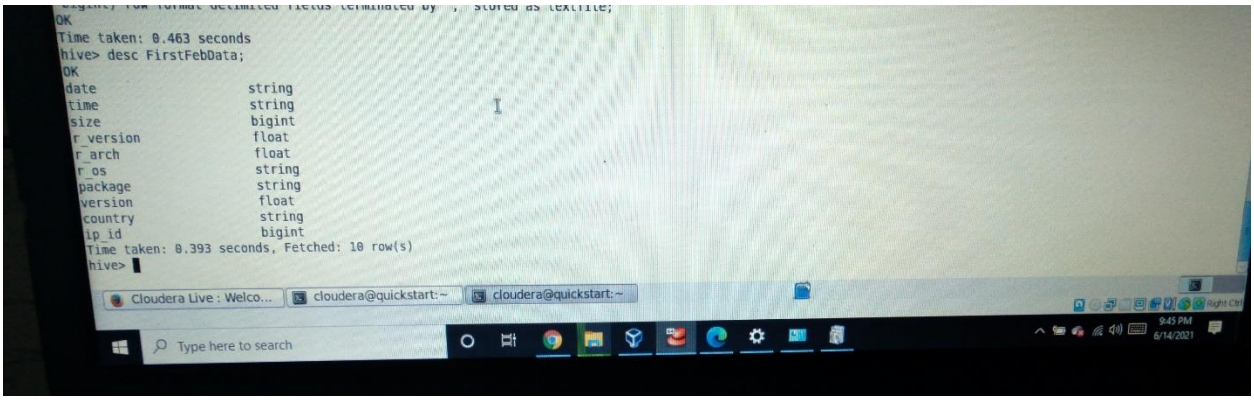
```
[cloudera@quickstart ~]$ hdfs dfs -mkdir /user/master
[cloudera@quickstart ~]$ hdfs dfs -mkdir /user/master/RLogFiles
[cloudera@quickstart ~]$ hdfs dfs -ls /user/master
Found 1 items
drwxr-xr-x   - cloudera supergroup          0 2021-06-13 22:24 /user/master/RLogFiles
[cloudera@quickstart ~]$ hdfs dfs -put /home/cloudera/Documents/RStudioLogs/*.csv /user/master/RLogFiles
[cloudera@quickstart ~]$ hdfs dfs -ls /user/master/RLogFiles
Found 7 items
-rw-r--r--   1 cloudera supergroup  230459165 2021-06-13 22:27 /user/master/RLogFiles/2019-02-01.csv
-rw-r--r--   1 cloudera supergroup  117303794 2021-06-13 22:27 /user/master/RLogFiles/2019-02-02.csv
-rw-r--r--   1 cloudera supergroup  119279156 2021-06-13 22:27 /user/master/RLogFiles/2019-02-03.csv
-rw-r--r--   1 cloudera supergroup  247621719 2021-06-13 22:28 /user/master/RLogFiles/2019-02-04.csv
-rw-r--r--   1 cloudera supergroup  261986660 2021-06-13 22:28 /user/master/RLogFiles/2019-02-05.csv
-rw-r--r--   1 cloudera supergroup  268666730 2021-06-13 22:30 /user/master/RLogFiles/2019-02-06.csv
-rw-r--r--   1 cloudera supergroup  266936135 2021-06-13 22:33 /user/master/RLogFiles/2019-02-07.csv
[cloudera@quickstart ~]$
```

Load Log Files of one day

Create table FirstFeBData in HIVE

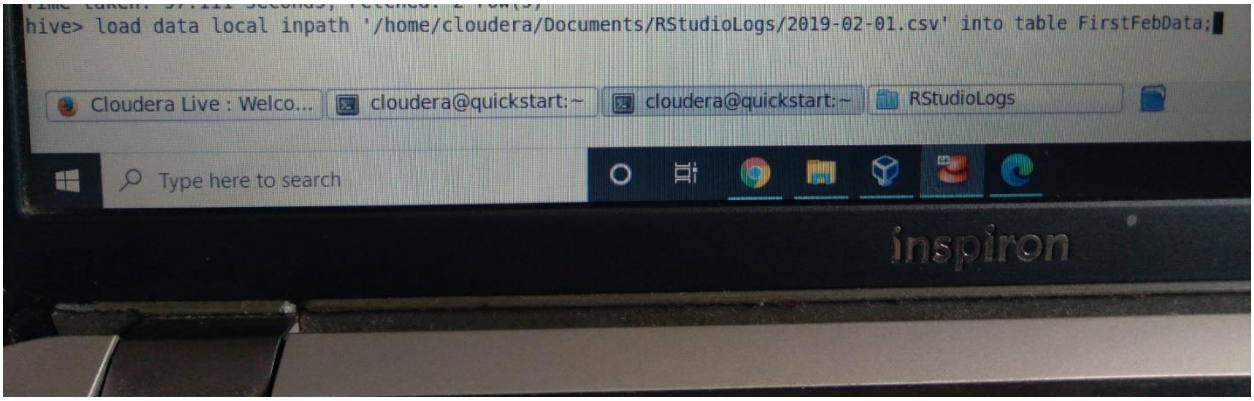
create table FirstFebData(date string, time string, size bigint, r_version float, r_arch float, r_os string, package string, version float, country string, ip_id bigint) row format delimited fields terminated by ',' stored as textfiles;





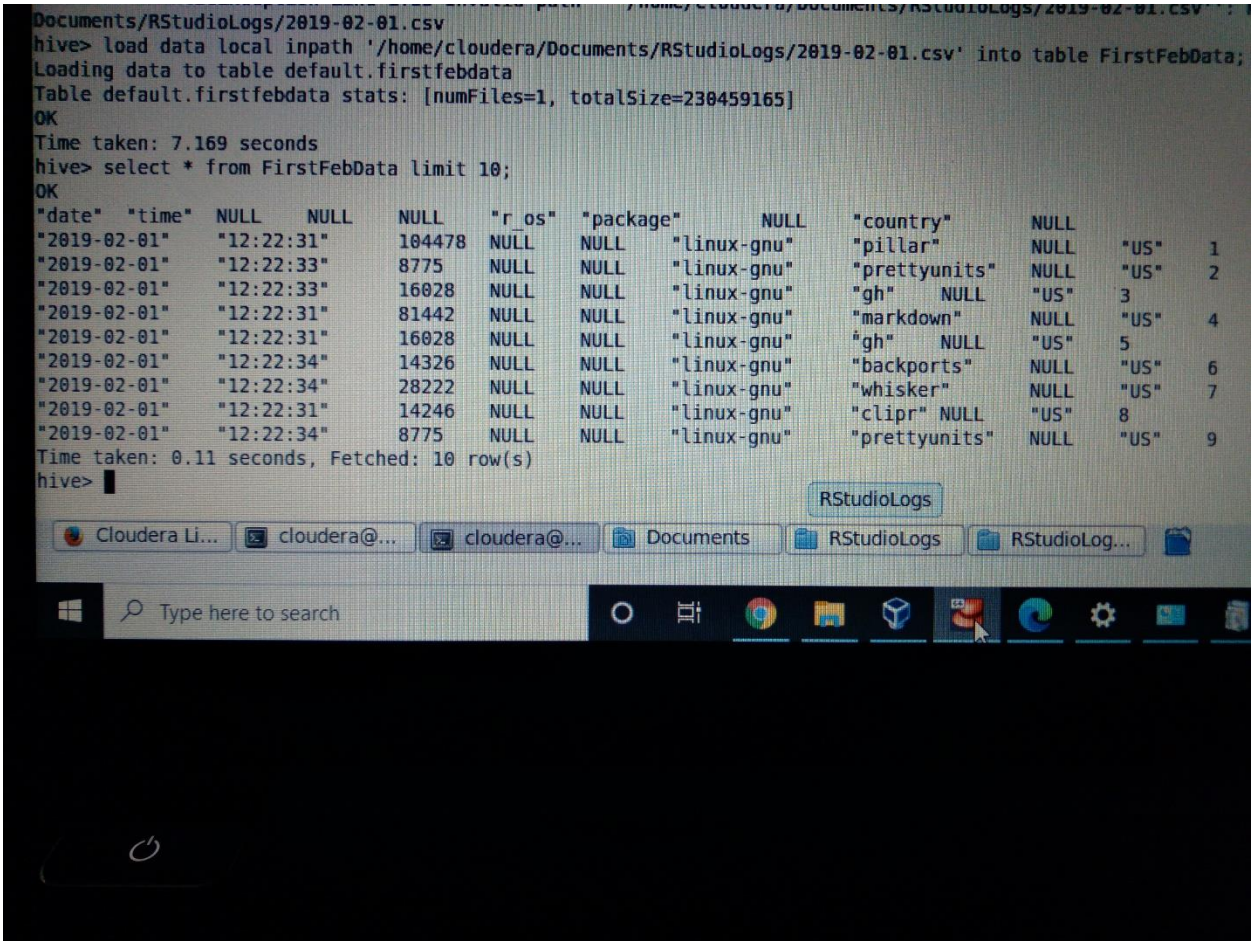
Load Single file to table FirstFebData

load data inpath ‘ /home/cloudera/Documents/RStudioLogs/2019-02-01.csv’ overwrite into table FirstFebData;



Dump the first 10 entries on screen to check if it works or not

select * from FirstFebData limit 10;



```
Documents/RStudioLogs/2019-02-01.csv
hive> load data local inpath '/home/cloudera/Documents/RStudioLogs/2019-02-01.csv' into table FirstFebData;
Loading data to table default.firstfebdata
Table default.firstfebdata stats: [numFiles=1, totalSize=230459165]
OK
Time taken: 7.169 seconds
hive> select * from FirstFebData limit 10;
OK
"date"  "time"  NULL    NULL    NULL    "r_os"  "package"       NULL    "country"       NULL
"2019-02-01"    "12:22:31"      104478  NULL    NULL    "linux-gnu"     "pillar"        NULL    "US"    1
"2019-02-01"    "12:22:33"      8775    NULL    NULL    "linux-gnu"     "prettyunits"   NULL    "US"    2
"2019-02-01"    "12:22:33"      16028   NULL    NULL    "linux-gnu"     "gh"    NULL    "US"    3
"2019-02-01"    "12:22:31"      81442   NULL    NULL    "linux-gnu"     "markdown"      NULL    "US"    4
"2019-02-01"    "12:22:31"      16028   NULL    NULL    "linux-gnu"     "gh"    NULL    "US"    5
"2019-02-01"    "12:22:34"      14326   NULL    NULL    "linux-gnu"     "backports"     NULL    "US"    6
"2019-02-01"    "12:22:34"      28222   NULL    NULL    "linux-gnu"     "whisker"       NULL    "US"    7
"2019-02-01"    "12:22:31"      14246   NULL    NULL    "linux-gnu"     "clipr" NULL    "US"    8
"2019-02-01"    "12:22:34"      8775    NULL    NULL    "linux-gnu"     "prettyunits"   NULL    "US"    9
Time taken: 0.11 seconds, Fetched: 10 row(s)
hive>
```

Count number of occurrences of different packages

Select package, count(*) from FirstFebData group by package;

```
Time taken: 97.111 seconds, Fetched: 2 row(s)
hive> select package, count(*) from FirstFebData group by package;
Query ID = cloudera_20210614223434_586726a1-f043-4fbf-b610-e752bd02adb4
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1623675726543_0008, Tracking URL = http://quickstart.cloudera
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1623675726543_0008
```

```
"zeitgebr"      8
"zeligverse"    62
"zendeskR"      8
"zenplots"      5
"zeroEQpart"    4
"zetadiv"       4
"zfa"   4
"zic"   7
"zip"   5139
"zipR"  4
"zipcode"       216
"zipfR" 22
"zipfextR"      5
"ziphsmm"       4
"zoeppritz"     5
"zoib"  7
"zonator"       7
"zoo"   14080
"zooaRch"       6
"zooaRchGUI"    4
"zoocat"        5
"zooimage"      5
"zoom"  275
"zoomgrid"      3
"zoon"  4
"zscorer"       6
"ztable"        86
"ztype" 5
"zyp"   29
NA      20
Time taken: 98.776 seconds, Fetched: 14320 row(s)
hive>
```

Count number of Occurrences of different packages by operating system

```
    > ;
OK
date                    string
time                    string
size                    bigint
r_version               float
r_arch                  float
r_os                    string
package                 string
version                 float
country                 string
ip_id                   bigint
Time taken: 0.308 seconds, Fetched: 10 row(s)
hive> select r_os, count(*) from FirstFebData group by r_os;
Query ID = cloudera_20210614215353_7c56f869-bd99-47a3-8e9b-30a7dc62b17d
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1623675726543_0004, Tracking URL = http://quickstart.cloudera:
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1623675726543_0004
```

Cloudera Li... | cloudera@... | cloudera@... | Documents | RStudioLogs

```
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 17.12 sec   HDFS Read: 230471242 HDFS
Total MapReduce CPU Time Spent: 17 seconds 120 msec
OK
"cygwin"        33
"darwin10.8.0"  146
"darwin11.0.0"  4
"darwin11.4.2"  38
"darwin13.4.0"  86800
"darwin14.5.0"  212
"darwin15.6.0"  302300
"darwin16.1.0"  2
"darwin16.5.0"  105
"darwin16.6.0"  11
"darwin16.7.0"  656
"darwin17.0.0"  2
"darwin17.2.0"  29
"darwin17.3.0"  98
"darwin17.4.0"  507
"darwin17.5.0"  292
"darwin17.6.0"  1053
"darwin17.7.0"  1527
"darwin18.0.0"  489
"darwin18.2.0"  4695
"freebsd11.2"   1
"freebsd12.0"   2
"linux-gnu"     883088
"linux-gnueabihf"       564
"mingw32"       1304342
"r_os"  1
NA      78977
Time taken: 89.523 seconds, Fetched: 27 row(s)
hive>
```