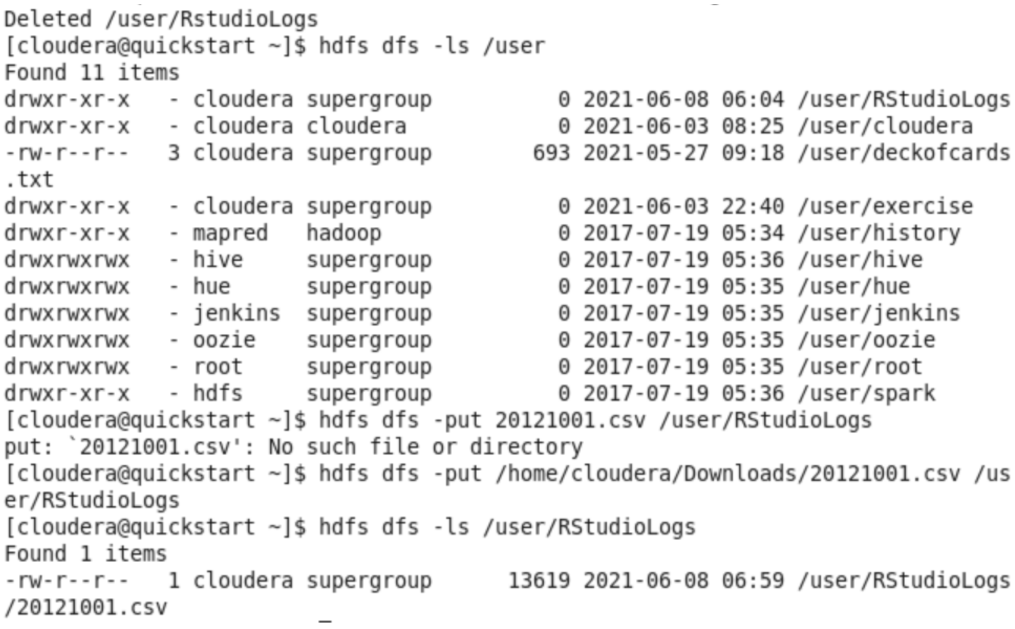
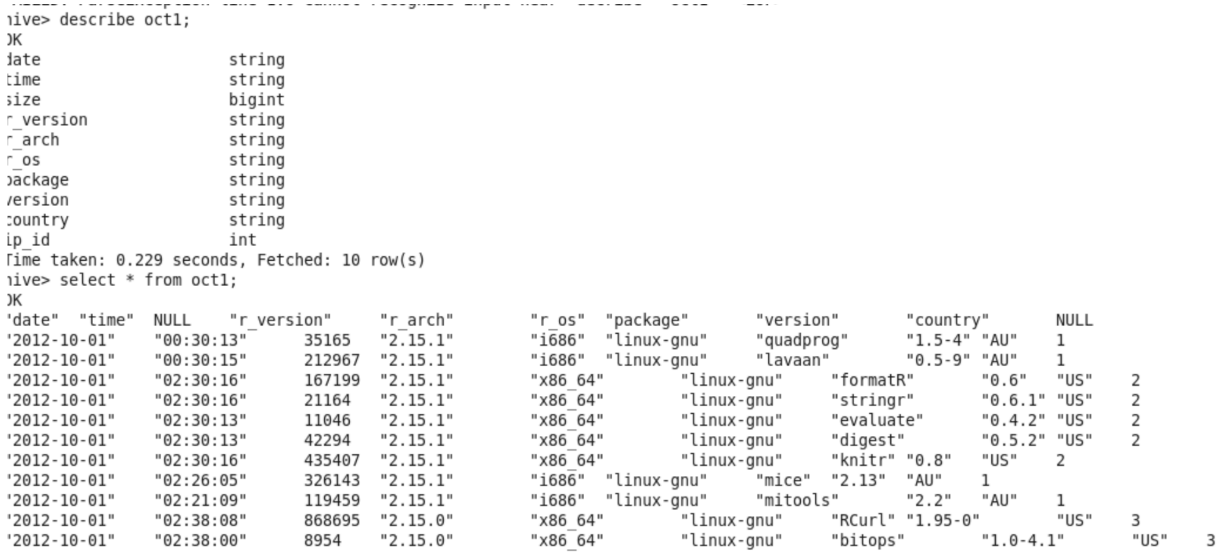
Hive Task 2:

1. Import RStudio Log Files from into HDFS.

Step 1 : Downloaded the file 2012-10-01
Step 2: Import the file in HDFS under RStudioLogs directory

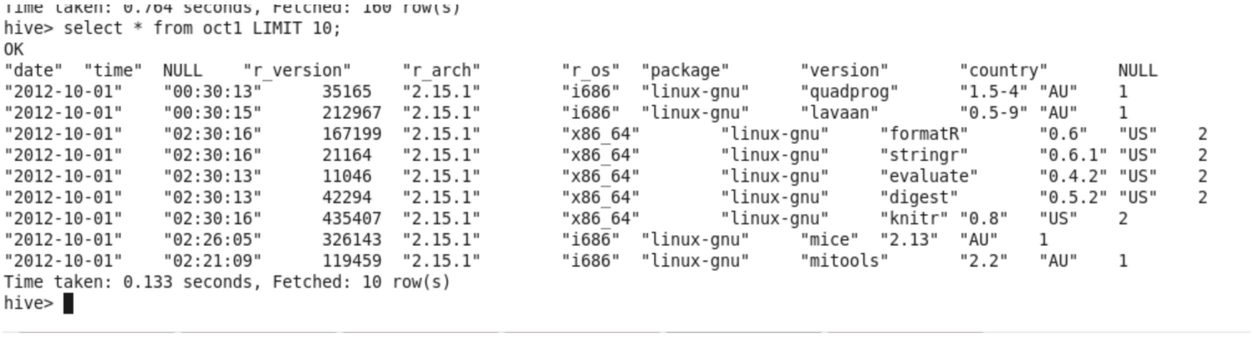


```
Deleted /user/RstudioLogs
[cloudera@quickstart ~]$ hdfs dfs -ls /user
Found 11 items
drwxr-xr-x   - cloudera supergroup          0 2021-06-08 06:04 /user/RStudioLogs
drwxr-xr-x   - cloudera cloudera            0 2021-06-03 08:25 /user/cloudera
-rw-r--r--   3 cloudera supergroup        693 2021-05-27 09:18 /user/deckofcards
.txt
drwxr-xr-x   - cloudera supergroup          0 2021-06-03 22:40 /user/exercise
drwxr-xr-x   - mapred   hadoop              0 2017-07-19 05:34 /user/history
drwxrwxrwx   - hive     supergroup          0 2017-07-19 05:36 /user/hive
drwxrwxrwx   - hue      supergroup          0 2017-07-19 05:35 /user/hue
drwxrwxrwx   - jenkins  supergroup          0 2017-07-19 05:35 /user/jenkins
drwxrwxrwx   - oozie    supergroup          0 2017-07-19 05:35 /user/oozie
drwxrwxrwx   - root     supergroup          0 2017-07-19 05:35 /user/root
drwxr-xr-x   - hdfs     supergroup          0 2017-07-19 05:36 /user/spark
[cloudera@quickstart ~]$ hdfs dfs -put 20121001.csv /user/RStudioLogs
put: `20121001.csv': No such file or directory
[cloudera@quickstart ~]$ hdfs dfs -put /home/cloudera/Downloads/20121001.csv /us
er/RStudioLogs
[cloudera@quickstart ~]$ hdfs dfs -ls /user/RStudioLogs
Found 1 items
-rw-r--r--   1 cloudera supergroup      13619 2021-06-08 06:59 /user/RStudioLogs
/20121001.csv
```



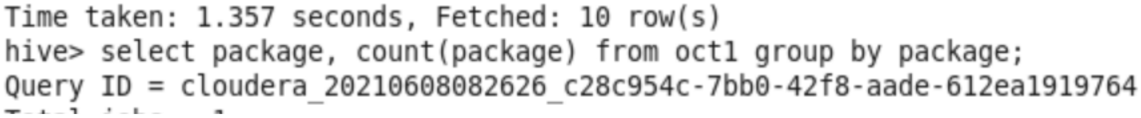
```
ıive> describe oct1;
)K
ıate                    string
time                    string
size                    bigint
r_version               string
r_arch                  string
r_os                    string
package                 string
version                 string
country                 string
ip_id                   int
Time taken: 0.229 seconds, Fetched: 10 row(s)
ıive> select * from oct1;
)K
'date"  "time"  NULL    "r_version"     "r_arch"        "r_os"  "package"       "version"       "country"       NULL
'2012-10-01"    "00:30:13"      35165   "2.15.1"        "i686"  "linux-gnu"     "quadprog"      "1.5-4" "AU"    1
'2012-10-01"    "00:30:15"      212967  "2.15.1"        "i686"  "linux-gnu"     "lavaan"        "0.5-9" "AU"    1
'2012-10-01"    "02:30:16"      167199  "2.15.1"        "x86_64"        "linux-gnu"     "formatR"       "0.6"   "US"    2
'2012-10-01"    "02:30:16"      21164   "2.15.1"        "x86_64"        "linux-gnu"     "stringr"       "0.6.1" "US"    2
'2012-10-01"    "02:30:13"      11046   "2.15.1"        "x86_64"        "linux-gnu"     "evaluate"      "0.4.2" "US"    2
'2012-10-01"    "02:30:13"      42294   "2.15.1"        "x86_64"        "linux-gnu"     "digest"        "0.5.2" "US"    2
'2012-10-01"    "02:30:16"      435407  "2.15.1"        "x86_64"        "linux-gnu"     "knitr" "0.8"   "US"    2
'2012-10-01"    "02:26:05"      326143  "2.15.1"        "i686"  "linux-gnu"     "mice"  "2.13"  "AU"    1
'2012-10-01"    "02:21:09"      119459  "2.15.1"        "i686"  "linux-gnu"     "mitools"       "2.2"   "AU"    1
'2012-10-01"    "02:38:08"      868695  "2.15.0"        "x86_64"        "linux-gnu"     "RCurl" "1.95-0"        "US"    3
'2012-10-01"    "02:38:00"      8954    "2.15.0"        "x86_64"        "linux-gnu"     "bitops"        "1.0-4.1"       "US"    3
```

Dump the first 10 entries on screen (attach a screen shot into your report) to check if it works or not

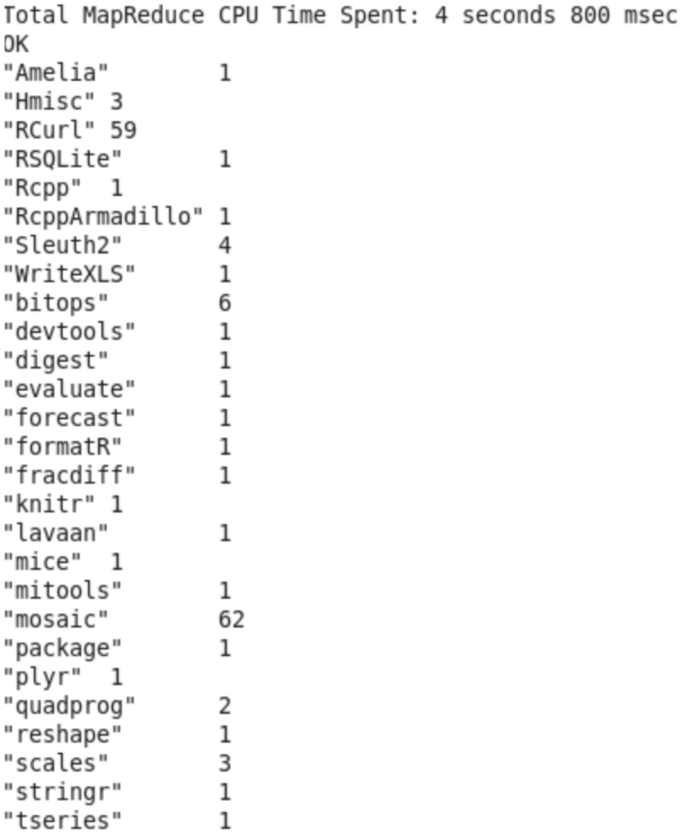


```
Time taken: 0.764 seconds, Fetched: 160 row(s)
hive> select * from oct1 LIMIT 10;
OK
"date"  "time"  NULL    "r_version"     "r_arch"        "r_os"  "package"       "version"       "country"       NULL
"2012-10-01"    "00:30:13"      35165   "2.15.1"        "i686"  "linux-gnu"     "quadprog"      "1.5-4" "AU"    1
"2012-10-01"    "00:30:15"      212967  "2.15.1"        "i686"  "linux-gnu"     "lavaan"        "0.5-9" "AU"    1
"2012-10-01"    "02:30:16"      167199  "2.15.1"        "x86_64"        "linux-gnu"     "formatR"       "0.6"   "US"    2
"2012-10-01"    "02:30:16"      21164   "2.15.1"        "x86_64"        "linux-gnu"     "stringr"       "0.6.1" "US"    2
"2012-10-01"    "02:30:13"      11046   "2.15.1"        "x86_64"        "linux-gnu"     "evaluate"      "0.4.2" "US"    2
"2012-10-01"    "02:30:13"      42294   "2.15.1"        "x86_64"        "linux-gnu"     "digest"        "0.5.2" "US"    2
"2012-10-01"    "02:30:16"      435407  "2.15.1"        "x86_64"        "linux-gnu"     "knitr" "0.8"   "US"    2
"2012-10-01"    "02:26:05"      326143  "2.15.1"        "i686"  "linux-gnu"     "mice"  "2.13"  "AU"    1
"2012-10-01"    "02:21:09"      119459  "2.15.1"        "i686"  "linux-gnu"     "mitools"       "2.2"   "AU"    1
Time taken: 0.133 seconds, Fetched: 10 row(s)
hive> █
```

Count the number of occurrences of different packages?

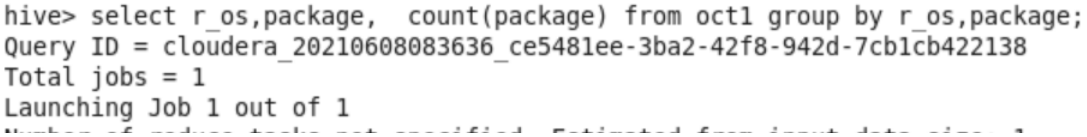


```
Time taken: 1.357 seconds, Fetched: 10 row(s)
hive> select package, count(package) from oct1 group by package;
Query ID = cloudera_20210608082626_c28c954c-7bb0-42f8-aade-612ea1919764
Total jobs = 1
```

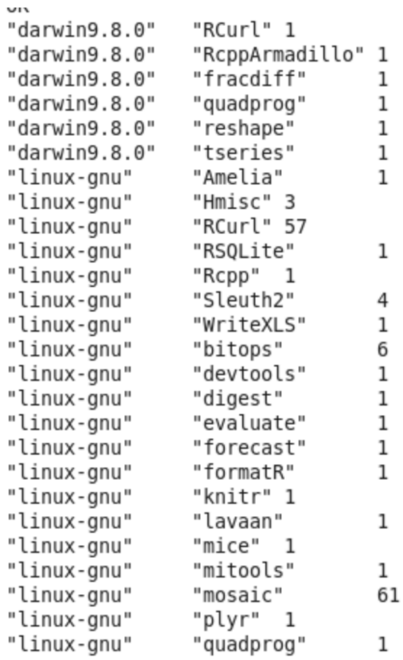


```
Total MapReduce CPU Time Spent: 4 seconds 800 msec
OK
"Amelia"        1
"Hmisc" 3
"RCurl" 59
"RSQLite"       1
"Rcpp"  1
"RcppArmadillo" 1
"Sleuth2"       4
"WriteXLS"      1
"bitops"        6
"devtools"      1
"digest"        1
"evaluate"      1
"forecast"      1
"formatR"       1
"fracdiff"      1
"knitr" 1
"lavaan"        1
"mice"  1
"mitools"       1
"mosaic"        62
"package"       1
"plyr"  1
"quadprog"      2
"reshape"       1
"scales"        3
"stringr"       1
"tseries"       1
```

Count the number of occurrences of different packages by operating system;



```
hive> select r_os,package,  count(package) from oct1 group by r_os,package;
Query ID = cloudera_20210608083636_ce5481ee-3ba2-42f8-942d-7cb1cb422138
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
```



```
OK
"darwin9.8.0"   "RCurl" 1
"darwin9.8.0"   "RcppArmadillo" 1
"darwin9.8.0"   "fracdiff"      1
"darwin9.8.0"   "quadprog"      1
"darwin9.8.0"   "reshape"       1
"darwin9.8.0"   "tseries"       1
"linux-gnu"     "Amelia"        1
"linux-gnu"     "Hmisc" 3
"linux-gnu"     "RCurl" 57
"linux-gnu"     "RSQLite"       1
"linux-gnu"     "Rcpp"  1
"linux-gnu"     "Sleuth2"       4
"linux-gnu"     "WriteXLS"      1
"linux-gnu"     "bitops"        6
"linux-gnu"     "devtools"      1
"linux-gnu"     "digest"        1
"linux-gnu"     "evaluate"      1
"linux-gnu"     "forecast"      1
"linux-gnu"     "formatR"       1
"linux-gnu"     "knitr" 1
"linux-gnu"     "lavaan"        1
"linux-gnu"     "mice"  1
"linux-gnu"     "mitools"       1
"linux-gnu"     "mosaic"        61
"linux-gnu"     "plyr"  1
"linux-gnu"     "quadprog"      1
```