

Big Data Management

Griffith College Dublin

Assignment 1

Rohin Mehra

Enrolment number: 3082862

MSc.Big Data Analysis & Management

1 April 2023

Abstract	3
Commands Used	4
Screenshots Of Hadoop Run	7
Mapper & Reducer Screenshots	12

Abstract

In this assignment, we are using Apache Hadoop to perform operations on a CSV file data which holds the following column names, which are as follows:

1. date: Download date
2. time: Download time (in UTC)
3. size: Package size (in bytes)
4. r_version: Version of R used to download package
5. r_arch: Processor architecture (i386 = 32 bit, x86_64 = 64 bit)
6. r_os: Operating System (darwin9.8.0 = mac, mingw32 = windows)
7. package: Name of the package downloaded
8. country: Two-letter ISO country code
9. ip_id: A unique daily id assigned to each IP address

By using this information we will perform the task of MapReduce.

We will also try to seek correct answers to the following questions during the process of MapReduce.

1. Show the number of downloads for package ggplot2.
2. List the highest number of downloads by country?
3. What were the top 10 most popular packages?
4. What is the most popular package in Ireland?
5. What OS is most popular among R programmers?

Commands Used

Q1.

```
$HADOOP_HOME/bin/hadoop jar $HADOOP_HOME/share/hadoop/tools/  
lib/hadoop-streaming-*jar -input /user/bdm/assignment/input -output /user/  
bdm/assignment/output -file /home/bdm/ assignment/mapper.py -file /  
home/bdm/assignment/reducer.py -mapper 'python3 mapper.py' -reducer  
'python3 reducer.py'
```

Output: Number of downloads for package ggplot2: 22360632

Q2.

```
$HADOOP_HOME/bin/hadoop jar $HADOOP_HOME/share/hadoop/tools/  
lib/hadoop-  
streaming-*jar -input /user/bdm/assignment/input -output /user/bdm/  
assignment/output2 -file /home/bdm/assignment/mapper.py -file /home/  
bdm/  
assignment/reducer2.py -mapper 'python3 mapper.py' -reducer 'python3  
reducer2.py'
```

Output: Highest number of downloads by a country: "NA" with 3225550
downloads

Q3.

```
$HADOOP_HOME/bin/hadoop jar $HADOOP_HOME/share/hadoop/tools/  
lib/hadoop-  
streaming*.jar -input /user/bdm/assignment/input -output /user/bdm/  
assignment/output3 -file /home/bdm/assignment/mapper.py -file /  
home/bdm/  
assignment/reducer3.py -mapper 'python3 mapper.py' -reducer 'python3  
reducer3.py'
```

Output: Top 10 most popular packages:

1. "NA": 3225550 downloads
2. ""mingw32"": 3194919 downloads
3. ""US"": 3061236 downloads
4. ""linux-gnu"": 778523 downloads
5. ""darwin17.0"": 648165 downloads
6. ""GB"": 569535 downloads
7. ""darwin20"": 328304 downloads
8. ""CN"": 282214 downloads
9. ""KR"": 254392 downloads
10. ""DE"": 236903 downloads

Q4.

```
$HADOOP_HOME/bin/hadoop jar $HADOOP_HOME/share/hadoop/tools/lib/hadoop-streaming*.jar -input /user/bdm/assignment/input -output /user/bdm/assignment/output4 -file /home/bdm/assignment/mapper.py -file /home/bdm/assignment/reducer4.py -mapper 'python3 mapper.py' -reducer 'python3 reducer4.py'
```

Output: Most popular package in Ireland: ""mingw32"" with 3194919 downloads

Q5.

```
$HADOOP_HOME/bin/hadoop jar $HADOOP_HOME/share/hadoop/tools/lib/hadoop-streaming*.jar -input /user/bdm/assignment/input -output /user/bdm/assignment/output5 -file /home/bdm/assignment/mapper.py -file /home/bdm/assignment/reducer5.py -mapper 'python3 mapper.py' -reducer 'python3 reducer5.py'
```

Output: Most popular OS among R programmers: ""mingw32"" with 3194919 downloads

Screenshots Of Hadoop Run

Question 1

```
ubuntu:/assignments$ SHADOOP_HOME/bin/hadoop jar SHADOOP_HOME/share/hadoop/tools/lib/hadoop-streaming-.jar -input /user/bdm/assignment/input -output /user/bdm/assignment/output -file /home/bdm/assignment/streamjob.py
14-01-02:40:38 578 WARN streaming.StreamJob: File execution is deprecated; please use generic option -files instead.
14-01-02:40:39 204 INFO client.DeAuthWithHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8082
14-01-02:40:39 342 INFO client.DeAuthWithHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8082
14-01-02:40:40 489 INFO mapred.FileInputFormat: Total input files to process : 1
14-01-02:40:40 555 INFO mapreduce.JobSubmitter: Number of splits:5
14-01-02:40:40 675 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1680310999416_0015
14-01-02:40:40 681 INFO mapreduce.Job: UserSpecifiedJobName is null
14-01-02:40:40 617 INFO resource.ResourceUtil: Unable to find 'resource-types.xml'.
14-01-02:40:40 684 INFO impl.YarnClientImpl: Submitted application application_1680310999416_0015
14-01-02:40:40 916 INFO mapreduce.Job: The url to track the job: http://ubuntu:8088/proxy/application_1680310999416_0015/
14-01-02:40:40 921 INFO mapreduce.Job: 2014-01-02 10:40:40 Job Id: job_1680310999416_0015
14-01-02:40:40 963 INFO mapreduce.Job: Job: job_1680310999416_0015 running in uber mode : false
14-01-02:40:40 957 INFO mapreduce.Job: map 0% reduce 0%
14-01-02:41:07 728 INFO mapreduce.Job: map 60% reduce 0%
14-01-02:41:13 985 INFO mapreduce.Job: map 73% reduce 0%
14-01-02:41:19 985 INFO mapreduce.Job: map 100% reduce 0%
14-01-02:41:20 372 INFO mapreduce.Job: map 100% reduce 0%
14-01-02:41:20 241 INFO mapreduce.Job: map 100% reduce 72%
14-01-02:41:30 284 INFO mapreduce.Job: map 100% reduce 97%
14-01-02:41:30 339 INFO mapreduce.Job: map 100% reduce 100%
14-01-02:41:30 539 INFO mapreduce.Job: Job job_1680310999416_0015 completed successfully
14-01-02:41:30 516 INFO mapreduce.Job: Counters
File System Counters
FILE: Number of bytes read=76151978
FILE: Number of bytes written=88819824
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=99919249
HDFS: Number of bytes written=51
HDFS: Number of read operations=20
HDFS: Number of large read operations=0
HDFS: Number of write operations=2
HDFS: Number of bytes read erasure-coded=0
Job Counters
Killed map tasks=0
Launched map tasks=1
Launched reduce tasks=1
Data-local map tasks=1
Total time spent by all maps in occupied slots (ms)=146278
Total time spent by all reduces in occupied slots (ms)=26739
Total time spent by all map tasks (ms)=146278
Total time spent by all reduce tasks (ms)=26739
Total vcore-milliseconds taken by all map tasks=14278
Total vcore-milliseconds taken by all reduce tasks=26739
Total megabyte-milliseconds taken by all map tasks=147988672
Total megabyte-milliseconds taken by all reduce tasks=27380736
Map-Reduce Framework
Map input records=7453844
Map output records=22368632
Map output bytes=258674277
Map output materialized bytes=303393771
Input split bytes=576
Combine input records=0
Combine output records=0
Reduce input groups=20835
Reduce shuffle bytes=303393771
Reduce input records=22368632
Map-Reduce Framework
Map input records=7453844
Map output records=22368632
Map output bytes=258674277
Map output materialized bytes=303393771
Input split bytes=576
Combine input records=0
Combine output records=0
Reduce input groups=20835
Reduce shuffle bytes=303393771
Reduce input records=22368632
Reduced Input records=22368632
Serialized Records=744356
Shuffled Maps =5
Failed Shuffles=0
Merged Map outputs=5
0.000000 ms=2922
CPU time spent (ms)=9178
Physical memory (bytes) snapshot=2386871296
Virtual memory (bytes) snapshot=14768511104
Total committed heap memory (bytes)=2386871296
Peak Map Physical memory (bytes)=482816900
Peak Map Virtual memory (bytes)=2473878784
Peak Reduce Physical memory (bytes)=494931968
Peak Reduce Virtual memory (bytes)=2478227456
Peak
Shuffle
BAD_ID=0
CONNECTION=0
IO_ERROR=0
WRONG_PARTITION=0
WRONG_MAP=0
WRONG_REDUCE=0
File Input Format Counters
Bytes Read=599518479
File Output Format Counters
Bytes Written=1
4-01-02:41:30 516 INFO streaming.StreamJob: Output directory: /user/bdm/assignment/output
ubuntu:/assignments$ hdfs dfs -cat /user/bdm/assignment/output/part-00000
of downloads for package ggplot2: 22368632
```

Task 1

Question 2

```
ubuntu:~/assignment$ $HADOOP_HOME/bin/hadoop jar $HADOOP_HOME/share/hadoop/tools/lib/hadoop-streaming-e.jar -input /user/bdm/assignment/input -output /user/bdm/assignment/output2 -file /home/bdm/mapper.py -file /home/bdm/assignment/reducer2.py -mapper 'python3 mapper.py' -reducer reducer2.py'
14-01 02:47:47,592 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead
14-01 02:47:48,364 INFO client.DefaultNoharmFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8082 []
14-01 02:47:48,558 INFO client.DefaultNoharmFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8082
14-01 02:47:49,787 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/bdm/.staging/job_1680310999416_0016
14-01 02:47:49,766 INFO mapreduce.JobInputFormat: Total input files to process : 1
14-01 02:47:49,787 INFO mapreduce.JobResourceUploader: Uploading tokens for job: job_1680310999416_0016
14-01 02:47:50,146 INFO mapreduce.JobResourceUploader: Submitting tokens for job: job_1680310999416_0016
14-01 02:47:50,298 INFO conf.Configuration: resource-types.xml not found
14-01 02:47:50,298 INFO resource.ResourcesUtil: Unable to find resource-types.xml.
14-01 02:47:50,298 INFO resource.ResourcesUtil: Using default resource-type configuration.job_1680310999416_0016
14-01 02:47:50,412 INFO mapreduce.Job: The url to track the job: http://ubuntu:8088/proxy/application_1680310999416_0016
14-01 02:47:50,584 INFO mapreduce.Job: Job job_1680310999416_0016 running in uber mode : false
14-01 02:47:50,584 INFO mapreduce.Job: map 5% reduce 0%
14-01 02:48:10,180 INFO mapreduce.Job: map 52% reduce 0%
14-01 02:48:10,278 INFO mapreduce.Job: map 5% reduce 0%
14-01 02:48:24,516 INFO mapreduce.Job: map 73% reduce 0%
14-01 02:48:24,516 INFO mapreduce.Job: map 5% reduce 0%
14-01 02:48:24,516 INFO mapreduce.Job: map 92% reduce 0%
14-01 02:48:36,814 INFO mapreduce.Job: map 100% reduce 0%
14-01 02:48:42,633 INFO mapreduce.Job: map 100% reduce 51%
14-01 02:48:48,297 INFO mapreduce.Job: map 100% reduce 68%
14-01 02:48:48,297 INFO mapreduce.Job: map 100% reduce 77%
14-01 02:49:06,780 INFO mapreduce.Job: map 100% reduce 65%
14-01 02:49:06,848 INFO mapreduce.Job: map 100% reduce 100%
14-01 02:49:06,858 INFO mapreduce.Job: Job job_1680310999416_0016 completed successfully
14-01 02:49:07,172 INFO mapreduce.Job: Counters: 55
File System Counters:
  FILE: Number of bytes read=75119978
  FILE: Number of bytes written=888198854
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
  HDFS: Number of bytes read=599519249
  HDFS: Number of bytes written=71
  HDFS: Number of read operations=20
  HDFS: Number of large read operations=0
  HDFS: Number of write operations=2
  HDFS: Number of bytes read erasure-coded=0
Job Counters
  Killed map tasks=1
  Launched map tasks=6
  Launched reduce tasks=1
  Data-local map tasks=6
  Total time spent by all map tasks in occupied slots (ms)=171835
  Total time spent by all map tasks in occupied slots (ms)=44647
  Total time spent by all map tasks (ms)=171835
  Total time spent by all reduce tasks (ms)=44647
  Total vcore-milliseconds taken by all map tasks=171835
  Total vcore-milliseconds taken by all reduce tasks=44647
  Total megabyte-milliseconds taken by all map tasks=175959849
  Total megabyte-milliseconds taken by all reduce tasks=45718528
Map-Reduce Framework
  Map input records=745044
  Map output records=2248632
  Map output bytes=2867477
  Map output materialized bytes=303393771
  Input split bytes=578
  Combine input records=0
Map-Reduce Framework
  FILE: Number of bytes read=575119978
  FILE: Number of bytes written=888198854
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
  HDFS: Number of bytes read=599519249
  HDFS: Number of bytes written=71
  HDFS: Number of read operations=20
  HDFS: Number of large read operations=0
  HDFS: Number of write operations=2
  HDFS: Number of bytes read erasure-coded=0
Job Counters
  Killed map tasks=1
  Launched map tasks=6
  Launched reduce tasks=1
  Data-local map tasks=6
  Total time spent by all maps in occupied slots (ms)=171835
  Total time spent by all reduces in occupied slots (ms)=44647
  Total time spent by all map tasks (ms)=171835
  Total time spent by all reduce tasks (ms)=44647
  Total vcore-milliseconds taken by all map tasks=171835
  Total vcore-milliseconds taken by all map tasks=44647
  Total megabyte-milliseconds taken by all map tasks=175959849
  Total megabyte-milliseconds taken by all reduce tasks=45718528
Map-Reduce Framework
  Map input records=745044
  Map output records=2248632
  Map output bytes=2867477
  Map output materialized bytes=303393771
  Input split bytes=578
  Combine input records=0
  Combine output records=8
  Reduce input groups=20035
  Reduce shuffle bytes=383393771
  Reduce output bytes=3368652
  Reduce output Records=1
  Spilled Records=6474436
  Shuffled Maps=5
  Failed Shuffles=0
  Merged Map outputs=5
  GC time elapsed (ms)=6071
  CPU time spent (ms)=115550
  Physical memory (bytes) snapshot=2567557128
  Virtual memory (bytes) snapshot=1795355536
  Total resident memory (bytes)=2462555632
  Peak Map Physical memory (bytes)=482521088
  Peak Map Virtual memory (bytes)=472869888
  Peak Reduce Physical memory (bytes)=467927408
  Peak Reduce Virtual memory (bytes)=2480345668
  Shuffle Errors
    BAD_ID=0
    CONNECTION=0
    IO_ERROR=0
    WRONG_LENGTH=0
    WRONG_MAP=0
    WRONG_REDUCE=0
  File Input Format Counters:
    BYTES读取=10951679
  File Output Format Counters:
    Bytes Written=71
14-01 02:49:07,173 INFO streaming.StreamJob: Output directory: /user/bdm/assignment/output
ubuntu:~/assignment$ hdfs dfs -cat /user/bdm/assignment/output/part-00000
it number of downloads by a country: "NA" with 3228556 downloads
```

Task 2

Question 3

```
ubuntu:~/assignments$ $HADOOP_HOME/bin/hadoop jar $HADOOP_HOME/share/hadoop/tools/lib/hadoop-streaming*.jar -input /user/bdm/assignment/input -output /user/bdm/assignment/output3 -file /home/bdm/mapper.py
File: /home/bdm/assignment/mapper.py, file does not exist. DeprecationWarning: Please use 'resource' instead.
aJobJar: [/home/bdm/assignment/mapper.py, /home/bdm/assignment/reducer3.py, /tmp/hadoop-unjar499467728579354224/] []
aJobID: job_168031099416_0017
aJobName: mapred.FileInputFormatTest
Total input files to process : 1
aJob: 02:54:14,278 INFO mapreduce.JobSubmitter: Number of splits:5
aJob: 02:54:14,342 INFO mapreduce.JobSubmitter: Executing with tokens: []
aJob: 02:54:14,342 INFO mapreduce.JobSubmitter: http://ubuntu:8088/proxy/application\_168031099416\_0017
aJob: 02:54:14,484 INFO resource.ResourceUtil: Unable to find 'resource-types.xml'.
aJob: 02:54:14,615 INFO impl.YarnClientImpl: Submitted application application_168031099416_0017
aJob: 02:54:14,651 INFO mapreduce.Job: The url to track the job: http://ubuntu:8088/proxy/application\_168031099416\_0017
aJob: 02:54:14,651 INFO mapreduce.Job: Running job: Job 168031099416_0017
aJob: 02:54:14,679 INFO mapreduce.Job: Job 168031099416_0017 running in uber mode : false
aJob: 02:54:14,770 INFO mapreduce.Job: Counters: 55
File System Counters
  FILE: Number of bytes read=76112970
  FILE: Number of bytes written=886190854
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
  HDFS: Number of bytes read=249
  HDFS: Number of bytes written=347
  HDFS: Number of read operations=20
  HDFS: Number of large read operations=2
  HDFS: Number of write operations=2
  HDFS: Number of bytes read erasure-coded=0
Job Counters
  Killed map tasks=1
  Launched map tasks=0
  Launched reduce tasks=1
  Data locality map tasks=0
  Total time spent by all mappers in occupied slots (ms)=168982
  Total time spent by all reducers in occupied slots (ms)=29786
  Total time spent by all map tasks (ms)=168982
  Total time spent by all reduce tasks (ms)=29786
  Total core-milliseconds taken by all map tasks=168982
  Total core-milliseconds taken by all reduce tasks=29786
  Total megabyte-milliseconds taken by all map tasks=17295648
  Total megabyte-milliseconds taken by all reduce tasks=36418944
Map-Reduce Framework
  Map input records=745344
  Map output records=2340632
  Map output bytes=258672477
  Map output materialized bytes=303393771
  Input split bytes=576
```

```
HDFS: Number of bytes read erasure-coded=0
Job Counters
  Killed map tasks=1
  Launched map tasks=0
  Launched reduce tasks=1
  Data locality map tasks=0
  Total time spent by all mappers in occupied slots (ms)=168982
  Total time spent by all reducers in occupied slots (ms)=29786
  Total time spent by all map tasks (ms)=168982
  Total time spent by all reduce tasks (ms)=29786
  Total core-milliseconds taken by all map tasks=168982
  Total core-milliseconds taken by all reduce tasks=29786
  Total megabyte-milliseconds taken by all map tasks=17295648
  Total megabyte-milliseconds taken by all reduce tasks=36418944
Map-Reduce Framework
  Map input records=745344
  Map output records=2340632
  Map output bytes=258672477
  Map output materialized bytes=303393771
  Input split bytes=576
  Combine input records=0
  Combine output records=0
  Reduce input groups=26635
  Reduce output bytes=2153775184
  Reduce output records=2340632
  Reduce output records=11
  Spilled Records=64744356
  Shuffled Maps = 0
  Failed Map files=0
  Merged Map outputs=5
  GC time elapsed (ms)=3212
  CPU time spent (ms)=108386
  Physical memory (bytes) snapshot=2309423872
  Virtual memory (bytes) snapshot=4756461784
  Total committed heap usage (bytes)=2153775184
  Peak Map Physical memory (bytes)=481894400
  Peak Map Virtual memory (bytes)=247141176
  Peak Reduce Physical memory (bytes)=3808800
  Peak Reduce Virtual memory (bytes)=24812205568
  Shuffle Errors
    BAD_ID=0
    CONNECTION=0
    IO_ERROR=0
    WRONG_LENGTH=0
    WRONG_MAP=0
    WRONG_REDUCE=0
  File Input Format Counters
    Bytes Read=599518679
  File Output Format Counters
    Bytes Written=349
4-01 02:55:17,771 INFO streaming.StreamJob: Output directory: /user/bdm/assignment/output
ubuntu:~/assignments$ hdfs dfs -cat /user/bdm/assignment/output3/part-00000
most popular packages:
*: 325568 downloads
linux32*: 3194919 downloads
*: 5314259 downloads
linux-gnu*: 778521 downloads
armwin17.0*: 648155 downloads
*: 50935 downloads
armwin20*: 392838 downloads
*: 2122321 downloads
R*: 254392 downloads
DE*: 236903 downloads
```

Task 3

Question 4

```

root@ubuntu-OptiPlex-5070:~/Assignment$ SHADOP HOME/bin/hadoop jar SHADOP_HOME/share/hadoop/tools/lib/hadoop-streaming.jar -input /user/bdm/assignment/input -output /user/bdm/assignment/output4 -file mapper.py -file reducer.py -mapper "python3 mapper.py" -reducer "python3 reducer.py"
4-01 :/assignment:WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
eJobJobs = ["/home/bdm/assignment/mapper.py", "/home/bdm/assignment/reducer.py", "/tmp/hadoop-unjarred/jar/jarfile.jar"]
4-01 :/assignment:INFO client.HDFSClient: Connecting to ResourceManager at /0.0.0.0:8088
4-01 :/assignment:INFO client.DefaultHDFSJobTracker: Connecting to ResourceManager at /0.0.0.0:8088
4-01 :/assignment:INFO client.HDFSClient: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/bdm/_staging/job_1688310999416_0028
4-01 :/assignment:INFO mapreduce.FileInputFormat: Total input files to process : 1
4-01 :/assignment:INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1688310999416_0028
4-01 :/assignment:INFO mapreduce.JobSubmitter: Executing with tokens: []
4-01 :/assignment:INFO conf.Configuration: types.xml not found
4-01 :/assignment:INFO conf.Configuration: types.xml not found
4-01 :/assignment:INFO mapreduce.Job: Unable to find resource types.xml.
4-01 :/assignment:INFO mapreduce.Job: The url to track the job: http://ubuntu:8088/proxy/application_1688310999416_0028
4-01 :/assignment:INFO mapreduce.Job: The url to track the job: https://ubuntu:8088/proxy/application_1688310999416_0028/
4-01 :/assignment:INFO mapreduce.Job: Running job: job_1688310999416_0028
4-01 :/assignment:INFO mapreduce.Job: map 32% reduce 0%
4-01 :/assignment:INFO mapreduce.Job: map 50% reduce 0%
4-01 :/assignment:INFO mapreduce.Job: map 57% reduce 0%
4-01 :/assignment:INFO mapreduce.Job: map 65% reduce 0%
4-01 :/assignment:INFO mapreduce.Job: map 73% reduce 0%
4-01 :/assignment:INFO mapreduce.Job: map 72% reduce 0%
4-01 :/assignment:INFO mapreduce.Job: map 73% reduce 0%
4-01 :/assignment:INFO mapreduce.Job: map 97% reduce 0%
4-01 :/assignment:INFO mapreduce.Job: map 100% reduce 0%
4-01 :/assignment:INFO mapreduce.Job: map 100% reduce 0%
4-01 :/assignment:INFO mapreduce.Job: map 100% reduce 6%
4-01 :/assignment:INFO mapreduce.Job: map 100% reduce 83%
4-01 :/assignment:INFO mapreduce.Job: map 100% reduce 100%
4-01 :/assignment:INFO mapreduce.Job: map 100% reduce 100%
4-01 :/assignment:INFO mapreduce.Job: Job: job_1688310999416_0028 completed successfully
4-01 :/assignment:INFO mapreduce.Job: Counters: 55
File System Counters
    File Number of bytes read=422251549
    File Number of bytes written=952027636
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=10249
    HDFS: Number of bytes written=69
    HDFS: Number of read operations=20
    HDFS: Number of large read operations=0
    HDFS: Number of write operations=2
    HDFS: Number of bytes read exactly-coded=0
    HDFS: Number of bytes read erasure-coded=0
Job Counters
    Killed map tasks=1
    Launched map tasks=5
    Launched reduce tasks=1
    Data-local map tasks=4
    Total time spent by all maps in occupied slots (ms)=190885
    Total time spent by all reduces in occupied slots (ms)=35170
    Total time spent by all map tasks=190885
    Total time spent by all reduce tasks=35170
    Total vcore-milliseconds taken by all map tasks=190835
    Total vcore-milliseconds taken by all reduce tasks=35170
    Total negate-milliseconds taken by all map tasks=19454248
    Total negate-milliseconds taken by all reduce tasks=35614080
Map-Reduce Framework
    Map input records=745354
    Map output records=22306632
    Map output bytes=28347764

```

```
FILE: Number of bytes read=422151459
FILE: Number of bytes written=9520277636
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=599519249
HDFS: Number of bytes written=69
HDFS: Number of read operations=20
HDFS: Number of large read operations=0
HDFS: Number of write operations=0
HDFS: Number of bytes read erasure-coded=0
Job Counters
    Killed map tasks=0
    Launched map tasks=4
    Launched reduce tasks=1
    Data-local map tasks=6
    Total time spent by all maps in occupied slots (ms)=198385
    Total time spent by all maps in used slots (ms)=35178
    Total time spent by all map tasks (ms)=198385
    Total time spent by all reduce tasks (ms)=35178
    Total vcore-milliseconds taken by all map tasks=198385
    Total vcore-milliseconds taken by all reduce tasks=35178
    Total map-milliseconds taken by all map tasks=194954246
    Total map-milliseconds taken by all reduce tasks=35014088
Map-Reduce Framework
    Map input records=745354
    Map output records=238199052
    Map output bytes=328199052
    Map output materialized bytes=328199058
    Input split bytes=578
    Combine input records=0
    Combiner input records=0
    Reduce input groups=23510
    Reduce shuffle bytes=238199858
    Reduce input records=2346632
    Reduce output records=1
    Spilled Records=4744356
    Shuffled Maps =5
    Failed Shuffles=0
    Merged Map outputs=5
    OS Total available=1024000000000
    CPU time spent (ms)=16778
    Physical memory (bytes) snapshot=2658144256
    Virtual memory (bytes) snapshot=14747336784
    Total committed heap usage (bytes)=2658144256
    Peak Map Physical memory (bytes)=2471149568
    Peak Map Virtual memory (bytes)=2471149568
    Peak Reduce Physical memory (bytes)=484331520
    Peak Reduce Virtual memory (bytes)=2479550644
Shuffle Errors
    BAD_ID=0
    CONNECTION=0
    IO_ERROR=0
    WRONG_LENGTH=0
    WRONG_MAP=0
    WRONG_REDUCE=0
File Input Format Counters
    Bytes Read=99518679
File Output Format Counters
    Bytes Written=0
    Bytes Written=0
4-01 03:44:21,727 INFO streaming.StreamJob: Output directory: /user/bdm/assignment/output4
$ hdfs dfs -cat /user/bdm/assignment/output4/part-00000
No file found in /user/bdm/assignment/output4/part-00000
```

Task 4

Question 5

```

ubuntu:/assignment$ $HADOOP_HOME/bin/hadoop jar $HADOOP_HOME/share/hadoop/tools/lib/hadoop-streaming*.jar -input /user/bdm/assignment/input -output /user/bdm/assignment/output5 -file /home/bdm
mapper.py -file /home/bdm/assignment/reducer5.py -mapper 'python3 mapper.py' -reducer 'python3 reducer5.py'
4-01 03:35:30,353 WARN streaming.StreamJob: /home/bdm/assignment/mapper.py, /home/bdm/assignment/reducer5.py, /tmp/hadoop-unjs5709719445131851/] [] /tmp/streamjob8726752793092545136.jar tmpDir=null
4-01 03:34:31,095 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8082
4-01 03:34:31,354 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8082
4-01 03:34:31,572 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/bdm..staging/job_1680310999416_0026
4-01 03:34:31,580 INFO mapreduce.JobSubmissionEventHandler: Total number of processes : 1
4-01 03:34:31,658 INFO mapreduce.JobSubmissionEventHandler: Number of splits: 1
4-01 03:34:31,669 INFO mapreduce.JobSubmissionEventHandler: Submitting tokens for job: job_1680310999416_0026
4-01 03:34:32,222 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1680310999416_0026
4-01 03:34:32,322 INFO mapreduce.JobSubmitter: Executing with tokens: []
4-01 03:34:32,368 INFO conf.Configuration: resource-types.xml not found
4-01 03:34:32,370 INFO conf.Configuration: resource-types.xml not found
4-01 03:34:32,371 INFO impl.YarnClientImpl: Submitted application application_1680310999416_0026
4-01 03:34:32,583 INFO mapreduce.Job: The url to track the job: http://ubuntu:8088/proxy/application_1680310999416_0026/
4-01 03:34:32,585 INFO mapreduce.Job: Running job: job_1680310999416_0026
4-01 03:34:32,766 INFO mapreduce.Job: Job 3 completed successfully
4-01 03:34:32,768 INFO mapreduce.Job: map 0% reduce 0%
4-01 03:34:32,769 INFO mapreduce.Job: map 28% reduce 0%
4-01 03:35:01,282 INFO mapreduce.Job: map 50% reduce 0%
4-01 03:35:01,436 INFO mapreduce.Job: map 57% reduce 0%
4-01 03:35:01,438 INFO mapreduce.Job: map 64% reduce 0%
4-01 03:35:01,461 INFO mapreduce.Job: map 71% reduce 0%
4-01 03:35:12,656 INFO mapreduce.Job: map 84% reduce 0%
4-01 03:35:13,678 INFO mapreduce.Job: map 88% reduce 0%
4-01 03:35:13,679 INFO mapreduce.Job: map 100% reduce 0%
4-01 03:35:13,680 INFO mapreduce.Job: map 100% reduce 74K
4-01 03:35:20,863 INFO mapreduce.Job: map 100% reduce 67K
4-01 03:35:24,911 INFO mapreduce.Job: map 100% reduce 100%
4-01 03:35:34,934 INFO mapreduce.Job: Job job_1680310999416_0026 completed successfully
4-01 03:35:34,936 INFO mapreduce.Job: Counters: 54
File System Counters
FILE: Number of bytes read=422151459
FILE: Number of bytes written=952027630
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=999519249
HDFS: Number of bytes written=73
HDFS: Number of read operations=0
HDFS: Number of large read operations=0
HDFS: Number of write operations=2
HDFS: Number of bytes read erasure-coded=0
Job Counters
launched map tasks=5
launched reduce tasks=1
Data-local map tasks=5
Total time spent by all maps in occupied slots (ms)=152851
Total time spent by all reduce tasks in occupied slots (ms)=28556
Total time spent by all map tasks (ms)=152851
Total time spent by all reduce tasks (ms)=28556
Total vcore-milliseconds taken by all map tasks=152851
Total vcore-milliseconds taken by all reduce tasks=28556
Total megabyte-milliseconds taken by all map tasks=156519424
Total megabyte-milliseconds taken by all reduce tasks=27241344
Map-Reduce Framework
Map input records=745344
Map output records=2236632
Map output bytes=477764
Map output materialized bytes=328199858
Input split bytes=570
Combine input records=0
Combine output records=0
Combine output records=0

File System Counters
FILE: Number of bytes read=422151459
FILE: Number of bytes written=952027630
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=999519249
HDFS: Number of bytes written=73
HDFS: Number of read operations=0
HDFS: Number of large read operations=0
HDFS: Number of write operations=2
HDFS: Number of bytes read erasure-coded=0
Job Counters
launched map tasks=5
launched reduce tasks=1
Data-local map tasks=5
Total time spent by all maps in occupied slots (ms)=152851
Total time spent by all reduce tasks in occupied slots (ms)=28556
Total time spent by all map tasks (ms)=152851
Total time spent by all reduce tasks (ms)=28556
Total vcore-milliseconds taken by all map tasks=152851
Total vcore-milliseconds taken by all reduce tasks=28556
Total megabyte-milliseconds taken by all map tasks=156519424
Total megabyte-milliseconds taken by all reduce tasks=27241344
Map-Reduce Framework
Map input records=745344
Map output records=2236632
Map output bytes=477764
Map output materialized bytes=328199858
Input split bytes=570
Combine input records=0
Combine output records=0
Combine output records=0
Reduce input records=23618
Reduce shuffle bytes=328199858
Reduce input records=2236632
Reduce output records=31
Shuffled Maps=5
Failed Shuffles=0
Merged Map outputs=5
Data locality miss=495
CPU time spent (ms)=197228
Physical memory (bytes) snapshot=2621042688
Virtual memory (bytes) snapshot=14748766208
Total committed heap usage (bytes)=2527068160
Peak Physical memory (bytes)=2701910144
Peak Map Virtual memory (bytes)=2701910144
Peak Reduce Physical memory (bytes)=664948736
Peak Reduce Virtual memory (bytes)=2478886912
Shuffle Errors
 0x0: 0+0
 0x1: 0+0
 0x2: CONNECTION=0
 0x3: IO_ERROR=0
 0x4: WRONG_LENGTH=0
 0x5: WRONG_TYPE=0
 0x6: WRONG_REDUCE=0
File Input Format Counters
  Bytes Read=599518079
File Output Format Counters
  Bytes Written=599518079
4-01 03:35:35,129 INFO streaming.StreamJob: Output directory: /user/bdm/assignment/output5
ubuntu:/assignment$ hdfs dfs -cat /user/bdm/assignment/output5/part-00000
popular OS among R programmers: "mingw32" with 3194919 downloads

```

Task 5

Mapper & Reducer Screenshots

Mapper is common in structure for all tasks it's just that reducer changes for every operation desired.

The screenshot shows a Jupyter Notebook interface with several tabs at the top: BDM, mapper.py, reducer.py, reducer2.py, reducer3.py, reducer4.py, and reducer5.py. The main area displays the contents of the mapper.py file. The code is a Python script that reads from standard input, splits each line into fields, and then performs specific operations based on the number of fields (10). It handles questions about package downloads, highest downloads by country, top 10 packages, most popular package in Ireland, and the most popular OS among R programmers. The code uses print statements to output results.

```
#!/usr/bin/env python3
import sys

# input comes from STDIN (standard input)
for line in sys.stdin:
    # remove leading and trailing whitespace
    line = line.strip()

    # split the line into fields
    fields = line.split(',')

    # check that the line has the expected number of fields
    if len(fields) == 10:
        # extract the fields
        package = fields[6]
        country = fields[7]
        os = fields[5]

        # Question 1: number of downloads for package ggplot2
        if package == 'ggplot2':
            print('ggplot2\t1')

        # Question 2: highest number of downloads by country
        print("%s\t1" % country)

        # Question 3: top 10 most popular packages
        print("%s\t1" % package)

        # Question 4: most popular package in Ireland
        if country == 'IE':
            print("%s\t1" % package)

        # Question 5: most popular OS among R programmers
        print("%s\t1" % os)
```

Mapper Code

For Question 1 reducer is as follows:

The screenshot shows a code editor interface with a dark theme. On the left, there's a sidebar with 'Project' and 'Bookmarks' sections. The main area displays a Python file named 'reducer.py'. The code is as follows:

```
1 #!/usr/bin/env python3
2
3 import sys
4
5 # Initialize the count for ggplot2 downloads
6 ggplot2_downloads = 0
7
8 # Input comes from STDIN (standard input)
9 for line in sys.stdin:
10     # Remove leading and trailing whitespace
11     line = line.strip()
12
13     # Split the line into package and count
14     package, count = line.split('\t')
15
16     # Convert count (currently a string) to int
17     count = int(count)
18
19     # Increment the count for ggplot2 downloads
20     ggplot2_downloads += count
21
22 # Print the results
23 print(f'Number of downloads for package ggplot2: {ggplot2_downloads}'
```

At the bottom, there are tabs for 'Version Control', 'Python Packages', 'TODO', 'Python Console', 'Problems', 'Terminal', 'Endpoints', and 'Services'. On the right side, there are several icons for GitHub Copilot, Database, SciView, and Notifications.

Reducer for Question 1

For Question 2 reducer is as follows:

The screenshot shows a Jupyter Notebook interface with multiple tabs at the top: 'Current File' (highlighted), 'mapper.py', 'reducer.py', 'reducer2.py' (current tab), 'reducer3.py', 'reducer4.py', and 'reducer5.py'. The code in the 'reducer2.py' tab is as follows:

```
1 #!/usr/bin/env python3
2
3 import sys
4
5 # Initialize the country count dictionary
6 country_count = {}
7
8 # Input comes from STDIN (standard input)
9 for line in sys.stdin:
10     # Remove leading and trailing whitespace
11     line = line.strip()
12
13     # Split the line into country and count
14     country, count = line.split('\t')
15
16     # Convert count (currently a string) to int
17     count = int(count)
18
19     # Increment the count for the country
20     if country in country_count:
21         country_count[country] += count
22     else:
23         country_count[country] = count
24
25 # Find the highest number of downloads by country
26 highest_country = max(country_count, key=country_count.get)
27
28 # Print the results
29 print(f'Highest number of downloads by a country: {highest_country} with {country_count[highest_country]} downloads')
```

At the bottom of the interface, there are several tabs: Version Control, Python Packages, TODO, Python Console, Problems, Terminal, Endpoints, and Services. The status bar indicates the file is 30 lines long and is located at /Users/rohimmehra/opt/anaconda3.

Reducer for Question 2

For Question 3 reducer is as follows:

```
BDM > reducer3.py
mapper.py × reducer.py × reducer2.py × reducer3.py × reducer4.py × reducer5.py ×
1  #!/usr/bin/env python3
2
3  import ...
4
5
6  # Initialize the package count dictionary
7  package_count = {}
8
9  # Input comes from STDIN (standard input)
10 for line in sys.stdin:
11     # Remove leading and trailing whitespace
12     line = line.strip()
13
14     # Split the line into package and count
15     package, count = line.split('\t')
16
17     # Convert count (currently a string) to int
18     count = int(count)
19
20     # Increment the count for the package
21     if package in package_count:
22         package_count[package] += count
23     else:
24         package_count[package] = count
25
26     # Find the top 10 most popular packages
27     top_10_packages = Counter(package_count).most_common(10)
28
29     # Print the results
30     print('Top 10 most popular packages:')
31     for i, (package, count) in enumerate(top_10_packages, start=1):
32         print(f'{i}. {package}: {count} downloads')
33
```

Reducer for Question 3

For Question 4 reducer is as follows:

```
1 #!/usr/bin/env python3
2
3 import sys
4
5 # Initialize the package count dictionary and the count for Ireland
6 package_count = {}
7 ireland_count = 0
8
9 # Input comes from STDIN (standard input)
10 for line in sys.stdin:
11     # Remove leading and trailing whitespace
12     line = line.strip()
13
14     # Split the line into package and count
15     package, count = line.split('\t')
16
17     # Convert count (currently a string) to int
18     count = int(count)
19
20     # Increment the count for the package
21     if package in package_count:
22         package_count[package] += count
23     else:
24         package_count[package] = count
25
26     # Check if the country is Ireland and increment the count
27     if line.startswith("IE"):
28         ireland_count += count
29
30     # Find the most popular package in Ireland
31     ireland_packages = {k: v for k, v in package_count.items() if k in package_count and k is not None}
32     most_popular_package = max(ireland_packages, key=ireland_packages.get)
33
34     # Print the results
35     print(f'Most popular package in Ireland: {most_popular_package} with {ireland_packages[most_popular_package]} downloads')
```

Reducer for Question 4

For Question 5 reducer is as follows:

The screenshot shows a Jupyter Notebook interface with multiple tabs at the top: mapper.py, reducer.py, reducer2.py, reducer3.py, reducer4.py, and reducer5.py. The reducer5.py tab is active, displaying the following Python code:

```
#!/usr/bin/env python3
import sys

# Initialize the OS count dictionary
os_count = {}

# Input comes from STDIN (standard input)
for line in sys.stdin:
    # Remove leading and trailing whitespace
    line = line.strip()

    # Split the line into OS and count
    os, count = line.split('\t')

    # Convert count (currently a string) to int
    count = int(count)

    # Increment the count for the OS
    if os in os_count:
        os_count[os] += count
    else:
        os_count[os] = count

# Find the most popular OS among R programmers
most_popular_os = max(os_count, key=os_count.get)

# Print the results
print(f'Most popular OS among R programmers: {most_popular_os} with {os_count[most_popular_os]} downloads')
```

The notebook interface includes a toolbar with icons for file operations, a search bar, and a status bar at the bottom showing "1:1 LF UTF-8 4 spaces /Users/rohinmehra/opt/anaconda3/lib/python3.7".

Reducer for Question 5

