CSE488 (Section 1)
[Spring 2023]

Lab Assignment Submission Report

<u>All Online MapReduce LAB</u>

Submitted by:
Sazzad Hossen
ID: 2019-1-60-063

LAB 1

1.

```
# 1. Find the Air craft, country flying mission and target country for those missions in which time on target was less than 1000.

query= """

SELECT TargetCountry, COUNT(TargetCountry)
FROM Bombing_Operations
WHERE TimeOnTarget < 1000
GROUP BY TargetCountry

"""
ass1=spark.sql(query)
ass1.show()
```

```
+--------------+--------------------+
| TargetCountry|count(TargetCountry)|
+--------------+--------------------+
|       CAMBODIA|              154639|
|WESTPAC WATERS|                  63|
|  SOUTH VIETNAM|             1578639|
|        UNKNOWN|                  48|
|    PHILLIPINES|                   1|
|       THAILAND|               53765|
|  NORTH VIETNAM|              432509|
|           LAOS|              703833|
+--------------+--------------------+
```

2.

```
[39] #2. Find the target country and the number of times they have been attacked. Show the result in descending order.


    query= """
    SELECT MissionDate,TargetCountry
    FROM Bombing_Operations B
    ORDER BY B.TimeOnTarget DESC,
             B.MissionDate DESC
    """
    ass2=spark.sql(query)
    ass2.show()
```

```
+----------+-------------+
|MissionDate|TargetCountry|
+----------+-------------+
| 1965-11-25|NORTH VIETNAM|
| 1966-02-18|SOUTH VIETNAM|
| 1966-02-18|SOUTH VIETNAM|
| 1966-02-05|SOUTH VIETNAM|
| 1966-02-05|SOUTH VIETNAM|
| 1966-01-31|SOUTH VIETNAM|
| 1965-10-05|SOUTH VIETNAM|
| 1965-10-05|SOUTH VIETNAM|
| 1965-10-05|SOUTH VIETNAM|
| 1965-10-05|SOUTH VIETNAM|
| 1965-10-05|SOUTH VIETNAM|
```

3.

```
[55]  # 3.Find the target country which had been attacked most.
      query= """
      SELECT TargetCountry,
             COUNT(*) TimeOnTarget
      FROM Bombing_Operations
      GROUP BY TargetCountry
      HAVING COUNT(TargetCountry) > 5
      """


      ass3=spark.sql(query)
      ass3.show()
```

```
+--------------+------------+
| TargetCountry|TimeOnTarget|
+--------------+------------+
|      CAMBODIA|      218788|
|WESTPAC WATERS|          74|
| SOUTH VIETNAM|     2172403|
|       UNKNOWN|          59|
|      THAILAND|       60486|
| NORTH VIETNAM|      694269|
|          LAOS|     1254695|
+--------------+------------+
```

4.

```
[57]  # 4. Find the country flying mission which attacked most.
      query= """
      SELECT ContryFlyingMission,
             COUNT(*) TimeOnTarget
      FROM Bombing_Operations
      GROUP BY ContryFlyingMission
      HAVING COUNT(TimeOnTarget) > 1000
      """


      ass4=spark.sql(query)
      ass4.show()
```

```
+-------------------+------------+
| ContryFlyingMission|TimeOnTarget|
+-------------------+------------+
|    VIETNAM (SOUTH)|      622013|
|      KOREA (SOUTH)|       24469|
|UNITED STATES OF ...|     3708997|
|          AUSTRALIA|       12519|
|               LAOS|       32777|
+-------------------+------------+
```

5.

```
[58]  # 5. Find the take off location and how many times they had used.
      query= """
      SELECT TakeoffLocation,
              COUNT(TimeOnTarget) TimeOnTarget
      FROM Bombing_Operations
      GROUP BY TakeoffLocation

      """


      ass5=spark.sql(query)
      ass5.show()
```

```
+-------------------+------------+
|    TakeoffLocation|TimeOnTarget|
+-------------------+------------+
|       LOUANG PHRAB|        5164|
|       TAN SON NHUT|      311871|
|          KHE SANH|          27|
|         NHA TRANG|       56818|
|        WESTPAC 59|         481|
|        KADENA AFB|        5413|
|       TONKIN GULF|      194182|
|         VINH BINH|          10|
```

6.

```
[87]  # 6. Find the number of fighter jets from Aircraft_Glossary.
      Aircraft_Glossary.createOrReplaceTempView("Aircraft_Glossary")
      query= """ SELECT COUNT(*) AirCraft FROM Aircraft_Glossary WHERE AirCraftType =   'Fighter Jet'"""

      ass6= spark.sql(query)
      ass6.show()
```

```
+--------+
|AirCraft|
+--------+
|      10|
+--------+
```

7.

```
[98] # 7. Find the number of different types air craft for each air craft type.

    query= """ SELECT DISTINCT Aircraft FROM Aircraft_Glossary """

    ass7= spark.sql(query)
    ass7.show()
```

```
+--------+
|Aircraft|
+--------+
|     E-2|
|  LC-130|
|     F-5|
|   OV-10|
|     E-3|
|   HC-47|
|    HH-3|
|    TA-4|
|  EC-121|
|  AC-123|
|    C-76|
|    OV-1|
|    B-57|
|    U-17|
|   EB-66|
```
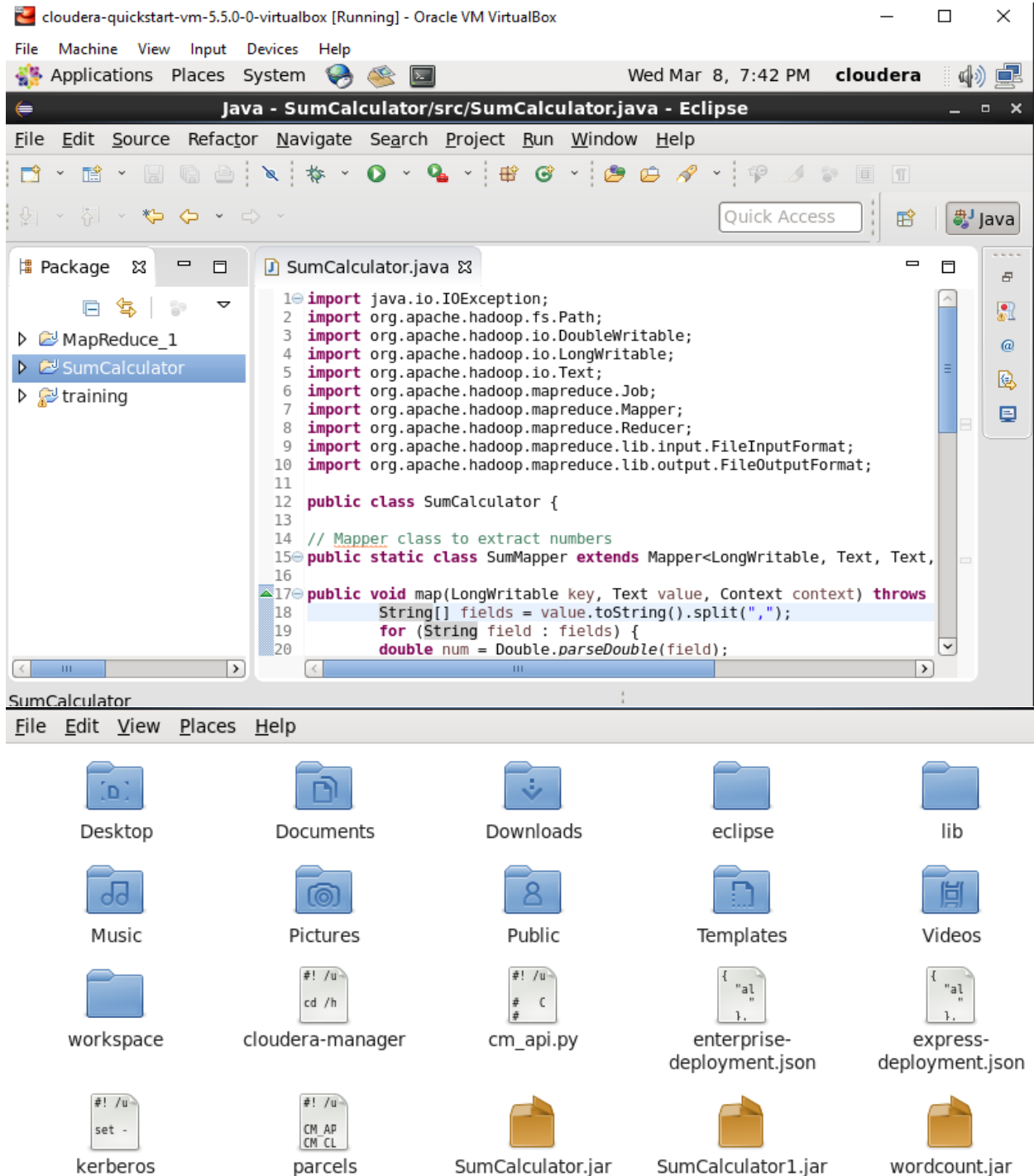
8.

```
p5=Bombing_Operations.join(Aircraft_Glossary,Bombing_Operations.AirCraft == Aircraft_Glossary.AirCraft,"inner") \.show(truncate=False)
```

AirCraft	ContryFlyingMission	MissionDate	OperationSupported	PeriodOfDay	TakeoffLocation	TargetCountry	TimeOnTarget	WeaponType	WeaponsLoadedWeight
EC-47	UNITED STATES OF AMERICA	1971-06-05	null	D	TAN SON NHUT	CAMBODIA	1005.0	null	0
EC-47	UNITED STATES OF AMERICA	1972-12-26	null	D	NAKHON PHANOM	SOUTH VIETNAM	530.0	null	0
RF-4	UNITED STATES OF AMERICA	1973-07-28	null	D	UDORN AB	LAOS	730.0	null	0
A-1	UNITED STATES OF AMERICA	1970-02-02	null	N	NAKHON PHANOM	LAOS	1415.0	BLU27 FIRE BOMB (750)	17400
A-37	VIETNAM (SOUTH)	1970-10-08	null	D	DANANG	SOUTH VIETNAM	1240.0	null	0
F-4	UNITED STATES OF AMERICA	1970-11-25	null	D	UBON AB	LAOS	650.0	MK 82 GP BOMB (500) LD	31860
A-4	UNITED STATES OF AMERICA	1972-03-08	null	D	TONKIN GULF	LAOS	1005.0	null	0
F-4	UNITED STATES OF AMERICA	1971-12-27	null	null	UDORN AB	LAOS	0.0	null	0
A-7	UNITED STATES OF AMERICA	1972-05-24	null	null	TONKIN GULF	NORTH VIETNAM	0.0	null	0
EC-47	UNITED STATES OF AMERICA	1972-09-12	null	D	TAN SON NHUT	SOUTH VIETNAM	710.0	null	0
CH-53	UNITED STATES OF AMERICA	1974-06-13	null	N	NAKHON PHANOM	THAILAND	1800.0	null	0
CH-53	UNITED STATES OF AMERICA	1974-12-19	null	D	NAKHON PHANOM	THAILAND	800.0	null	0
O-1	VIETNAM (SOUTH)	1973-10-24	null	D	NHA TRANG	SOUTH VIETNAM	800.0	null	0
UH-1	VIETNAM (SOUTH)	1974-03-19	null	D	PHU CAT	SOUTH VIETNAM	800.0	CARGO (TONS)	0

✓ 0s completed at 11:34 PM

LAB 3

Sum Calculator:

```
[cloudera@quickstart ~]$ hadoop jar SumCalculator1.jar SumCalculator /user/cloudera/wordcountfiles
23/03/08 19:37:40 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
23/03/08 19:37:43 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not perfo
oolRunner to remedy this.
23/03/08 19:37:44 INFO input.FileInputFormat: Total input paths to process : 1
23/03/08 19:37:44 INFO mapreduce.JobSubmitter: number of splits:1
23/03/08 19:37:44 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1678328889585_0001
23/03/08 19:37:45 INFO impl.YarnClientImpl: Submitted application application_1678328889585_0001
23/03/08 19:37:45 INFO mapreduce.Job: The url to track the job: http://quickstart.cloudera:8088/pr
23/03/08 19:37:45 INFO mapreduce.Job: Running job: job_1678328889585_0001
23/03/08 19:38:15 INFO mapreduce.Job: Job job_1678328889585_0001 running in uber mode : false
23/03/08 19:38:15 INFO mapreduce.Job:  map 0% reduce 0%
23/03/08 19:38:26 INFO mapreduce.Job:  map 100% reduce 0%
23/03/08 19:38:35 INFO mapreduce.Job:  map 100% reduce 100%
23/03/08 19:38:35 INFO mapreduce.Job: Job job_1678328889585_0001 completed successfully
23/03/08 19:38:35 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=90
                FILE: Number of bytes written=222989
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=146
                HDFS: Number of bytes written=15
                HDFS: Number of read operations=6
                HDFS: Number of large read operations=0
```

Java - SumCalculator/src/SumCalculator.java - Eclipse   ns=2

```
                Combine input records=0
                Combine output records=0
                Reduce input groups=1
                Reduce shuffle bytes=90
                Reduce input records=6
                Reduce output records=1
                Spilled Records=12
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=214
                CPU time spent (ms)=1360
                Physical memory (bytes) snapshot=330555392
                Virtual memory (bytes) snapshot=3008303104
                Total committed heap usage (bytes)=226365440
        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=11
        File Output Format Counters
                Bytes Written=15
```

Hue - File Browser - Mozilla Firefox

🦊 Hue - ...   🌐 Java - ...   🖥 Comp...   📁 /   📁 [home]   🏠 cloudera   🖥 cloude...   📁   🖥

about:s...estore   Hue - Fi... ×   East West ...   CSE488 [S...   CSE488_L...   Google Ac...

quickstart.cloudera:8888/filebrowser/view=/user/   Search

Cloudera Hue Hadoop ∨ HBase ∨ Impala ∨ Spark ∨ Solr Oozie Cloudera Manager

Query Editors ∨   Data Browsers ∨   Workflows ∨   Search   Security ∨

File Browser

Search for file name   ⚙ Actions ∨   ✖ Move to trash ∨

	Name	Size	User	Group	Permissions	Date
📁	⬆		cloudera	cloudera	drwxr-xr-x	March 08, 2023 07:38 PM
📁	.		cloudera	cloudera	drwxr-xr-x	March 08, 2023 07:38 PM
📄	_SUCCESS	0 bytes	cloudera	cloudera	-rw-r--r--	March 08, 2023 07:38 PM
📄	part-r-00000	15 bytes	cloudera	cloudera	-rw-r--r--	March 08, 2023 07:38 PM

Cloudera Hue Hadoop ∨ HBase ∨ Impala ∨ Spark ∨ Solr Oozie Cloudera Manager

Query Editors ∨   Data Browsers ∨   Workflows ∨   Search   Security ∨

File Browser
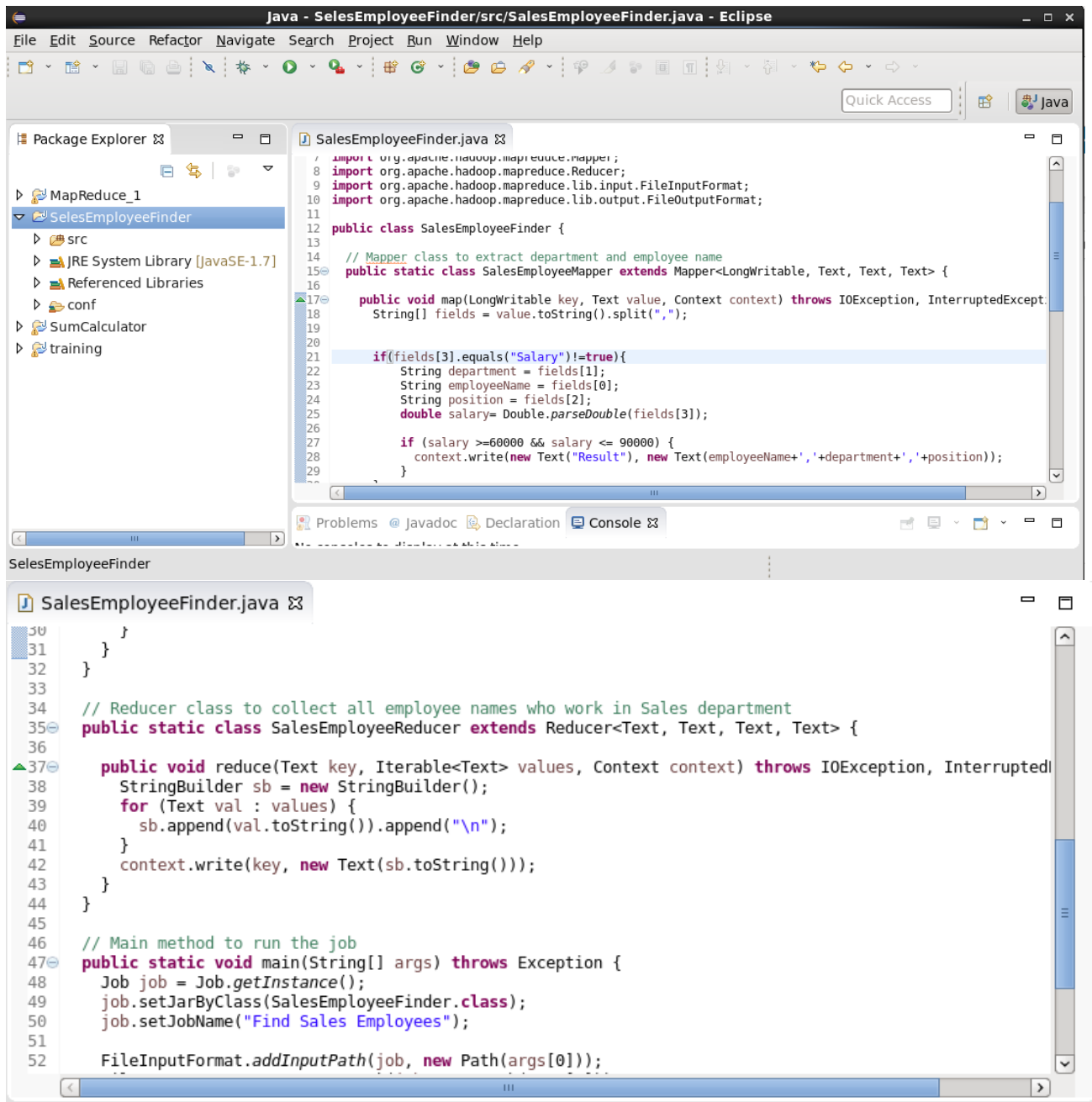
ACTIONS

▥ View as binary

✏ Edit file

⬇ Download

⌂ Home   Page 1 of 1 |◀ ◀◀ ▶▶ ▶|

/ user / cloudera / output / **part-r-00000**

```
Total sum     23.0
```

LAB 4

INPUT:


```java
import org.apache.hadoop.mapreduce.Mapper;
import org.apache.hadoop.mapreduce.Reducer;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;

public class SalesEmployeeFinder {

    // Mapper class to extract department and employee name
    public static class SalesEmployeeMapper extends Mapper<LongWritable, Text, Text, Text> {

        public void map(LongWritable key, Text value, Context context) throws IOException, InterruptedExcept
            String[] fields = value.toString().split(",");

            if(fields[3].equals("Salary")!=true){
                String department = fields[1];
                String employeeName = fields[0];
                String position = fields[2];
                double salary= Double.parseDouble(fields[3]);

                if (salary >=60000 && salary <= 90000) {
                    context.write(new Text("Result"), new Text(employeeName+','+department+','+position));
                }
```

```java
        }
    }
}

    // Reducer class to collect all employee names who work in Sales department
    public static class SalesEmployeeReducer extends Reducer<Text, Text, Text, Text> {

        public void reduce(Text key, Iterable<Text> values, Context context) throws IOException, Interrupted
            StringBuilder sb = new StringBuilder();
            for (Text val : values) {
                sb.append(val.toString()).append("\n");
            }
            context.write(key, new Text(sb.toString()));
        }
    }

    // Main method to run the job
    public static void main(String[] args) throws Exception {
        Job job = Job.getInstance();
        job.setJarByClass(SalesEmployeeFinder.class);
        job.setJobName("Find Sales Employees");

        FileInputFormat.addInputPath(job, new Path(args[0]));
```

File Edit View Places Help

Desktop   Documents   Downloads   eclipse   lib   Music   Pictures   Public

Templates   Videos   workspace   cloudera-manager   cm_api.py   enterprise-deployment.json   express-deployment.json   kerberos

parcels   SelesEmployeeFinder.jar   SelesEmployeeFinder1.jar   sum.jar   wordcountsp2.jar

---

Cloudera   Hue   Hadoop ˅   HBase ˅   Impala ˅   Spark ˅   Solr   Oozie   Cloudera Manager   Getting Started

Hue   Query Editors ˅   Data Browsers ˅   Workflows ˅   Search   Security ˅   File Browser

File Browser

ACTIONS

View as binary

Edit file

Download

View file location

Refresh

INFO

Last modified
March 15, 2023 7:51
p.m.

Home   / user / cloudera / seles / **sample_data.csv**

```
Name,Department,Position,Salary,Start_Date
John Smith,Engineering,Software Developer,75000,1/15/2019
Emily Johnson,Marketing,Marketing Coordinator,55000,3/12/2018
Michael Lee,Finance,Financial Analyst,65000,7/1/2020
Sarah Chen,Human Resources,HR Manager,85000,9/5/2017
David Kim,Operations,Operations Manager,95000,4/2/2016
Jasmine Wong,Sales,Sales Representative,60000,2/21/2021
Matthew Johnson,Engineering,Engineering Manager,100000,11/10/2015
Ava Martinez,Marketing,Marketing Manager,90000,6/23/2014
Jessica Park,Finance,Financial Controller,110000,2/15/2013
Richard Lee,Operations,Operations Coordinator,55000,1/1/2020
```

Output:

Cloudera   Hue   Hadoop ˅   HBase ˅   Impala ˅   Spark ˅   Solr   Oozie   Cloudera Manager   Getting Started

Hue   Query Editors ˅   Data Browsers ˅   Workflows ˅   Search   Security ˅   File Browser

File Browser

ACTIONS

View as binary

Edit file

Download

View file location

Refresh
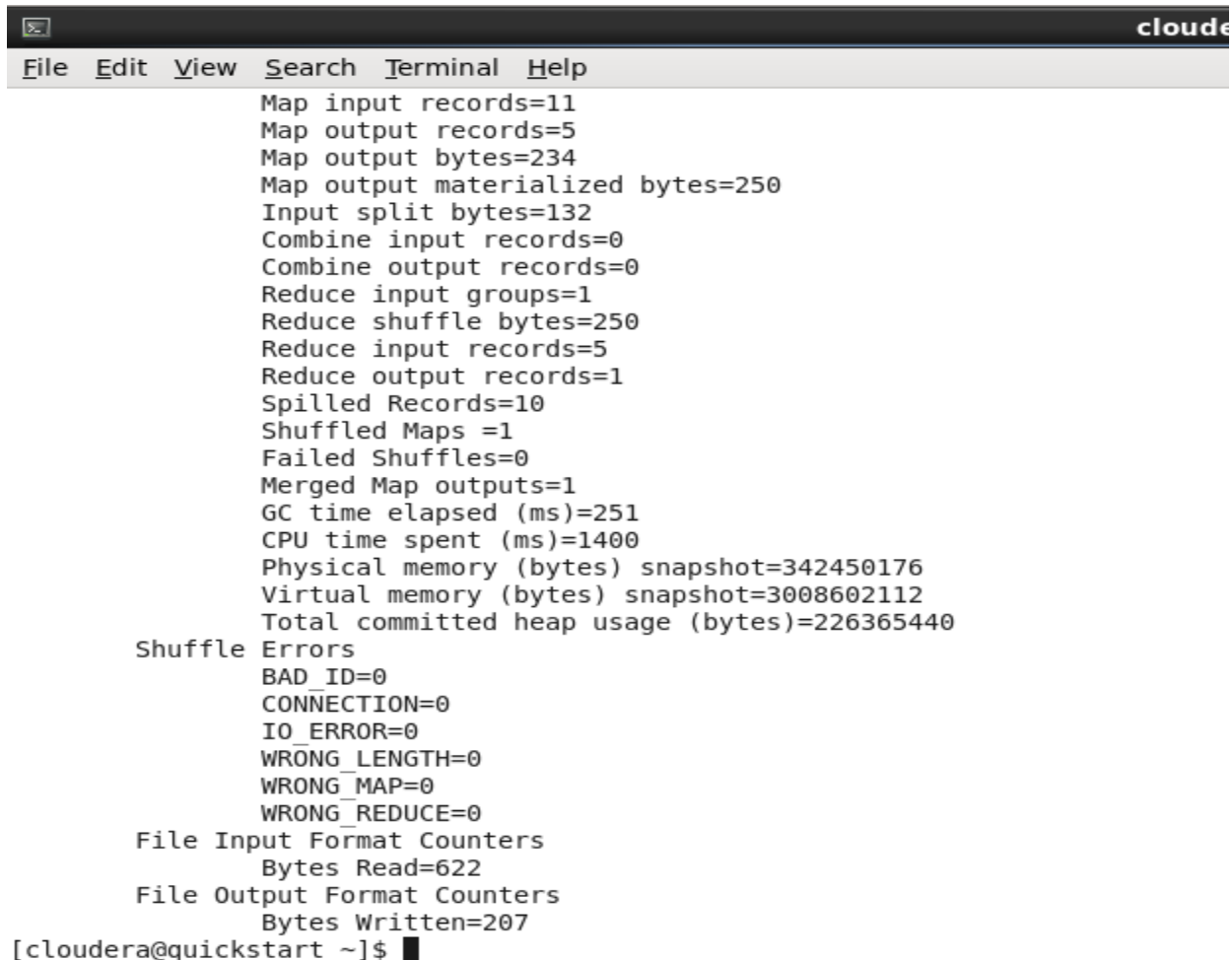
INFO

Last modified

Home   / user / cloudera / selesoutput2 / **part-r-00000**

```
Result  Ava Martinez,Marketing,Marketing Manager
Jasmine Wong,Sales,Sales Representative
Sarah Chen,Human Resources,HR Manager
Michael Lee,Finance,Financial Analyst
John Smith,Engineering,Software Developer
```

```
[cloudera@quickstart ~]$ hadoop jar SelesEmployeeFinder1.jar SalesEmployeeFinder
 /user/cloudera/seles /user/cloudera/selesoutput2
23/03/15 20:32:05 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0
:8032
23/03/15 20:32:06 WARN mapreduce.JobResourceUploader: Hadoop command-line option
 parsing not performed. Implement the Tool interface and execute your applicatio
n with ToolRunner to remedy this.
23/03/15 20:32:07 INFO input.FileInputFormat: Total input paths to process : 1
23/03/15 20:32:07 INFO mapreduce.JobSubmitter: number of splits:1
23/03/15 20:32:07 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_16
78934192891_0003
23/03/15 20:32:07 INFO impl.YarnClientImpl: Submitted application application_16
78934192891_0003
23/03/15 20:32:07 INFO mapreduce.Job: The url to track the job: http://quickstar
t.cloudera:8088/proxy/application_1678934192891_0003/
23/03/15 20:32:07 INFO mapreduce.Job: Running job: job_1678934192891_0003
23/03/15 20:32:21 INFO mapreduce.Job: Job job_1678934192891_0003 running in ub
er mode : false
23/03/15 20:32:21 INFO mapreduce.Job:  map 0% reduce 0%
23/03/15 20:32:40 INFO mapreduce.Job:  map 100% reduce 0%
23/03/15 20:32:53 INFO mapreduce.Job:  map 100% reduce 100%
23/03/15 20:32:53 INFO mapreduce.Job: Job job_1678934192891_0003 completed suc
cessfully
23/03/15 20:32:53 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=250
                FILE: Number of bytes written=223361
                FILE: Number of read operations=0
```

File Edit View Search Terminal Help

```
                Map input records=11
                Map output records=5
                Map output bytes=234
                Map output materialized bytes=250
                Input split bytes=132
                Combine input records=0
                Combine output records=0
                Reduce input groups=1
                Reduce shuffle bytes=250
                Reduce input records=5
                Reduce output records=1
                Spilled Records=10
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=251
                CPU time spent (ms)=1400
                Physical memory (bytes) snapshot=342450176
                Virtual memory (bytes) snapshot=3008602112
                Total committed heap usage (bytes)=226365440
        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=622
        File Output Format Counters
                Bytes Written=207
[cloudera@quickstart ~]$
```
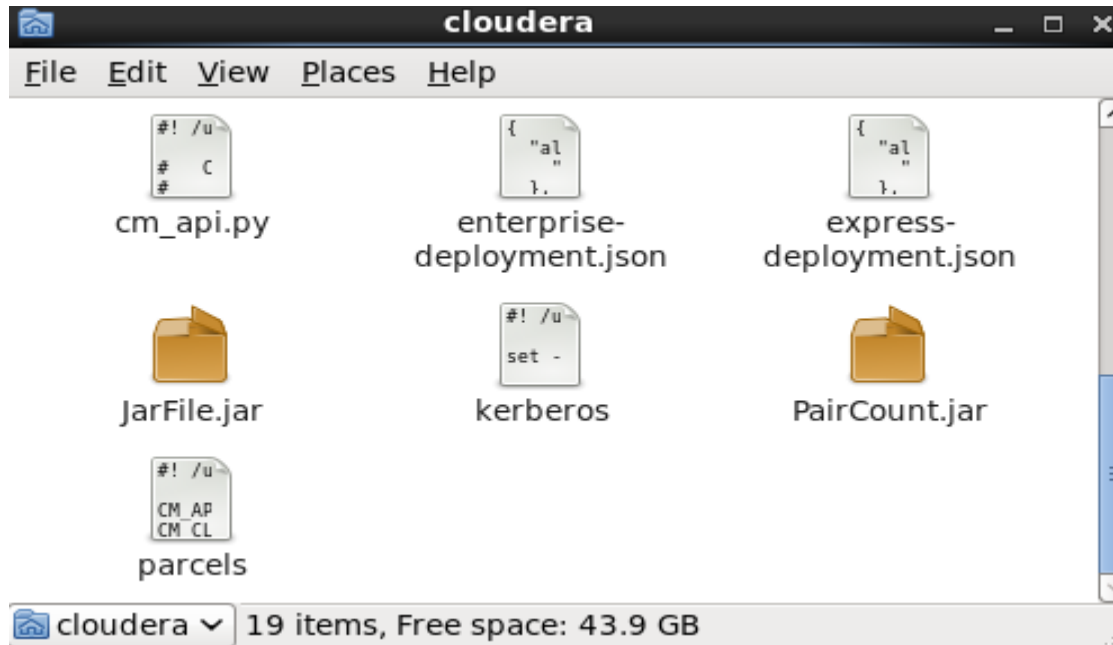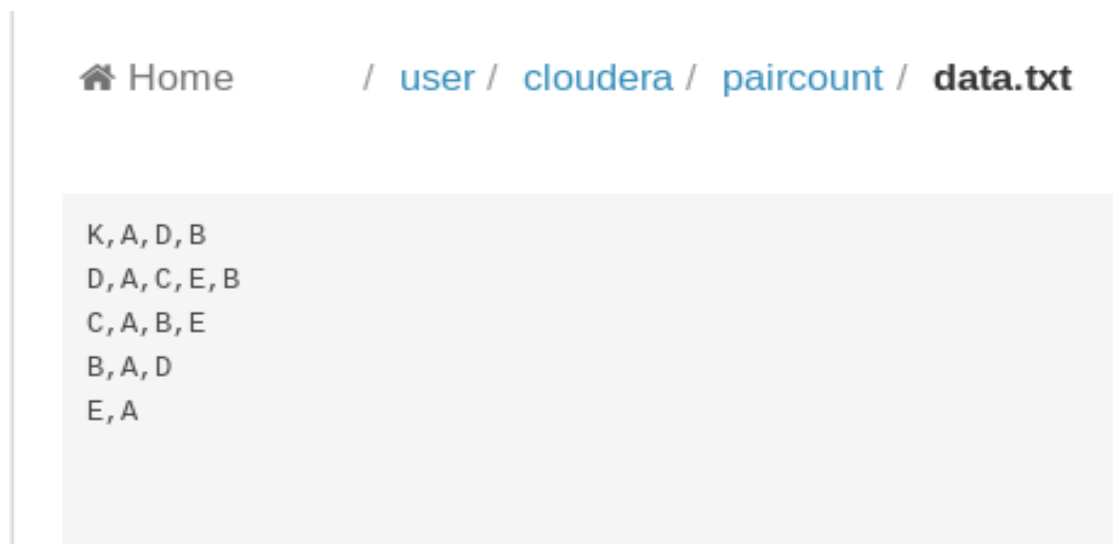
LAB 5

JAVA FILE:

PairCount.java ⊠

```java
22        private Text p = new Text();
23
24        public void map(LongWritable key, Text value, Context context) throws IOException, InterruptedException {
25            String line = value.toString();
26            StringTokenizer tokenizer = new StringTokenizer(line);
27            String[] items = tokenizer.nextToken().split(",");
28
29            Arrays.sort(items);
30
31            for (int i = 0; i < items.length - 1; i++) {
32                for (int j = i + 1; j < items.length; j++) {
33
34                    if(items[i].charAt(0)<items[j].charAt(0)){
35
36                        p.set(items[i] + "," + items[j]);
37                        context.write(p, o);
38                    }
39                    else{
40                        p.set(items[j] + "," + items[i]);
41                        context.write(p, o);
42                    }
43
44                }
45            }
46        }
47    }
48
49    public static class PairCountReducer extends Reducer<Text, IntWritable, Text, IntWritable> {
50
51        private IntWritable count = new IntWritable();
52
53        public void reduce(Text key, Iterable<IntWritable> values, Context context) throws IOException, InterruptedException {
54            int sum = 0;
55
56            Iterator<IntWritable> iterator = values.iterator();
57
58            while (iterator.hasNext()) {
59                sum += iterator.next().get();
60            }
61            count.set(sum);
62            context.write(key, count);
63        }
64    }
65
66    public static void main(String[] args) throws Exception {
67        Configuration conf = new Configuration();
68        Job job = Job.getInstance(conf, "pair count");
69
```

JAR file Create "Pair Count":

Dataset input in Hue Cloudera:


```
K, A, D, B
D, A, C, E, B
C, A, B, E
B, A, D
E, A
```

Run in Terminal:

```
                Bytes Written=78
[cloudera@quickstart ~]$ hadoop jar PairCount.jar PairCount /user/cloudera/pairc
ount /user/cloudera/paircountoutput6
23/03/22 20:22:03 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0
:8032
23/03/22 20:22:04 WARN mapreduce.JobResourceUploader: Hadoop command-line option
 parsing not performed. Implement the Tool interface and execute your applicatio
n with ToolRunner to remedy this.
23/03/22 20:22:04 INFO input.FileInputFormat: Total input paths to process : 1
23/03/22 20:22:04 INFO mapreduce.JobSubmitter: number of splits:1
23/03/22 20:22:04 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_16
77120724034_0009
23/03/22 20:22:05 INFO impl.YarnClientImpl: Submitted application application_16
77120724034_0009
23/03/22 20:22:05 INFO mapreduce.Job: The url to track the job: http://quickstar
t.cloudera:8088/proxy/application_1677120724034_0009/
23/03/22 20:22:05 INFO mapreduce.Job: Running job: job_1677120724034_0009
23/03/22 20:22:14 INFO mapreduce.Job: Job job_1677120724034_0009 running in uber
 mode : false
23/03/22 20:22:14 INFO mapreduce.Job:  map 0% reduce 0%
23/03/22 20:22:23 INFO mapreduce.Job:  map 100% reduce 0%
23/03/22 20:22:32 INFO mapreduce.Job:  map 100% reduce 100%
23/03/22 20:22:33 INFO mapreduce.Job: Job job_1677120724034_0009 completed succe
ssfully
```

```
                  Reduce shuffle bytes=136
                  Reduce input records=13
                  Reduce output records=13
                  Spilled Records=26
                  Shuffled Maps =1
                  Failed Shuffles=0
                  Merged Map outputs=1
                  GC time elapsed (ms)=181
                  CPU time spent (ms)=1350
                  Physical memory (bytes) snapshot=326934528
                  Virtual memory (bytes) snapshot=3008466944
                  Total committed heap usage (bytes)=226365440
          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=35
          File Output Format Counters
                  Bytes Written=78
[cloudera@quickstart ~]$ █
```

Output:

File Browser

ACTIONS

- View as binary
- Edit file
- Download
- View file location
- Refresh

INFO

Last modified
March 22, 2023 8:22 p.m.

Home / user / cloudera / paircountoutput6 / **part-r-00000**

```
A,B     4
A,C     2
A,D     3
A,E     3
A,K     1
B,C     2
B,D     3
B,E     2
B,K     1
C,D     1
C,E     2
D,E     1
```