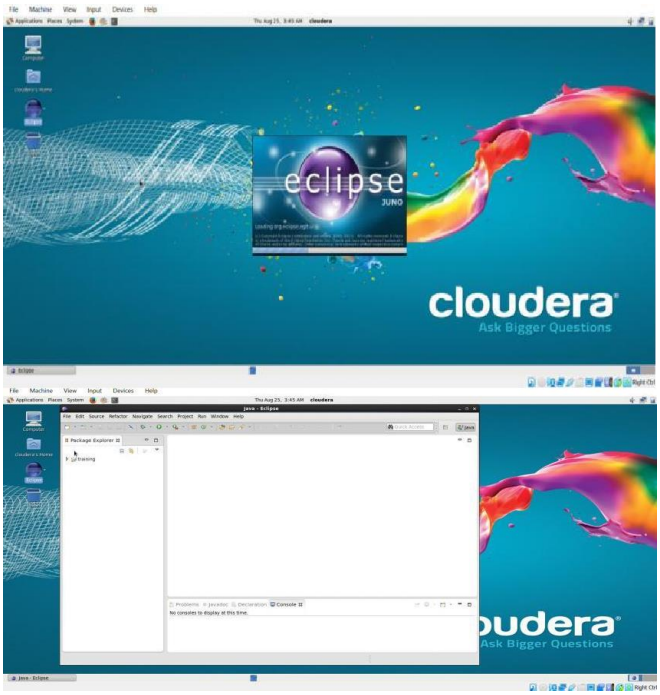
Experiment No:8

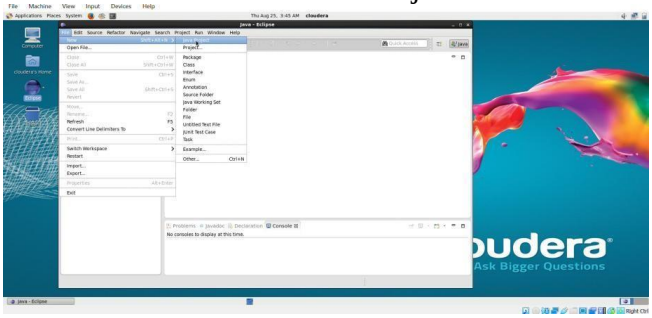
Aim: Implement User Defined functions in PIG.

Procedure:

1. Open Oracle VM Virtual box -> click start -> open cloudera -> open eclipse.

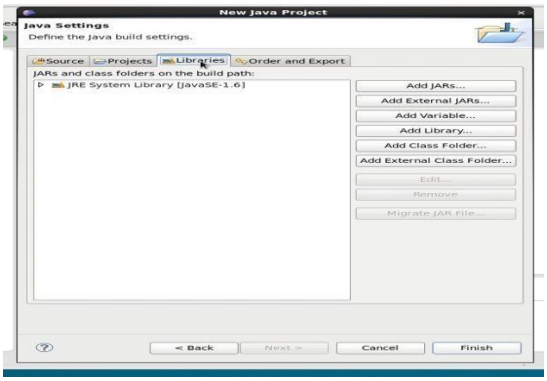


2. Click on file->New->Java Project.

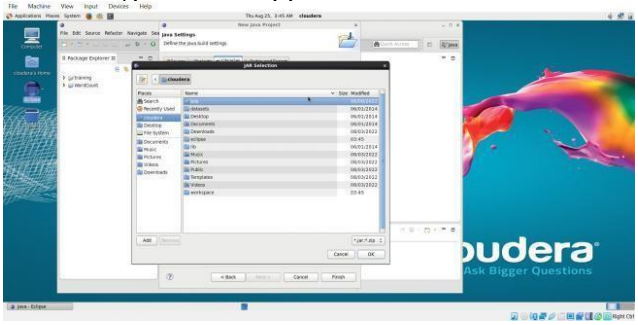


3. Give your project name as "**UDF**" and click on next.

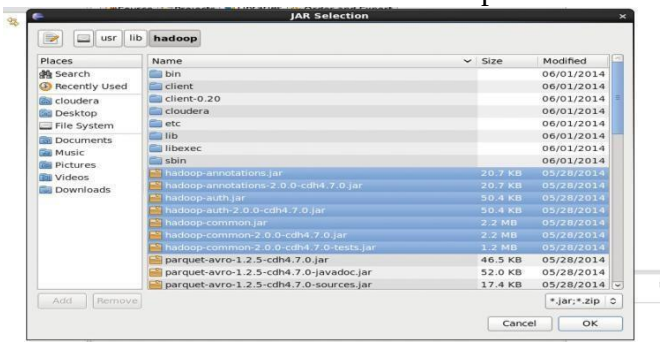
4. After that, click on libraries on the top and click on add external jars on the right side.



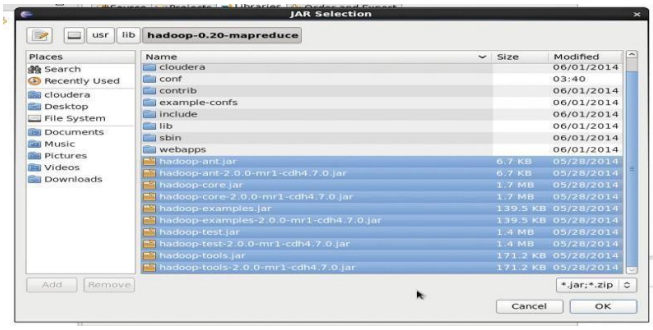
5. A dialog box will appear.



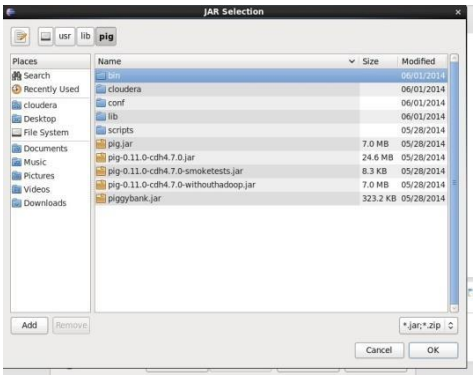
6. Click on file system on the left, and then usr->lib->hadoop.
Select all the files named with hadoop and click ok.



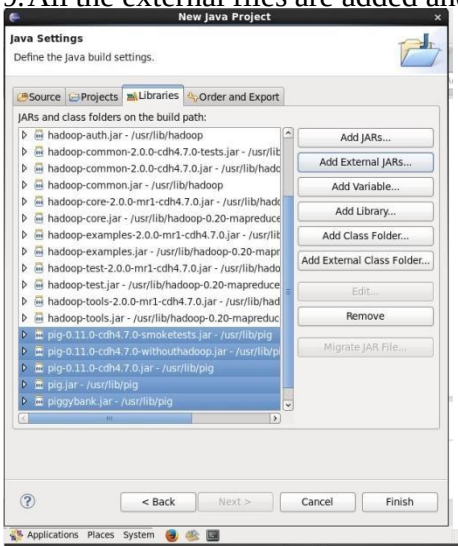
7. Go back to lib and click **hadoop-0.20-mapreduce** and select all the files named with hadoop and click ok.

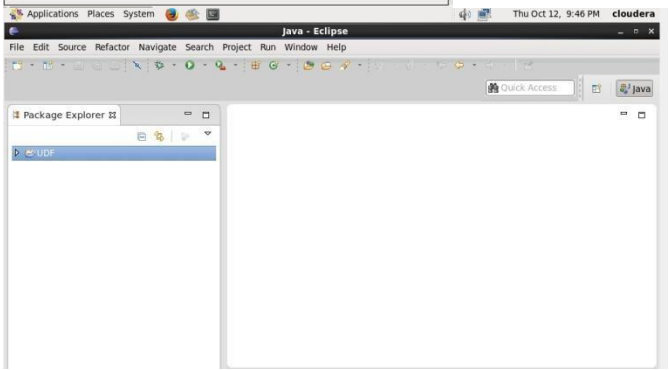


8. Go back to lib and click **pig** and select all the files named with PIG and click ok.

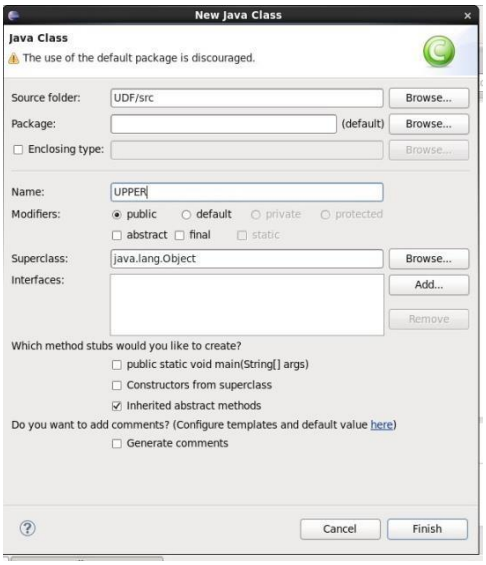


9. All the external files are added and now click on finish. The project is created successfully.





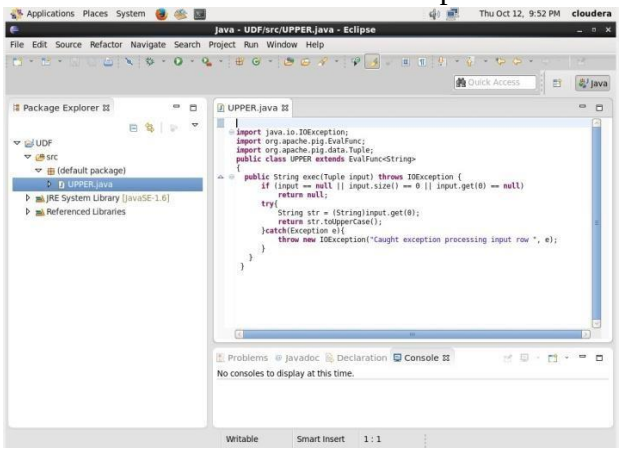
10. Create a class with name "**UPPER**" and click on finish.



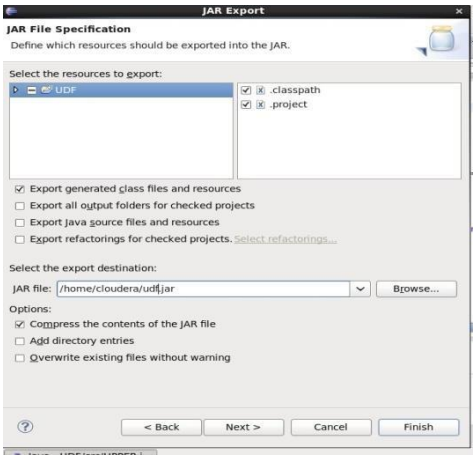
11. Open chrome and search for PIG UDF in Apache pig and copy the program into eclipse.

```
import java.io.IOException;
import org.apache.pig.EvalFunc;
import org.apache.pig.data.Tuple;
public class UPPER extends EvalFunc<String>
{
  public String exec(Tuple input) throws IOException {
      if (input == null || input.size() == 0 || input.get(0) == null)
          return null;
      try{
          String str = (String)input.get(0);
        return str.toUpperCase();
     }catch(Exception e){
        throw new IOException("Caught exception processing input row ", e);
      }
  }
}
```

12. Copy and paste the program 4 lines at a time-If copy and paste doesn't work then goto device in cloudera and set both shared clipboard and drag and drop as **bidirectional.**



13. After finishing all the programs, right click on your project and select export Select java-> JAR file->next, Click on browse and a dialog box will appear, give any name for the jar file and click ok and click on finish.



14. Now go to cloud environment →click on terminal.Check whether the jar file is added on local system or not.

15. Create a directory in the local system named as **"pigexample"** .
16. Create a text file in local system named as **"pigsample2.txt"** and put the text file into directory.

```
[cloudera@localhost ~]$ touch pigsample2.txt
[cloudera@localhost ~]$ vi pigsample2.txt
[cloudera@localhost ~]$ cat pigsample2.txt
1,Divya,CSE
2,Padhu,CSE
3,Mouni,CSE
4,Nuthana,IT
5,Suma,IT
6,Swathi,AIDS
7,Raju,ECE
8,Manoj,ECE
9,Rakesh,EEE
10,Suresh,AIML
[cloudera@localhost ~]$
```

```
[cloudera@localhost ~]$ hadoop fs -mkdir pigexample
[cloudera@localhost ~]$ hadoop fs -put /home/cloudera/pigsample2.txt /user/cloud
era/pigexample
```

17. Open Pig Environment.

```
[cloudera@localhost ~]$ pig
2023-10-12 21:59:21,918 [main] INFO  org.apache.pig.Main - Apache Pig version 0.
11.0-cdh4.7.0 (rexported) compiled May 28 2014, 11:05:48
2023-10-12 21:59:21,918 [main] INFO  org.apache.pig.Main - Logging error message
s to: /home/cloudera/pig_1697173161917.log
2023-10-12 21:59:21,940 [main] INFO  org.apache.pig.impl.util.Utils - Default bo
otup file /home/cloudera/.pigbootup not found
2023-10-12 21:59:22,137 [main] WARN  org.apache.hadoop.conf.Configuration - fs.d
efault.name is deprecated. Instead, use fs.defaultFS
2023-10-12 21:59:22,137 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.HExecutionEngine - Connecting to hadoop file system at: hdfs://localhost.loca
ldomain:8020
2023-10-12 21:59:22,551 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.HExecutionEngine - Connecting to map-reduce job tracker at: localhost.localdo
main:8021
2023-10-12 21:59:22,552 [main] WARN  org.apache.hadoop.conf.Configuration - fs.d
efault.name is deprecated. Instead, use fs.defaultFS
grunt>
```

18. Register the jar file.

Syntax:REGISTER jarfile location

```
grunt> REGISTER /home/cloudera/udf.jar
2023-10-12 22:02:57,762 [main] WARN  org.apache.hadoop.conf.Configuration - dfs.
df.interval is deprecated. Instead, use fs.df.interval
2023-10-12 22:02:57,762 [main] WARN  org.apache.hadoop.conf.Configuration - dfs.
max.objects is deprecated. Instead, use dfs.namenode.max.objects
2023-10-12 22:02:57,762 [main] WARN  org.apache.hadoop.conf.Configuration - hado
op.native.lib is deprecated. Instead, use io.native.lib.available
```

19. Define UDF and load text file.

Syntax:DEFINE function-name class-name();

```
grunt> DEFINE myfun UPPER();
grunt> stu= load '/user/cloudera/pigexample/pigsample2.txt' using PigStorage(','
) as (sid:int,sname:chararray,dept:chararray);
grunt> dump stu;
2023-10-12 22:05:24,009 [main] INFO  org.apache.pig.tools.pigstats.ScriptState -
 Pig features used in the script: UNKNOWN
2023-10-12 22:05:24,072 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? fal
se
2023-10-12 22:05:24,084 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
```

20. Output:

```
2023-10-12 22:05:36,649 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MapReduceLauncher - Success!
2023-10-12 22:05:36,651 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Ke
y [pig.schematuple] was not set... will not generate code.
2023-10-12 22:05:36,667 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileI
nputFormat - Total input paths to process : 1
2023-10-12 22:05:36,667 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.util.MapRedUtil - Total input paths to process : 1
(1,Divya,CSE)
(2,Padhu,CSE)
(3,Mouni,CSE)
(4,Nuthana,IT)
(5,Suma,IT)
(6,Swathi,AIDS)
(7,Raju,ECE)
(8,Manoj,ECE)
(9,Rakesh,EEE)
(10,Suresh,AIML)
grunt>
```

21. Use FOREACH to generate your function.

Syntax:variable= FOREACH data generate function-name(attribute);

```
grunt> stu_UPPER= FOREACH stu GENERATE myfun(sname);
grunt> dump stu_UPPER;
2023-10-12 22:09:17,346 [main] INFO  org.apache.pig.tools.pigstats.ScriptState -
 Pig features used in the script: UNKNOWN
2023-10-12 22:09:17,348 [main] INFO  org.apache.pig.newplan.logical.rules.Column
PruneVisitor - Columns pruned for stu: $0, $2
2023-10-12 22:09:17,351 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? fal
se
```