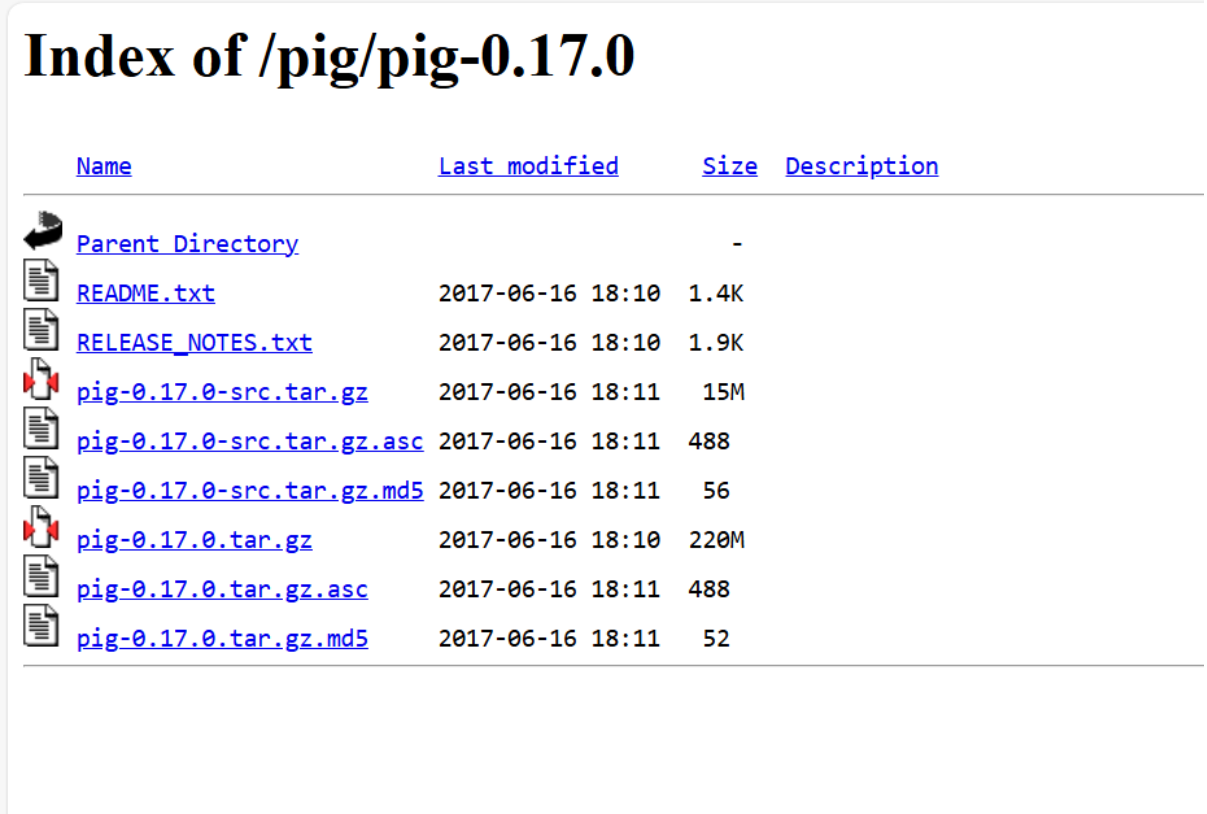
Create UDF (User Defined Functions) in Apache Pig and execute it in MapReduce / HDFS mode

Download Pig from https://downloads.apache.org/pig/pig-0.17.0/

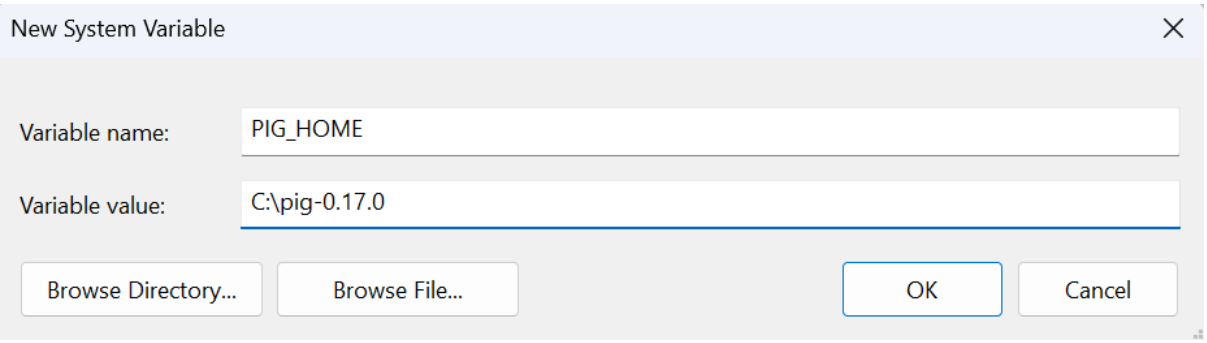


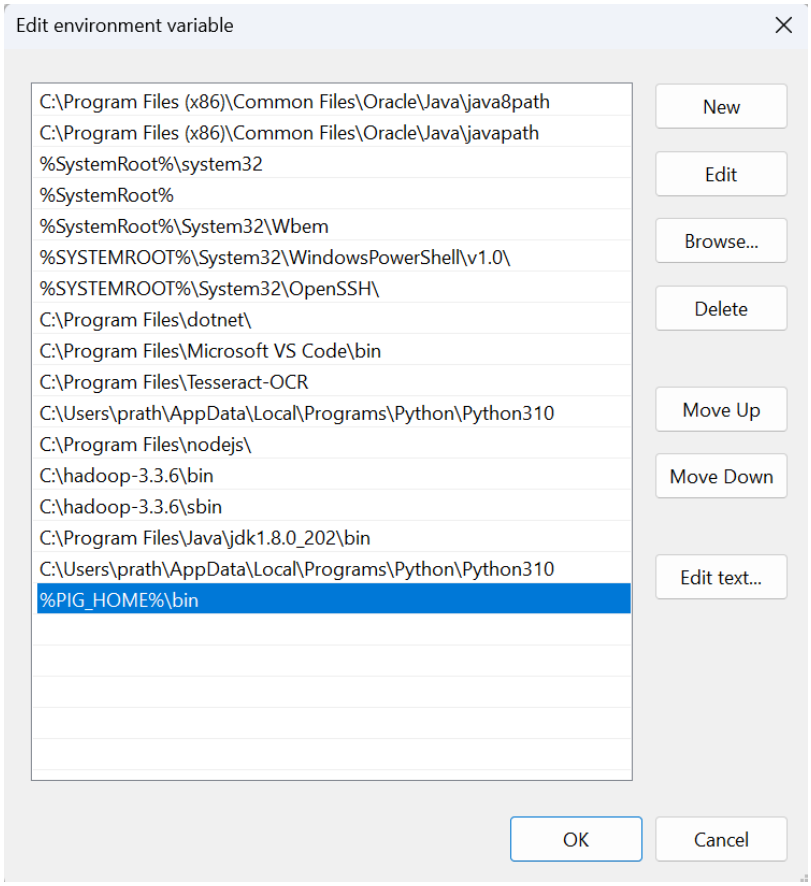
Index of /pig/pig-0.17.0

Name	Last modified	Size	Description
Parent Directory		-	
README.txt	2017-06-16 18:10	1.4K	
RELEASE_NOTES.txt	2017-06-16 18:10	1.9K	
pig-0.17.0-src.tar.gz	2017-06-16 18:11	15M	
pig-0.17.0-src.tar.gz.asc	2017-06-16 18:11	488	
pig-0.17.0-src.tar.gz.md5	2017-06-16 18:11	56	
pig-0.17.0.tar.gz	2017-06-16 18:10	220M	
pig-0.17.0.tar.gz.asc	2017-06-16 18:11	488	
pig-0.17.0.tar.gz.md5	2017-06-16 18:11	52	

Extract the files and save it in the desired location.

Add environment variable for Pig.





Go to C:\pig-0.16.0\bin and open pig (Windows Command Script)

```
set HADOOP_BIN_PATH=%HADOOP_HOME%\libexec
```

Open Windows Powershell and type “pig –x local”

```
PS C:\Users\prath> pig -x local
2024-09-11 13:50:35,862 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-09-11 13:50:35,862 INFO pig.ExecTypeProvider: Picked LOCAL as the ExecType
2024-09-11 13:50:36,121 [main] INFO  org.apache.pig.Main - Apache Pig version 0.17.0 (r1797386) compiled Jun 02 2017, 15
:41:58
2024-09-11 13:50:36,121 [main] INFO  org.apache.pig.Main - Logging error messages to: C:\hadoop-3.3.6\logs\pig_172604283
6118.log
2024-09-11 13:50:36,140 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file C:\Users\prath/.pigbootup not
found
2024-09-11 13:50:36,508 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated
. Instead, use mapreduce.jobtracker.address
2024-09-11 13:50:36,511 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hado
op file system at: file:///
2024-09-11 13:50:36,579 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is depreca
ted. Instead, use dfs.bytes-per-checksum
2024-09-11 13:50:36,605 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-5b9eeaeb-7019
-4e7f-827c-caa4a14fcd43
2024-09-11 13:50:36,605 [main] WARN  org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set
to false
grunt>
```

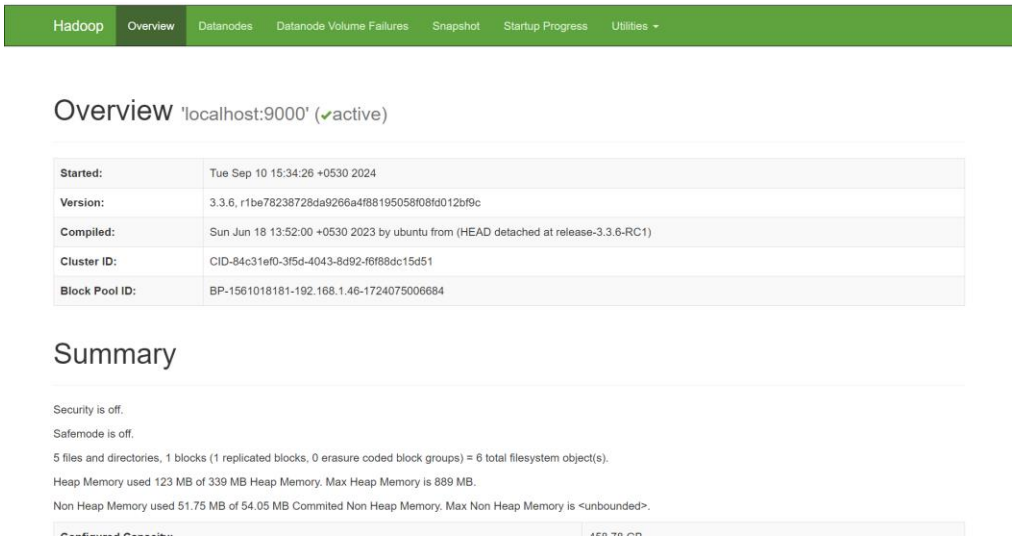
Open command prompt and run as administrator

Start Hadoop services by typing in the following commands:

- start-dfs.cmd
- start-yarn.cmd

```
C:\Windows\System32>jps
14212 Jps

C:\Windows\System32>start-dfs.cmd

C:\Windows\System32>jps
12000 DataNode
16488 Jps
24904 NameNode

C:\Windows\System32>start-yarn.cmd
starting yarn daemons

C:\Windows\System32>jps
12000 DataNode
6384 NodeManager
31300 Jps
24904 NameNode
29036 ResourceManager

C:\Windows\System32>
```

Open the browser and go to the URL localhost:9870



Create a text file “pig_udf_text.txt”:

```
1,hello
2,pig
3,user
4,apache
```

Create a Directory in HDFS and copy the Input File to HDFS using the following commands:

hdfs dfs -mkdir -p /pig/hadoop/input

hadoop fs -put C:/Semester7/DataAnalytics/Lab/Ex4/pig_udf_text.txt /pig/hadoop/input/

```
C:\hadoop-3.3.6\sbin>hdfs dfs -mkdir -p /pig/hadoop/input

C:\hadoop-3.3.6\sbin>hadoop fs -put C:/Semester7/DataAnalytics/Lab/Ex4/pig_udf_text.txt /pig/hadoop/input/

C:\hadoop-3.3.6\sbin>
```

Create a Python file “uppercase_udf.py”

```python
def uppercase(text):
    return text.upper()

if __name__=="__main__":
    import sys
    for line in sys.stdin:
        line=line.strip()
        result=uppercase(line)
        print(result)
```

Create a Directory in HDFS and copy the Python File to HDFS using the following commands:

hdfs dfs -mkdir -p /pig/hadoop/udf

hadoop fs -put C:/Semester7/DataAnalytics/Lab/Ex4/uppercase_udf.py /pig/hadoop/udf/

```
C:\hadoop-3.3.6\sbin>hdfs dfs -mkdir -p /pig/hadoop/udf

C:\hadoop-3.3.6\sbin>hadoop fs -put C:/Semester7/DataAnalytics/Lab/Ex4/uppercase_udf.py /pig/hadoop/udf/

C:\hadoop-3.3.6\sbin>
```

Create pig file “script.pig”:

```
REGISTER "hdfs:///pig/hadoop/udf/uppercase_udf.py" USING jython AS udf;
data = LOAD "hdfs:///pig/hadoop/input/pig_udf_text.txt" AS (text:chararray);
uppercased_data = FOREACH data GENERATE udf.uppercase(text) AS uppercase_text;
STORE uppercased_data INTO "hdfs:///pig/hadoop/output";
```

Execute the pig file using the command:

pig -f script.pig

```
C:\Windows\System32>pig -f C:/Semester7/DataAnalytics/Lab/Ex4/script.pig
```

View the output using the following command:

```
C:\Windows\System32>hdfs dfs -cat /pig/hadoop/output/part-m-00000
1,HELLO
2,PIG
3,USER
4,APACHE
```

View the output in the file system on the browser

Hadoop Overview Datanodes Datanode Volume Failures Snapshot Startup Progress Utilities

Browse Directory

/pig/hadoop/output Go!

Show 25 entries Search:

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	prath	supergroup	0 B	Sep 11 14:46	1	128 MB	_SUCCESS
-rw-r--r--	prath	supergroup	30 B	Sep 11 14:46	1	128 MB	part-m-00000

Showing 1 to 2 of 2 entries Previous 1 Next

Hadoop, 2023.

File contents

```
1,HELLO
2,PIG
3,USER
4,APACHE
```