

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

Aim:

To create UDF in Apache Pig and execute it in MapReduce/HDFS mode.

Procedure:

Pig Download and installation:

1. Download Pig:

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	

2. Add the environment variable for Pig:

Edit System Variable

Variable name: PIG_HOME

Variable value: C:\pig-0.17.0

Browse Directory... Browse File... OK Cancel

Edit environment variable

C:\Program Files (x86)\Common Files\Oracle\Java\javapath
C:\Program Files (x86)\Common Files\Oracle\Java\javapath
C:\Program Files\Python311\Scripts\
C:\Program Files\Python311\
%SystemRoot%\system32
%SystemRoot%
%SystemRoot%\System32\Wbem
%SYSTEMROOT%\System32\WindowsPowerShell\v1.0\
%SYSTEMROOT%\System32\OpenSSH\
C:\Users\Admin\AppData\Roaming\Python\Python311\Scripts
C:\Program Files\nodejs\
D:\Admin\Git\cmd
C:\Java\jdk-1.8\bin
C:\Hadoop\bin
C:\Hadoop\sbin
C:\Python39\
%PIG_HOME%\bin

New Edit Browse... Delete Move Up Move Down Edit text... OK Cancel

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

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

4. Open Windows Powershell and type “pig –x local” and check whether pig grunt appears.

Pig is successfully installed.

Create UDF:

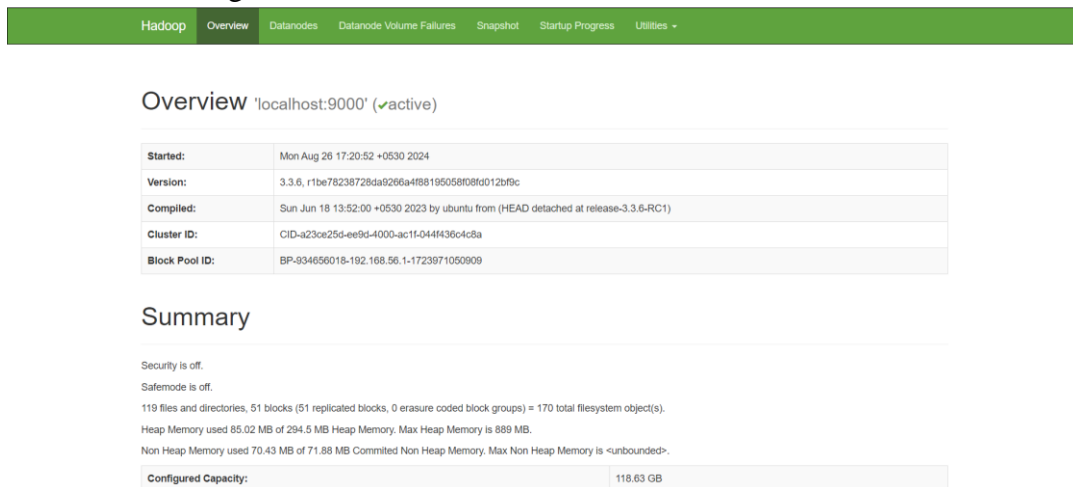
1. Start Hadoop services:

Open command prompt as an administrator

```
start-dfs.cmd
```

```
start-yarn.cmd
```

2. Open the browser and go to the URL “localhost:9870”

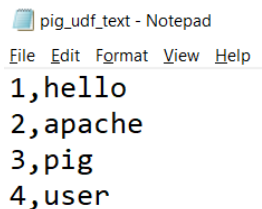


The screenshot shows the Hadoop Overview page at localhost:9870. The page has a green header with navigation links: Hadoop, Overview, Datanodes, Datanode Volume Failures, Snapshot, Startup Progress, and Utilities. The main content area is titled 'Overview' and shows the cluster status as 'localhost:9800' (active). Below this, there is a table with the following information:

Property	Value
Started:	Mon Aug 26 17:20:52 +0530 2024
Version:	3.3.6, r1be78238728da9266a4f88195058f0d012bffc
Compiled:	Sun Jun 18 13:52:00 +0530 2023 by ubuntu from (HEAD detached at release-3.3.6-RC1)
Cluster ID:	CID-a23ce25d-ee9d-4000-ac1f-044f436c4c8a
Block Pool ID:	BP-934656018-192.168.56.1-1723971050909

Below the table, there is a 'Summary' section. It states that Security is off, Safemode is off, and provides details about the filesystem objects, heap memory usage, and non-heap memory usage. At the bottom, it shows the 'Configured Capacity' as 118.63 GB.

3. Create a text file “pig_udf_text.txt”:



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

4. Create a Directory in HDFS and copy the Input File to HDFS

```
hdfs dfs -mkdir /user/Admin/home/hadoop/piginput
```

```
hadoop fs -put C:/Hadoop/piginput/pig_udf_text.txt /user/Admin/home/hadoop/piginput/
```

```
C:\>hdfs dfs -mkdir /user/Admin/home/hadoop/piginput
```

```
C:\>hadoop fs -put C:/Hadoop/piginput/pig_udf_text.txt /user/Admin/home/hadoop/piginput/
```

5. Create a Python file “uppercase_udf.py”:

```
uppercase_udf - Notepad
File Edit Format View Help
def uppercase(text):
    return text.upper()

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

6. Create a Directory in HDFS and copy the Input File to HDFS

```
hdfs dfs -mkdir /user/Admin/home/hadoop/udfs
```

```
hadoop fs -put C:/Users/Admin/uppercase_udf.py /user/Admin/home/hadoop/udfs/
```

```
C:\>hdfs dfs -mkdir /user/Admin/home/hadoop/udfs
C:\>hadoop fs -put C:/Users/Admin/uppercase_udf.py /user/Admin/home/hadoop/udfs/
```

7. Create pig file “script.pig”:

```
script - Notepad
File Edit Format View Help
-- Register the Python UDF script
REGISTER 'hdfs:///user/Admin/home/hadoop/udfs/uppercase_udf.py' USING jython AS udf;
-- Load some data
data = LOAD 'hdfs:///user/Admin/home/hadoop/piginput/pig_udf_text.txt' AS (text:chararray);
-- Use the Python UDF
uppercased_data = FOREACH data GENERATE udf.uppercase(text) AS uppercase_text;
-- Store the result
STORE uppercased_data INTO 'hdfs:///user/Admin/home/hadoop/pig_output_data';
```

8. Execute Pig file:

```
pig -f C:/Users/Admin/script.pig
```

```
C:\>pig -f C:/Users/Admin/script.pig
2024-08-26 19:02:52,575 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-08-26 19:02:52,578 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2024-08-26 19:02:52,579 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2024-08-26 19:02:53,142 [main] INFO org.apache.pig.Main - Apache Pig version 0.17.0 (r1797386) compiled Jun 02 2017, 15:41:58
2024-08-26 19:02:53,142 [main] INFO org.apache.pig.Main - Logging error messages to: C:\Hadoop\logs\pig_1724679173133.log
2024-08-26 19:02:53,674 [main] INFO org.apache.pig.impl.util.Utils - Default bootup file C:\Users\Admin/.pigbootup not found
2024-08-26 19:02:53,794 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use
2024-08-26 19:02:53,794 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system
2024-08-26 19:02:54,842 [main] INFO org.apache.pig.PigServer - Pig Script ID for the session: PIG-script.pig-d6bf1250-8d4f-4b65-8519-
2024-08-26 19:02:54,846 [main] WARN org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
2024-08-26 19:02:55,554 [main] INFO org.apache.pig.scripting.jython.JythonScriptEngine - created tmp python.cachedir=C:\Users\Admin\
2024-08-26 19:03:11,501 [JobControl] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2024-08-26 19:03:11,502 [JobControl] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
2024-08-26 19:03:11,540 [JobControl] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths (combined) to process : 1
2024-08-26 19:03:12,073 [JobControl] INFO org.apache.hadoop.mapreduce.JobSubmitter - number of splits:1
```

9. View the Output

```
hdfs dfs -ls /user/Admin/home/hadoop/pig_output_data
```

```
C:\>hdfs dfs -ls /user/Admin/home/hadoop/pig_output_data
Found 2 items
-rw-r--r--  1 Admin supergroup      0 2024-08-26 19:04 /user/Admin/home/hadoop/pig_output_data/_SUCCESS
-rw-r--r--  1 Admin supergroup    30 2024-08-26 19:04 /user/Admin/home/hadoop/pig_output_data/part-m-00000
```

```
hdfs dfs -cat /user/Admin/home/hadoop/pig_output_data/part-m-00000
```

```
C:\>hdfs dfs -cat /user/Admin/home/hadoop/pig_output_data/part-m-00000
1,HELLO
2,APACHE
3,PIG
4,USER
C:\>
```

10. Once the map reduce operations are performed successfully, the output will be present in the specified directory.

“/user/Admin/home/hadoop/pig_output_data/part-m-00000”

The screenshot shows the HDFS web interface. At the top, the path "/user/Admin/home/hadoop/pig_output_data" is entered in the search bar. Below the search bar, there are icons for file operations. A table lists the files in the directory:

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	Admin	supergroup	0 B	Aug 26 19:04	1	128 MB	_SUCCESS
-rw-r--r--	Admin	supergroup	30 B	Aug 26 19:04	1	128 MB	part-m-00000

Below the table, it says "Showing 1 to 2 of 2 entries". To the right of the table are "Previous", "1", and "Next" buttons. Below the table, there is a section titled "File contents" which shows the content of the selected file "part-m-00000":

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

11. Stop Hadoop Services

```
stop-dfs.cmd
```

```
stop-yarn.cmd
```

Result:

Thus, UDF in Apache Pig has been created and executed in MapReduce/HDFS mode successfully.