Ex No 4

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:

1. Install Apache PigDownload Pig:

1. DownloadPigfromtheApachePigdownloadpage:

Link: Apache Pig 0.17.0 Download

Extract the downloaded file (assuming you downloaded pig-0.17.0.tar.gz):

```
tar -xzf pig-0.17.0.tar.gz
```

Move the extracted folder to a directory, such as /usr/local/:

```
sudo mv pig-0.17.0 /usr/local/pig
```

2. Set Up Environment Variables for Pig

Edit your ~/._profileor ~/.zshrc to include Pig in the PATH. nano ~/.zshrc

Add the following lines:

```
export PIG_HOME=/usr/local/pig export
PATH=$PIG_HOME/bin:$PATH
```

Apply the changes:

```
source ~/.zshrc
```

3. Verify Pig Installation

Run the following command to check if Pig is installed correctly: pig -x local

You should see the Pig Grunt shell prompt:

grunt>

Type quitto exit the shell.

4. Start Hadoop Services

Make sure your Hadoop is up and running. Start the required services:

cd /usr/local/hadoop/sbin
./start-dfs.sh
./start-yarn.sh

5. Prepare Input Data **ex4.txt**( )

Create a sample text file for testing the UDF, named ex4.txt: nano ex4.txt

Example content:

1,John
2,Soniya
3,Vijay 4,Sonu

Upload the file to HDFS:

hdfs dfs -mkdir /UDF hdfs dfs -put
ex4.txt /UDF/

6. Create UDF in Python

Now, you need to write your Python UDF. CreateaPythonfile

uppercase_udf.py:

nano uppercase_udf.py

Add the following code to uppercase_udf.py:

```
#!/usr/bin/python3 def
uppercase(text): return
text.upper()

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

Upload the Python UDF to HDFS:

```
hdfs dfs -mkdir /UDF/udfs hdfs dfs -put uppercase_udf.py /UDF/udfs/
```

Make sure the file is in the correct HDFS directory by running: `hdfs dfs -ls /UDF/udfs`

7. Write Pig Script (**UDF.pig**)

Create a Pig script to apply your UDF.

Create `UDF.pig`: `nano UDF.pig`

Add the following Pig script to `UDF.pig`:

```
-- Register the UDF
REGISTER hdfs:///UDF/udfs/uppercase_udf.py USING jython AS myudfs;

-- Load the ex4.txt file from HDFS
data = LOAD 'hdfs:///UDF/ex4.txt' USING PigStorage(',') AS (id:int,name:chararray);

-- Apply the UDF to each line
uppercase_data = FOREACH data GENERATE myudfs.uppercase(name) AS upper_line;
-- Store the result in HDFS
STORE uppercase_data INTO 'hdfs:///UDF/output' USING PigStorage(',');
```

Save the file and exit.

8. Run the Pig Script in MapReduce Mode

Now that everything is set up, execute the Pig script in MapReduce mode:

hdfs dfs -chmod 755 /UDF/udfs/uppercase_udf.py
hdfs dfs -chmod 755 /UDF hdfs dfs -chmod 755
/UDF/ex4.txt

pig -x mapreduce UDF.pig

9. Check the Output

After the job finishes, you can view the output in HDFS.

List the output directory:

hdfs dfs -ls /UDF/output

You should see something like:

Found 1 items

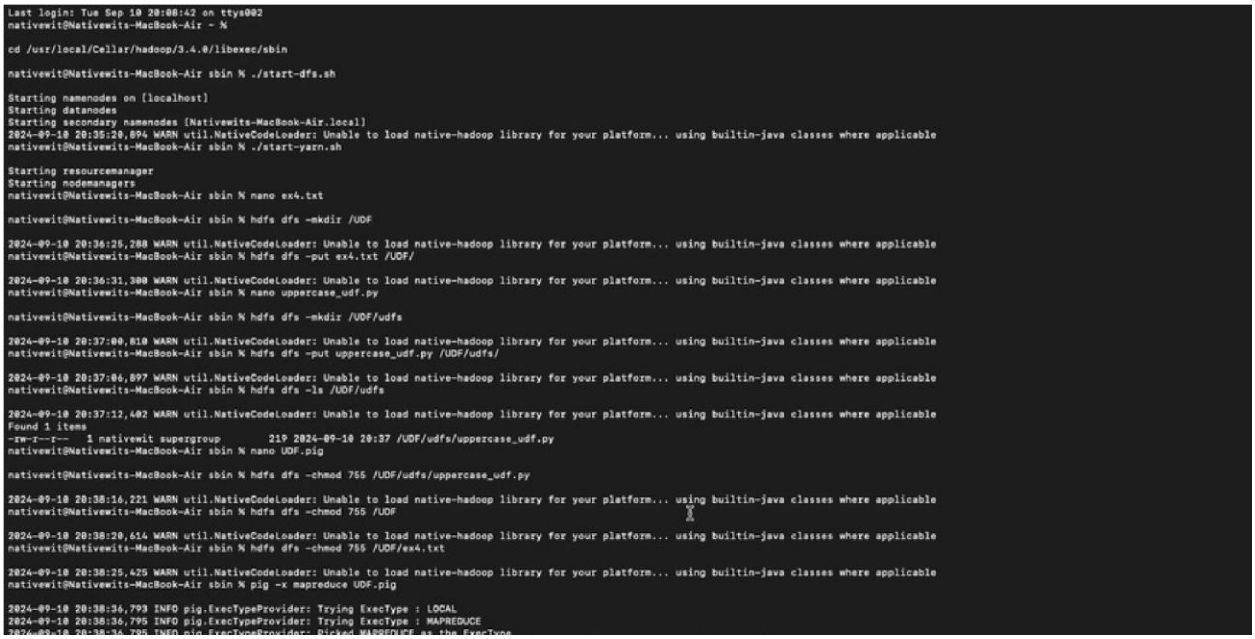
-rw-r--r-- 3usergroup 123 2024-09-11 12:00 /UDF/output/part-m-00000

View the output file:

hdfs dfs -cat /UDF/output/part-m-00000

You should see the content in uppercase

OUTPUT:



RESULT:

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