Experiment 10 : Working with Pig

Aim: Running Apache Pig Script in Local Mode

Procedure:

Step 1: Install OpenJDK 8

Step 2: Download and Extract Apache Pig

Step 3: Set Environment Variables

Step 4: Define the Pig Script

Step 5: Save the Pig Script to a File

Step 6: Specify Input Data Path

Step 7: Construct the Pig Command

Step 8: Run the Pig Script

Step 9: Check Exit Status

Code:

```python
import os
import subprocess

# Install OpenJDK 8 (if not already installed)
java_installation_command = "apt-get install openjdk-8-jre-headless -qq > /dev/null"
subprocess.run(java_installation_command, shell=True, check=True)

# Download and extract Apache Pig
pig_download_url = "https://downloads.apache.org/pig/pig-0.17.0/pig-0.17.0.tar.gz"
subprocess.run(["wget", pig_download_url])
subprocess.run(["tar", "-xzf", "pig-0.17.0.tar.gz"])

# Set environment variables
os.environ['PIG_HOME'] = '/content/pig-0.17.0'
os.environ['PATH'] = os.environ['PIG_HOME'] + '/bin:' +
os.environ['PATH']
os.environ['JAVA_HOME'] = '/usr/lib/jvm/java-8-openjdk-amd64'
os.environ['HADOOP_CONF_DIR'] = '/path/to/hadoop/config/dir'  # Adjust
the path if necessary

# Define the Pig script file
pig_script_content = """
A = LOAD 'students.txt' AS (name:chararray, age:int, gpa:float);
```

```python
B = GROUP A BY age;
C = FOREACH B GENERATE group, A.name;
DUMP C;
"""

# Save the Pig script to a file
with open('/content/students.pig', 'w') as pig_script_file:
    pig_script_file.write(pig_script_content)

input_data_path = '/content/students.txt'

pig_script_path = '/content/students.pig'

pig_command = f'pig -x local -f {pig_script_path} -param input_data={input_data_path}'

try:
    # Run the Pig script
    result = subprocess.run(pig_command, shell=True, text=True, stdout=subprocess.PIPE, stderr=subprocess.PIPE)

    # Check if the exit status is 0 (success)
    if result.returncode == 0:
        print("Apache Pig script executed successfully.")
    else:
        print("Apache Pig script executed with errors. Check the output for details.")

    # Print the Pig script output
    print("Standard Output:")
    print(result.stdout)


except subprocess.CalledProcessError as e:
    print(f"Error: {e}")
```

Output:

```
Apache Pig script executed successfully.
Standard Output:
2023-09-04 06:47:53,946 INFO  [main] pig.ExecTypeProvider
(ExecTypeProvider.java:selectExecType(41)) - Trying ExecType : LOCAL
2023-09-04 06:47:53,949 INFO  [main] pig.ExecTypeProvider
(ExecTypeProvider.java:selectExecType(43)) - Picked LOCAL as the ExecType
(19,{(Doug),(Sally)})
(21,{(John)})
(22,{(Alice)})
(26,{(Susan)})
```