IMAAD HAJWANE

202101132 / 23

BIG DATA ANALYTICS

ASSIGNMENT: INSTALLATION OF PIG

What is Apache Pig?

Apache Pig is a high-level platform for processing large datasets in Hadoop. It provides a scripting language called **Pig Latin** that simplifies writing data analysis programs. Pig converts these scripts into MapReduce jobs to process data stored in HDFS. Pig is especially suitable for programmers who need to process massive data in an easy-to-understand language without writing complex Java-based MapReduce code.

Key Features of Apache Pig:

- **Pig Latin**: A high-level language similar to SQL but designed for data flow (ETL: Extract, Transform, Load) and complex data analysis.
- **Flexibility**: Can handle both structured and unstructured data.
- **Optimization**: Automatically optimizes scripts and converts them into efficient MapReduce jobs.
- **Extensibility**: Allows users to create their own functions (UDFs) for custom processing.

Architecture of Apache Pig

Apache Pig has a multi-layered architecture that simplifies the process of converting high-level scripts into MapReduce jobs. Below are the key components of Pig's architecture:

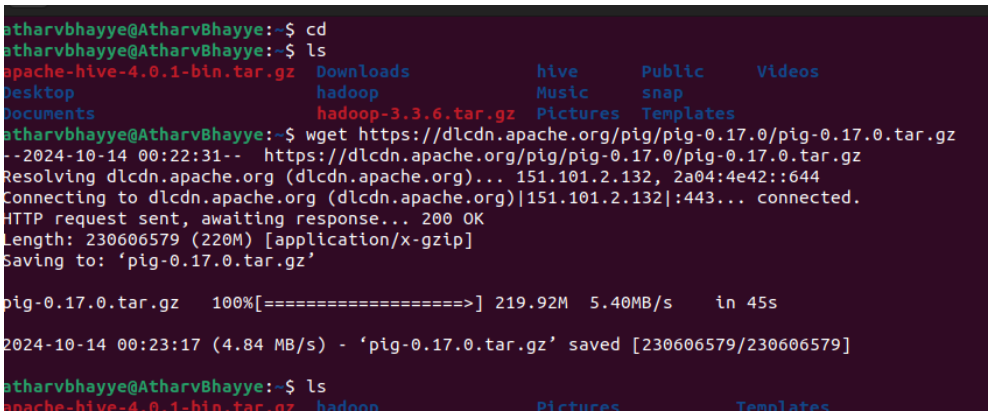
1. **Pig Latin Language**:
    - The scripting language used in Pig, which consists of a series of transformations applied to the data.
    - Each statement in Pig Latin describes a step in the data flow, such as loading data, transforming it, and storing the results.
2. **Parser**:
    - When a Pig Latin script is submitted, the **Parser** first checks the script for syntax and semantic errors.
    - If the script passes the checks, the parser generates a **logical plan** (a series of logical operators) for the data flow.
3. **Optimizer**:
    - The **logical plan** is passed to the **optimizer**, which applies optimization techniques like removing redundant operations, reordering steps, etc.
    - The goal is to make the script more efficient before converting it into physical operations.
4. **Compiler**:
    - The optimized plan is then passed to the **compiler**, which translates the logical operators into a **physical plan**.

- This physical plan consists of MapReduce jobs or other execution models (such as Tez or Spark).

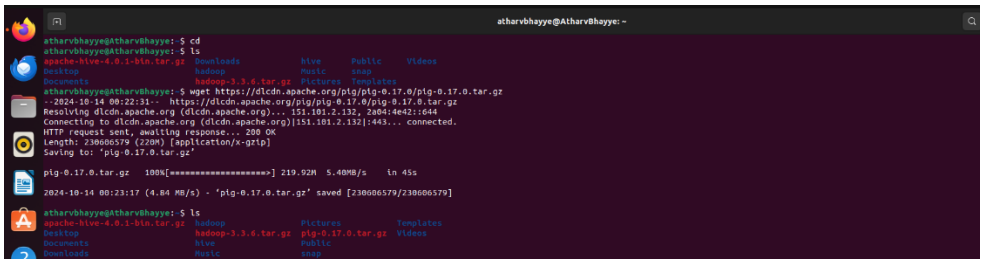
5. **Execution Engine**:
   - The final plan (usually in the form of MapReduce jobs) is executed on the Hadoop cluster.
   - Pig interacts with the **Hadoop Job Tracker** and **Task Tracker** to monitor and control the progress of the jobs.
6. **HDFS**:
   - Pig interacts with Hadoop's HDFS to read input data and write output data.

INSTALLATION STEPS:

- Downloading pig zip file



- Locating pig file in folders



- Extraction, Renaming & Updating bashrc file

```
Downloads                       Music                   snap
atharvbhayye@AtharvBhayye:~$ tar -xzf pig-0.17.0.tar.gz
atharvbhayye@AtharvBhayye:~$ ls
apache-hive-4.0.1-bin.tar.gz  hadoop                  Pictures            snap
Desktop                       hadoop-3.3.6.tar.gz     pig-0.17.0          Templates
Documents                     hive                    pig-0.17.0.tar.gz   Videos
Downloads                     Music                   Public
atharvbhayye@AtharvBhayye:~$ sudo mkdir -p /usr/local/pignew
[sudo] password for atharvbhayye:
atharvbhayye@AtharvBhayye:~$ mv pig-0.17.0 pig
atharvbhayye@AtharvBhayye:~$ cd pig
atharvbhayye@AtharvBhayye:~/pig$ ls
bin          docs      lib-src                 README.txt          test
build.xml    ivy       license                 RELEASE_NOTES.txt   tutorial
CHANGES.txt  ivy.xml   LICENSE.txt             scripts
conf         legacy    NOTICE.txt              shims
contrib      lib       pig-0.17.0-core-h2.jar  src
atharvbhayye@AtharvBhayye:~/pig$ sudo mv * /usr/local/pignew
atharvbhayye@AtharvBhayye:~/pig$ ls
atharvbhayye@AtharvBhayye:~/pig$ cd
atharvbhayye@AtharvBhayye:~$ nano .bashrc
atharvbhayye@AtharvBhayye:~$ source .bashrc
atharvbhayye@AtharvBhayye:~$ pig
2024-10-14 00:34:16,955 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-10-14 00:34:16,958 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2024-10-14 00:34:16,958 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2024-10-14 00:34:17,148 [main] INFO  org.apache.pig.Main - Apache Pig version 0.17.0 (r1797386) compiled Jun 02 2017, 15:41:58
2024-10-14 00:34:17,148 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/atharvbhayye/pig_1728846257140.log
```

- Final Execution

```
atharvbhayye@AtharvBhayye:~/pig$ sudo mv * /usr/local/pignew
atharvbhayye@AtharvBhayye:~/pig$ ls
atharvbhayye@AtharvBhayye:~/pig$ cd
atharvbhayye@AtharvBhayye:~$ nano .bashrc
atharvbhayye@AtharvBhayye:~$ source .bashrc
atharvbhayye@AtharvBhayye:~$ pig
2024-10-14 00:34:16,955 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-10-14 00:34:16,958 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2024-10-14 00:34:16,958 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2024-10-14 00:34:17,148 [main] INFO  org.apache.pig.Main - Apache Pig version 0.17.0 (r1797386) compiled Jun 02 2017, 15:41:58
2024-10-14 00:34:17,148 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/atharvbhayye/pig_1728846257140.log
2024-10-14 00:34:17,195 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/atharvbhayye/.pigbootup not found
2024-10-14 00:34:18,194 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2024-10-14 00:34:18,194 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://localhost:9000
2024-10-14 00:34:20,908 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-61741b10-60bd-489f-bff7-42fc23407e15
2024-10-14 00:34:20,909 [main] WARN  org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
grunt>
```