**DESCRIPTION OF PIG**

## Apache Pig – History

In **2006**, Apache Pig was developed as a research project at Yahoo, especially to create and execute MapReduce jobs on every dataset. In **2007**, Apache Pig was open sourced via Apache incubator. In **2008**, the first release of Apache Pig came out. In **2010**, Apache Pig graduated as an Apache top-level project.

**PIG INSTALLATION**

**Prerequisites**

It is essential that you have Hadoop and Java installed on your system before you go for Apache Pig. Therefore, prior to installing Apache Pig, install Hadoop and Java.

**Download Apache Pig**

First of all, download the latest version of Apache Pig from the following website − https://pig.apache.org/

**Install Apache Pig**

After downloading the Apache Pig software, install it in your Linux by extracting the downloaded tar files as shown below.

$ tar zxvf pig-0.15.0-src.tar.gz

$ tar zxvf pig-0.15.0.tar.gz

**Configure Apache Pig**

After installing Apache Pig, we have to configure it. To configure, we need to edit bashrc

$ gedit .bashrc

Add these lines at the end-

export PIG\_HOME = /home/Hadoop/Pig

export PATH = PATH:/home/Hadoop/pig/bin

export PIG\_CLASSPATH = $HADOOP\_HOME/conf

**What is Apache Pig?**

Apache Pig is an abstraction over MapReduce. It is a tool/platform which is used to analyze larger sets of data representing them as data flows. Pig is generally used with Hadoop; we can perform all the data manipulation operations in Hadoop using Apache Pig.

To write data analysis programs, Pig provides a high-level language known as Pig Latin. This language provides various operators using which programmers can develop their own functions for reading, writing, and processing data.

To analyze data using Apache Pig, programmers need to write scripts using Pig Latin language. All these scripts are internally converted to Map and Reduce tasks. Apache Pig has a component known as Pig Engine that accepts the Pig Latin scripts as input and converts those scripts into MapReduce jobs.

**Different statements in PIG**

**LOAD**

Relation\_name = LOAD 'Input file path' USING function as schema;

**STORE**

STORE Relation\_name INTO ' required\_directory\_path ' [USING function];

**DUMP**

Dump Relation\_Name

**DESCRIBE**

Describe Relation\_name

**EXPLAIN**

explain Relation\_name;

**ILLUSTRATE**

illustrate Relation\_name;

**GROUP**

Group\_data = GROUP Relation\_name BY field;

**COGROUP**

cogroup\_data = COGROUP relation\_1 by field, relation\_2 by field;

**SELF-JOIN**

Relation3\_name = JOIN Relation1\_name BY key, Relation2\_name BY key ;

**INNER-JOIN**

result = JOIN relation1 BY columnname, relation2 BY columnname;

**LEFT-OUTER JOIN**

Relation3\_name = JOIN Relation1\_name BY id LEFT OUTER, Relation2\_name BY customer\_id;

**RIGHT OUTER JOIN**

outer\_right = JOIN customers BY id RIGHT, orders BY customer\_id;

**FULL OUTER JOIN**

outer\_full = JOIN customers BY id FULL OUTER, orders BY customer\_id;

**CROSS**

Relation3\_name = CROSS Relation1\_name, Relation2\_name;

**UNION**

Relation\_name3 = UNION Relation\_name1, Relation\_name2;

**SPLIT**

SPLIT Relation1\_name INTO Relation2\_name IF (condition1), Relation2\_name (condition2);

**FILTER**

Relation2\_name = FILTER Relation1\_name BY (condition);

**DISTINCT**

Relation\_name2 = DISTINCT Relatin\_name1;

**GENERATE**

Relation\_name2 = FOREACH Relatin\_name1 GENERATE (required data);

**ORDER BY**

Relation\_name2 = ORDER Relatin\_name1 BY (ASC|DESC);

**LIMIT**

Result = LIMIT Relation\_name required number of tuples;

**Apache Pig Vs MapReduce**

Listed below are the major differences between Apache Pig and MapReduce.

|  |  |
| --- | --- |
| **Apache Pig** | **MapReduce** |
| Apache Pig is a data flow language. | MapReduce is a data processing paradigm. |
| It is a high level language. | MapReduce is low level and rigid. |
| Performing a Join operation in Apache Pig is pretty simple. | It is quite difficult in MapReduce to perform a Join operation between datasets. |
| Any novice programmer with a basic knowledge of SQL can work conveniently with Apache Pig. | Exposure to Java is must to work with MapReduce. |
| Apache Pig uses multi-query approach, thereby reducing the length of the codes to a great extent. | MapReduce will require almost 20 times more the number of lines to perform the same task. |
| There is no need for compilation. On execution, every Apache Pig operator is converted internally into a MapReduce job. | MapReduce jobs have a long compilation process. |