**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.

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.

Pig works in two execution modes:

•Local mode: In this mode, Pig runs in a single JVM and accesses the local file system. Local mode is actually a local simulation of Map Reduce in Hadoop’s LocalJobRunner class.

%pig –x local

•Map Reduce mode: In this mode, Pig relies on HDFS.

%pig

A single element of data is called field.

A collection of values is called tuple.

A collection of tuples is called bag.

Command used to output – DUMP(to screen) and STORE(to file)

To extract data from columns – FOREACH and GENERATE

DISTINCT to eliminate duplicates

Complex data types- tuple, bag, map

**Word count**

1. name = load '/home/user/Desktop/name.csv' USING PigStorage(',') AS (line:chararray);
2. words = FOREACH lines GENERATE FLATTEN(TOKENIZE(line)) as word;
3. word\_group = group words by word;
4. count\_word = foreach word\_group generate group,COUNT(words);
5. DUMP count\_word