Apache pig

pig is a high level platform or a tool which is used to process Large dataset.

It is an data flow system.

also it creates mapreduce programs used in Hadoop.

It translates into efficient sequences of one or more mapreduce jobs.

It offers high level language to write data analysis program. It is mainly used for non java programmer. which is called as pig Latin.

pig uses Both HDFS and map reduce for storage and processing.

features of pig:-

* Inbuild operators:

It provide very good set of operators for sort, join, filter etc.

* Automatic optimization:-

* handles all kind of data

* ease of programming.

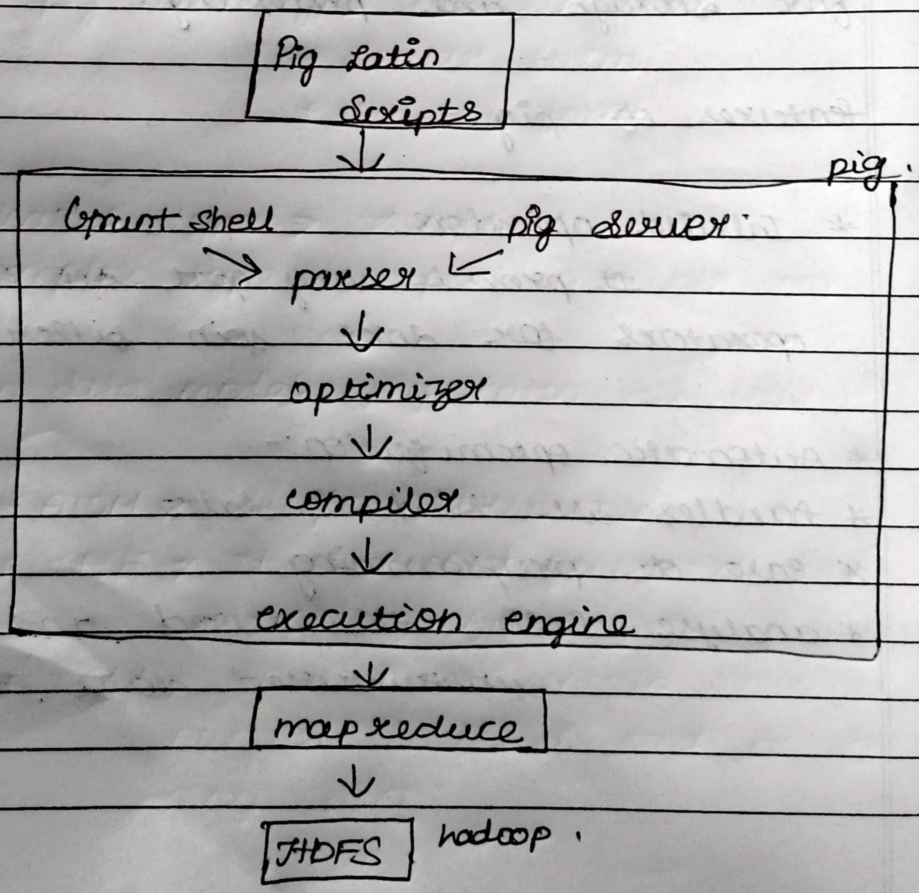
* analyse both structured and unstructured data.

Pig has two execution modes :-

* Local mode :-
To run pig in local mode, we need access to single machine. all files are installed and run using local host and file system.

* mapreduce mode :
To run pig in mapreduce mode we need access to hadoop cluster and HDFS installation. It is an default mode.

Pig architecture :-



components of Apache pig :-

1. parser : when pig latin script is sent to hadoop pig. It is first handled to parser.

parser checks the syntax of script. parser gives output in form of DAG which contain pig latin statements.

2. optimizer :-

After output from parser is retrieved, a logical plan for DAG passed to logical optimizer. It is used to carry logical optimization

3. compiler :-

The compiler comes when output from optimizer is received. The logical plan converted into series of mapreduce task or jobs.

4. Execution Engine:-

After logical plan converted into mapreduce jobs. These jobs are executed in hadoop for desired result.

Pig

↓

Load data

⇓

manipulate from HDFS.

⇓

DUMP data

⇓

Store result.

Pig data model :-

In pig , when data load the data model . any data load from disk into pig have specific schema and structure.

pig data type divided into two

* scalar form
* complex form.

scalar type ⇒ single value

complex type ⇒ tuple , container and map.

In data model, pig Laten has four types:-

* ATOM ⇒ single attribute
* TUPLE ⇒ It is record generated by series of field.
* BAG ⇒ each tuple contain arbitrary number of field and be of any sort.

* MAP → map set of pairs with main Values. It can store any type.

Pig Latin:-

pig latin statements are used to process the data.

* It can span multiple lines.
* Each statement end with semicolon.
* It include expression and schema.

Pig Latin statement work with relation A relation can be defined as.

* A relation is a bag
* A bag is a collection of tuple.
* A tuple is ordered set of ~~pair~~ field.
* A field is a piece of data.

Pig Latin data types:

* INT
* LONG
* FLOAT
* DOUBLE
* BOOLEAN.

Advantages:

- * Faster execution.
- * process Large amount of data
- * strong documentation to Learn pig Latin.

Disadvantages:

- * slow start up
- * NOT suitable for OLAP
- * complex application.