Pig

* What is Pig ?

→ Pig is a scripting lang that runs on hadoop cluster, designed to process and analyze large datasets
→ Pig operates on various types of data like structured, semi-s, unstru...

Map Reduce	Hive	Pig
→ Compiled lang	→ SQL like Query	→ Scripting lang
→ long complex code	→ no need	→ no need
→ process all types of data	→ structured	→ all
→ lower level of abstraction	→ higher	→ higher
→ supports partitioning	→ supports	→ no
→ uses JAVA, python	→ Hive QL	→ Pig Latin
→ programmers	→ data analysts	→ researchers & prog
→ code performance is good	→ less than MR & pig	→ more than Hive less than MR

* Pig Components :

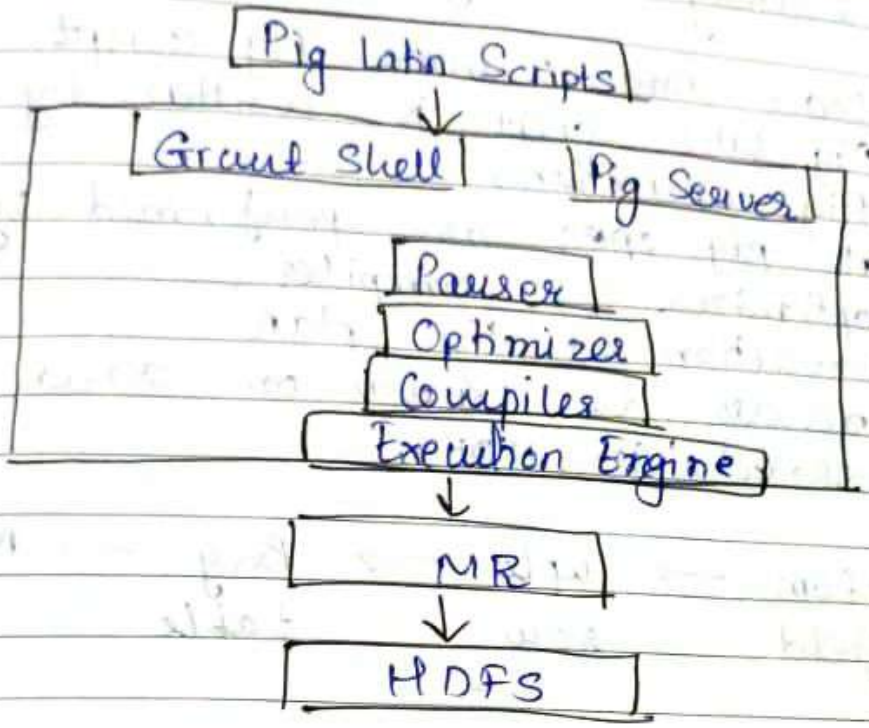
Pig Latin

→ is procedural data flow lang used in Pig to analyze data
→ easy to program
→ similar to SQL.

Runtime Engine

→ execution env. create to run Pig Latin prog
→ also a compiler to produce MR prog
→ uses HDFS for storing & retriving data.

* Pig Architecture :-



→ Programmers write a script in Pig Latin to analyze data using Pig.
→ Grunt Shell is Pig's interactive shell which is used to execute all Pig scripts
→ If the Pig script is written in a script file, the execution is done by Pig Server
→ Parser checks the syntax of Pig script output will be DAG.
→ DAG is passed to logical optimizer where optimization take place.
→ The Compiler converts the DAG into MR jobs.
→ MR jobs are executed at Execution Engine results are displayed using 'DUMP' statements and stored in HDFS using 'STORE' statement.

* Working of Pig :-

1) Load data & write Pig script
   Pig Latin script is written by users
2) Pig operations
   all pig opes are performed by parser, optimizer & compiler.
3) Execution of the plan
   results are shown on screen otherwise stored in HDFS

→ Atom → Tuple → Bag → Map
  field row table