



Recap

2

- ▶ Impala ad-hoc query processing tool
- ▶ MapReduce optimization
- ▶ MapReduce chaining

Agenda for today

3

- ▶ Pig
- ▶ Hive
- ▶ Impala



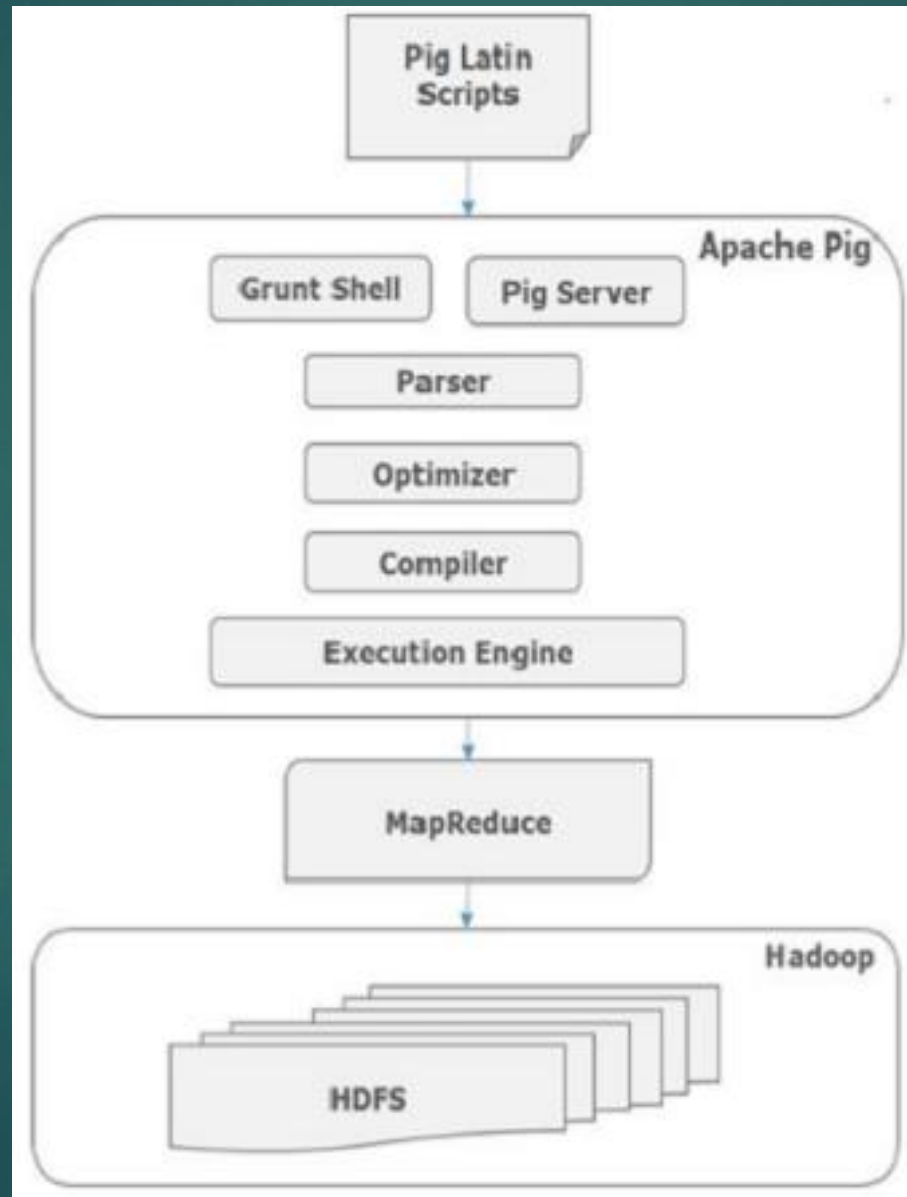
Introduction

5

- ▶ High Level Scripting Language developed by Yahoo originally
- ▶ Transforms SQL like language called Pig Latin into Java code
- ▶ Follows lazy evaluation
- ▶ Supports UDF written in multiple languages

Architecture

6



Execution

7

▶ Accessing approaches:

1. Batch mode: submit a script directly
2. Interactive mode: Grunt, the pig shell
3. PigServer for Java program

▶ Execution mode:

1. Local mode: `pig -x local`
2. Mapreduce mode: `pig -x mapreduce`

Applications

8

- ▶ Process web logs
- ▶ Build user behavior models
- ▶ Process images
- ▶ Build maps of the web
- ▶ Do research on large data sets

Data types

- ▶ Scalar Types: Int, long, float, double, boolean, null, chararray, bytearray
- ▶ Complex Types: fields, tuples, bags, relations

Operator: LOAD

10

▶ To load data from storage system
`lines=LOAD 'myfile' AS (line: chararray);`

▶ Supports various loader formats

1. PigStorage
2. TextLoader
3. BinStorage

Operator: LOAD cont...

11

- ▶ Load data without schema

```
relXYZ = LOAD 'yourfile.csv' USING PigStorage(',');
```

- ▶ Load data with schema

```
relXYZ = LOAD 'yourfile.csv' USING PigStorage(',') as  
(col1:datatype, col2:datatype,...);
```

Operator: LIMIT

12

- ▶ To take sample records

New_Rel = LIMIT RelationName <Sample Count>;

Operator: DUMP

13

- ▶ Print the data on console

DUMP RelationName;

Operator: FOREACH

14

- ▶ Select specific columns

New_Rel = FOREACH RelationName GENERATE
driverId, eventTime, eventType;

Operator: JOIN

15

- ▶ Joins two relations/datasets

```
join_data = JOIN relation1 BY (column1), relation2  
BY (column1);
```

Operator: SORT

16

- ▶ Sort a relation based on key

New_rel = ORDER RelationName BY ColumnName
asc;

Operator: FILTER

17

- ▶ Filter the dataset

New_rel = FILTER RelationName BY (Condition);

Operator: DISTINCT

18

- ▶ Remove duplicates

New_rel = DISTINCT RelationName;

Operator: STORE

19

- ▶ Store the output

```
STORE relationName INTO 'output_directory' USING  
PigStorage(',');
```

UDF

20

- ▶ Prepare a Jar file
- ▶ Register the Jar
- ▶ Define alias
- ▶ Use it

https://www.tutorialspoint.com/apache_pig/apache_pig_user_defined_functions.htm