Lesson 7

PIG characteristics and Functions

Apache® PIG

- An abstraction over MapReduce program
- Performs the data manipulation in files at data nodes in Hadoop
- An execution framework for Big Data parallel processing in HDFS environment

PIG execution engine

- Internally compiles and runs codes as MapReduce codes
- Reduces the length of codes using multi-query approach
- Pig code of 10 lines equal to MapReduce code of 200 lines

PIG Features

- A high-level data-flow language
- A PIG operation node takes the inputs and generates output for the next node─ similar to MapReduce.
- PIG makes writing the codes simpler compared MapReduce

…PIG Features

- Processes any kind of data, structured, semi-structured or unstructured data, coming from various sources
- Handles inconsistent schema in case of unstructured data

… PIG Features

- PIG performs automatic optimization of tasks before execution
- Programmer need to concentrate upon the whole operation with no need of creating mapper and reducer tasks separately

… PIG Features

- Read the input data files from HDFS or the data files from other sources such as local file system, stores the intermediate data and writes back the output in HDFS

… PIG Features

- Programmers writes complex data transformations using scripts (without using Java)

Applications of Pig

- ETL (Extracts the data, Transforms on operations on that data and Loads (dumps) the data in the required format in HDFS
- Processing time sensitive data loads

Applications of Pig

- Analyzing large data-sets,
- Executing tasks involving ad-hoc processing
- Processing large data sources such as web logs and streaming online data
- Data processing for search platforms

Optimization in Map-Reduce Task

- Section of codes need to run after mapping and before reducing to output result.
- Those codes can run at Mapper or Reducer.
- Optimization means distributing codes among Mapper and Reducer

PIG interactive shell

- Grunt shell to write Pig Latin codes
- Scripts using Pig Latin analyze data

PG Latin

- A language very similar to SQL
- Possess a rich set of built-in operators such as Group, Join, Filter, Limit, Order by, Parallel, Sort and Split.

PIG Latin Scripts

- Internally converts to Map and Reduce tasks with the help of the component known as Pig Engine
- Accepts the Pig Latin scripts as input and converts those scripts into MapReduce jobs.

PIG ‘User defined functions’ (UDFs)

- Custom functions, not available in Pig
- DUFF can be in other programming languages such as, Java, Python, Ruby, Python, JRuby
- DUFF codes easily embed into Pig scripts

Pig Characteristics

- Read data,
- Processing,
- Programming the UDFs in multiple languages,
- Programming multiple queries by fewer codes. This causes fast processing.

Pig Characteristics

- Fast processing
- Pig derives guidance from four philosophies, live anywhere, take anything, domestic and run as if flying.

Differences between PIG ad MapReduce

- Table 4.14
- Data flow high level language
- Easy programming being SQL like support, Join, Sort, Order
- Scripting using Grunt Shell and PIG Latin

Differences between PIG ad MapReduce

- Multi-query approach
- Nested data types.
- Tuple, Bag, Map

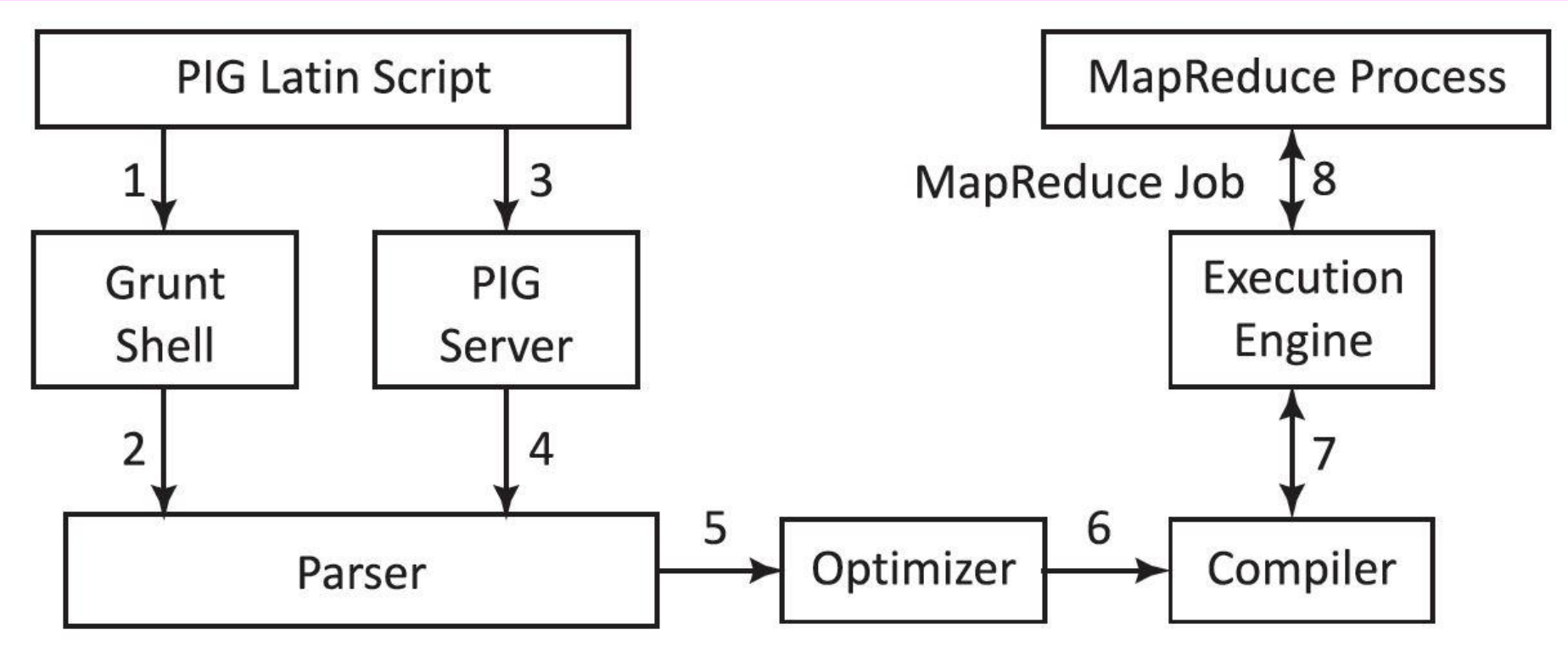
Differences between PIG and SQL

- Table 4.15
- Procedure oriented instead of declarative
- Nested relational model in place of Flat relational mode
- Schema optional– can store without schema unlike SQL
- Limited Query Optimization

Differences between Pig and Hive

- Table 4.16
- Hive mostly used for structured data, HiveQL declarative like SQL, query processing language instead of data-flow

Figure 4.12 Pig architecture for scripts data-flow and processing



Pig Architecture

- Three ways to execute scripts are:
- Grunt Shell commands, Use Script file and Embed scripts

Three Ways

1. **Grunt Shell:** An interactive shell of Pig, Shell executes the scripts (Section 4.6.1 Apache Pig - Grunt Shell)
2. **Script File:** Pig commands written in a script file which executes at PIG Server.

Three ways

3. Embedded Script: Create UDFs for the functions unavailable in Pig built-in operators

DUFF can be in other programming languages.

- The UDFs can embed in Pig Latin Script file

Installing Pig

- Refer Section 4.6.2

Summary

We learnt PIG:

- Performs the data manipulation in files at data nodes in Hadoop
- An execution framework for Big Data parallel processing in HDFS environment

Summary

We learnt :

- Processes Structured, unstructured, schema-less data,
- Data types: Tuple, Bag and Map
- Grunt Shell
- Embed Scripts: UDFs
-

Summary

We learnt PIG:

- SQL like possess a rich set of built-in operators such as Group, Join, Filter, Limit, Order by, Parallel, Sort and Split

End of Lesson 7 on PIG Characteristics and Functions