Big Data Hadoop and Spark Developer

Pig: Data Analysis Tool

Learning Objectives

By the end of this lesson, you will be able to:

- Define Pig and explain its execution flow
- Describe why Apache Pig is needed
- Recognize how Pig complements Hadoop
- Compare Pig operation with MapReduce

Introduction to Pig

What Is Pig?

- Apache Pig is a high-level programming language that is used for analyzing huge datasets.
- Apache Pig was created by Yahoo Research and is used along with Hadoop to perform a variety of data management tasks.
- Apache Pig uses the programming language *Pig Latin* to write data analysis programs.

Need for Pig

Pig Latin can be extended by writing UDFs (user-defined functions) in Python or JavaScript and then calling them directly from the language.

Pig Latin allows programmers to run MapReduce tasks without having to type complex Java codes.

Apache Pig Latin uses a multi-query approach that can reduce the code's length.

For example, a process that would require 100 lines of code in Java may be completed with 5 lines of code with Apache Pig Latin, which reduces development time.

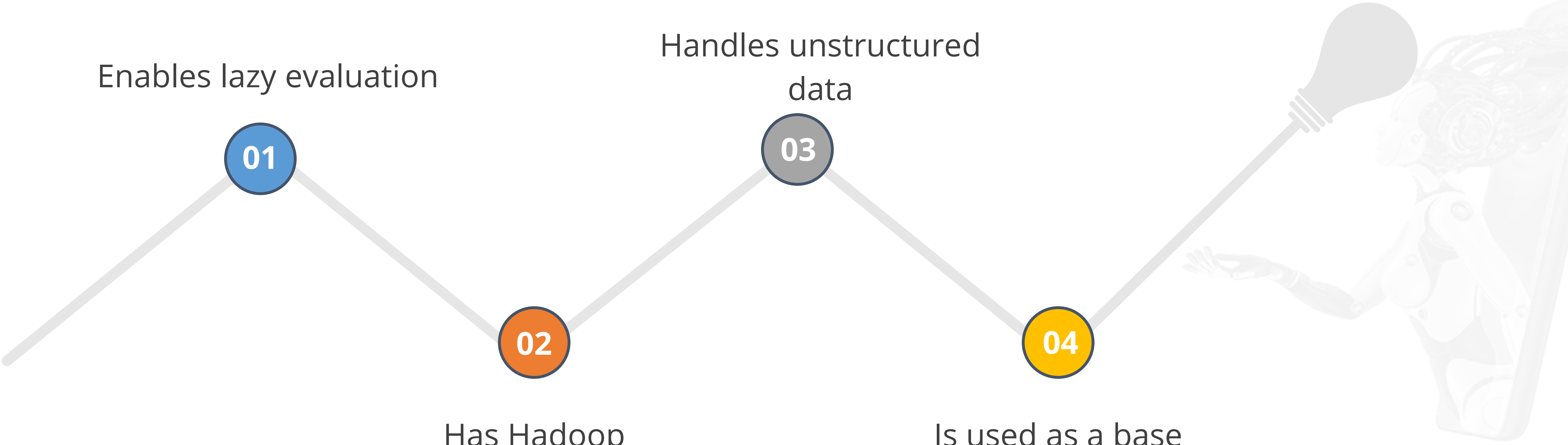
How Pig Complements Hadoop

Many built-in operators in Apache Pig provide data operations, such as join, filter, and ordering. It also includes nested data types, such as tuples, bags, and maps, which are not present in MapReduce.

Pig Latin is a high-level data flow language, whereas MapReduce is a low-level data processing paradigm.

Programmers can quickly achieve the same results using Pig Latin instead of implementing complex Java code in MapReduce.

Advantages of Pig



01 Enables lazy evaluation

02 Has Hadoop characteristics

03 Handles unstructured data

04 Is used as a base pipeline

Advantages of Pig

1. **Enables lazy evaluation**

   Pig can be used to optimize the program from start to finish. The data is processed when *STORE* and *DUMP* commands are encountered.

2. **Has Hadoop characteristics**

   Pig is commonly used with Hadoop and can perform Hadoop's data manipulation tasks.

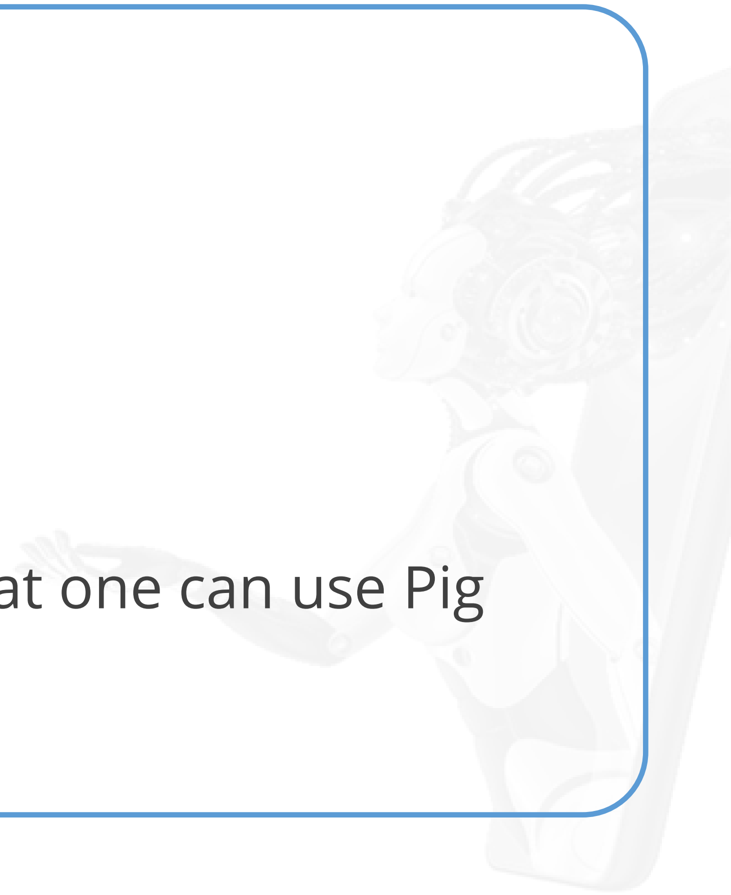
Advantages of Pig

3. **Handles unstructured data**

   Pig Latin can convert huge amounts of unstructured data into structured data.

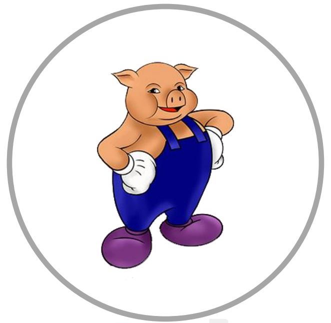
4. **Is used as a base pipeline**

   For huge amounts of data, Pig has UDFs (user-defined functions). This implies that one can use Pig as a base pipeline to handle the heavy work.



Components of Pig

Components of Pig



The Pig has two major components, which are listed below.

A runtime engine

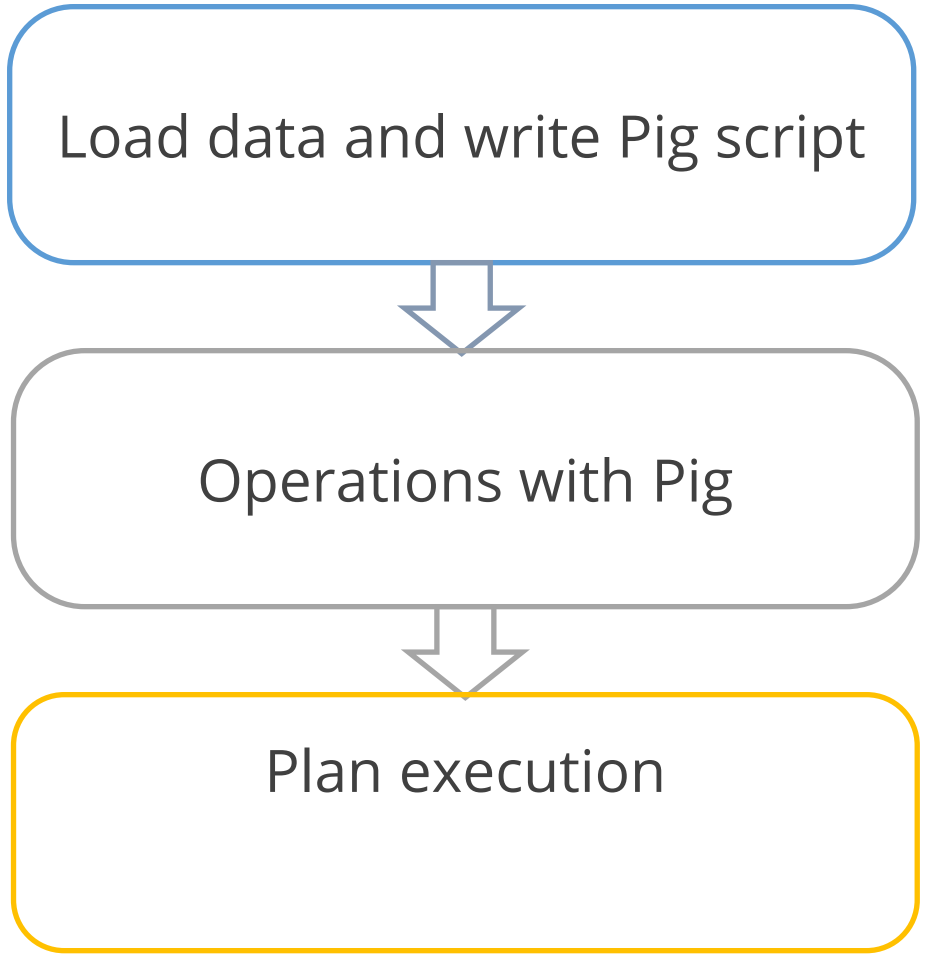
The runtime engine is a compiler that generates MapReduce program sequences. It stores and retrieves data using HDFS. It is also used to communicate with the Hadoop platform (HDFS and MapReduce).

Pig Latin script language

Pig Latin is a procedural data flow programming language. It includes syntax and instructions that can be used to create business logic.

The Stages of Pig Operations and Execution Flow

The operation of a Pig can be divided into the following stages:



Load data and write Pig script

Operations with Pig

Plan execution

The Stages of Pig Operations and Execution Flow

Step 1: Load data and write Pig script.

Example

```
student = LOAD '/user/testdemomay1301mailinator
/student_details.txt' USING PigStorage(',')
AS(id:int,firstname:chararray,lastname:chararray,
phone:chararray,city:chararray);
student_order = ORDER student BY city DESC;
student_limit = LIMIT student_order 4;
Dump student_limit;
Output:
(4,Isla,Ale,29,98872233)
(3,Anna,Mathew,28,98802239)
(2,Calvin,Joseph,25,98802238)
(1,John,edy,27,9848022337)
```

Steps to perform:

- Open the pig shell in "**Webconsole**" by typing the given commands.

Command:

- Pig
- Type an expression or a statement and click on "enter".
- Every expression and statement that is typed is evaluated and executed immediately.

The Stages of Pig Operations and Execution Flow

Steps to perform:

Step 1: Log in to the “Webconsole” and run the below command to enter pig shell mode.

```
pig
```

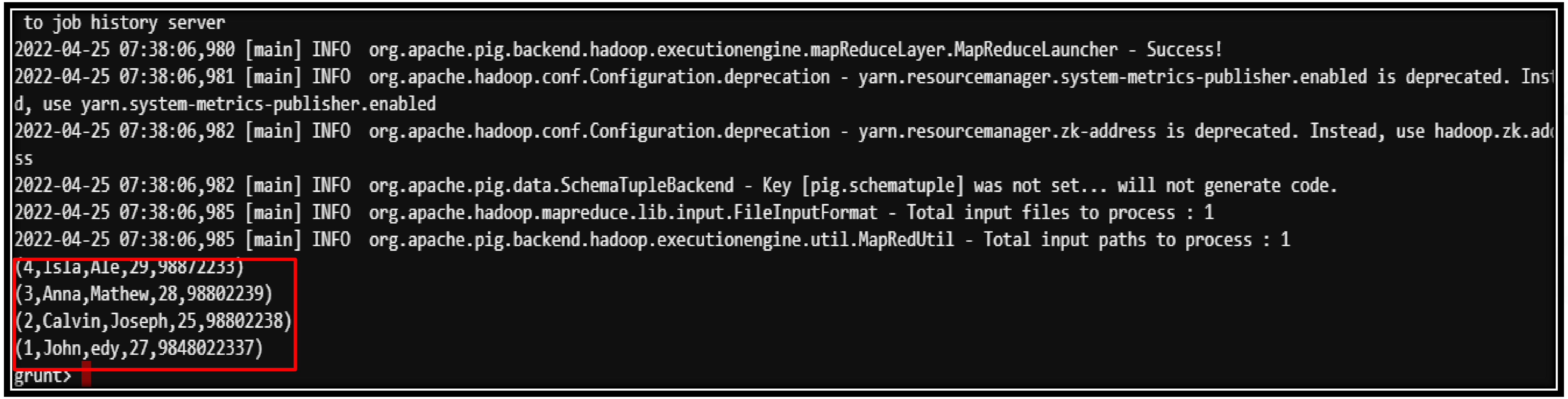
Step 2: Create a student_details text file on the desktop including (ID, First name, Last name, phone, city) information inside it and upload it inside HDFS in “Hue.”

Step 3: In the “Weconsole” type the commands are given below to get the maximum of 4 student details output except for the city.

```
Student = LOAD '/user/testdemomay1301mailinator/student_details.txt' USING PigStorage(‘,’)
        AS(id:int,firstname:chararray,lastname:chararray,phone:chararray,
        city:chararray);
        student_order = ORDER student BY city DESC;
        student_limit = LIMIT student_order 4;
        Dump student_limit;
```

The Stages of Pig Operations and Execution Flow

Output:



```
 to job history server
2022-04-25 07:38:06,980 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2022-04-25 07:38:06,981 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Inst
d, use yarn.system-metrics-publisher.enabled
2022-04-25 07:38:06,982 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - yarn.resourcemanager.zk-address is deprecated. Instead, use hadoop.zk.ad
ss
2022-04-25 07:38:06,982 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2022-04-25 07:38:06,985 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2022-04-25 07:38:06,985 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(4,Isla,Ale,29,98872233)
(3,Anna,Mathew,28,98802239)
(2,Calvin,Joseph,25,98802238)
(1,John,edy,27,9848022337)
grunt>
```

The Stages of Pig Operations and Execution Flow

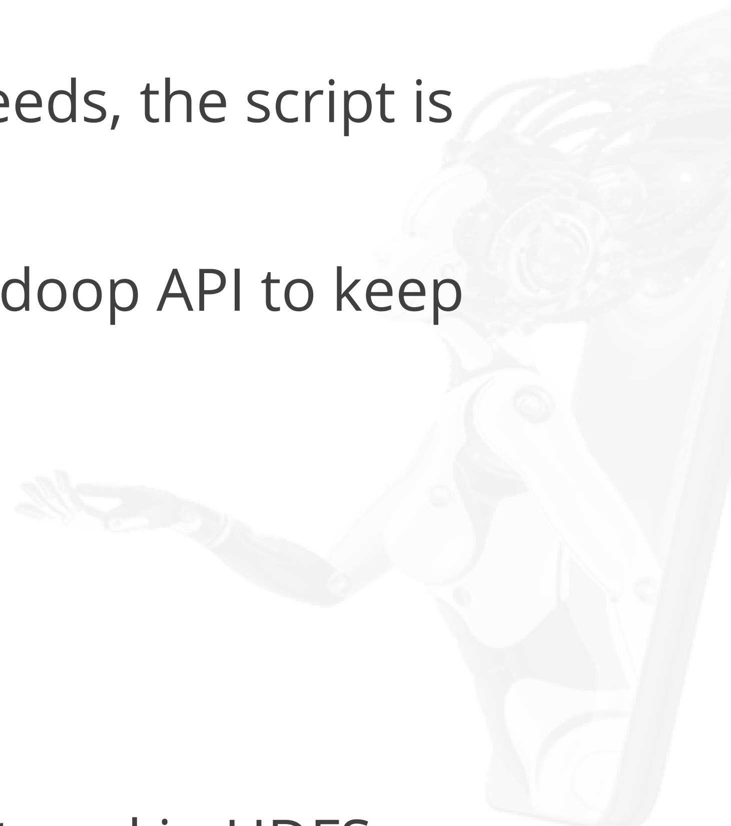
Step 2: Operations with Pig.

The Pig execution engine parses and checks the script in the second step. If it succeeds, the script is optimized, and a logical and physical execution plan is generated.

The job is defined as a MapReduce Task and submitted to Hadoop. Pig uses the Hadoop API to keep track of the job's progress and reports back to the client.

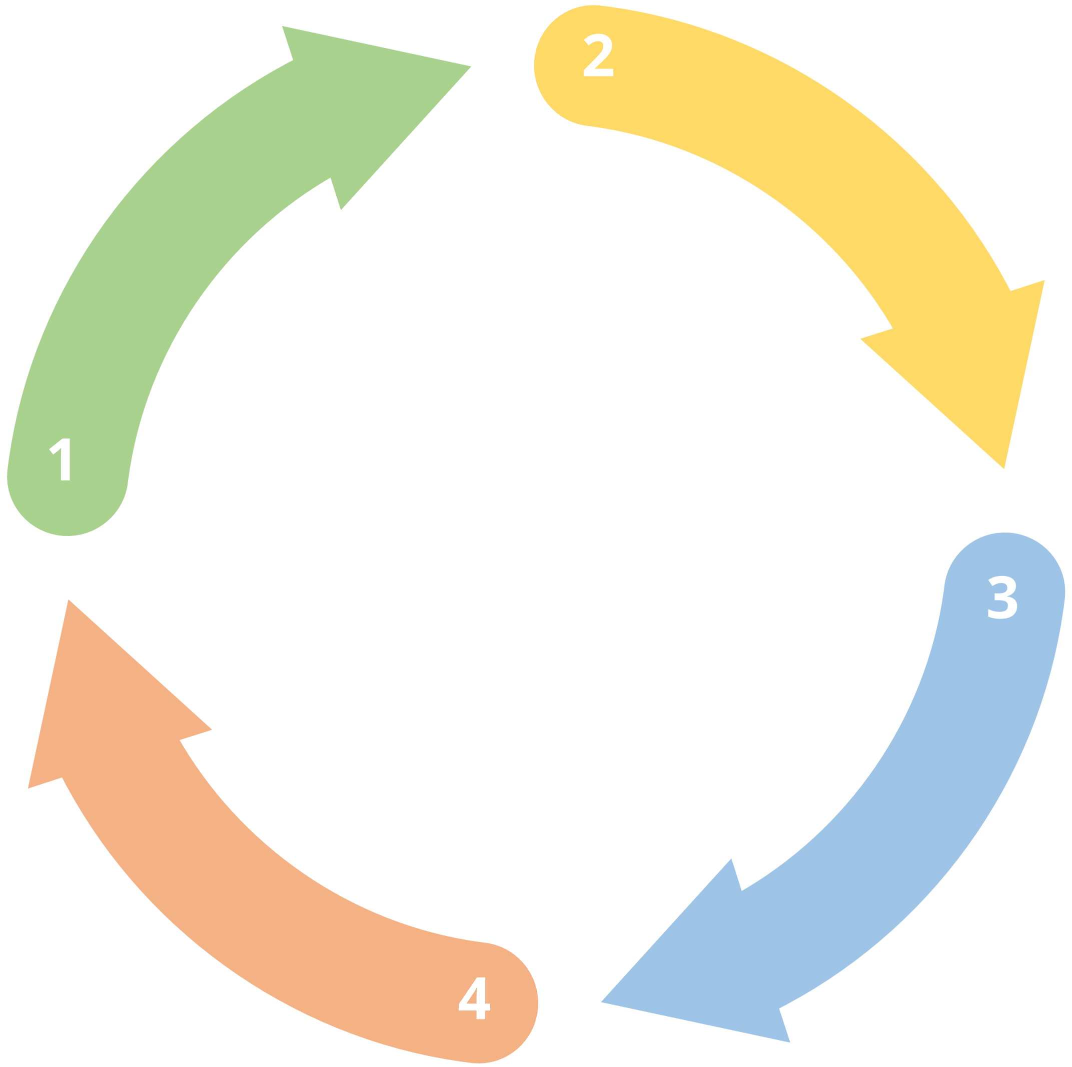
Step 3: Plan execution.

Depending on the user instruction, the results are either spilled on the section or stored in HDFS at the final stage.



Script Interpretation

Pig processes Pig Latin statements in the following manner:

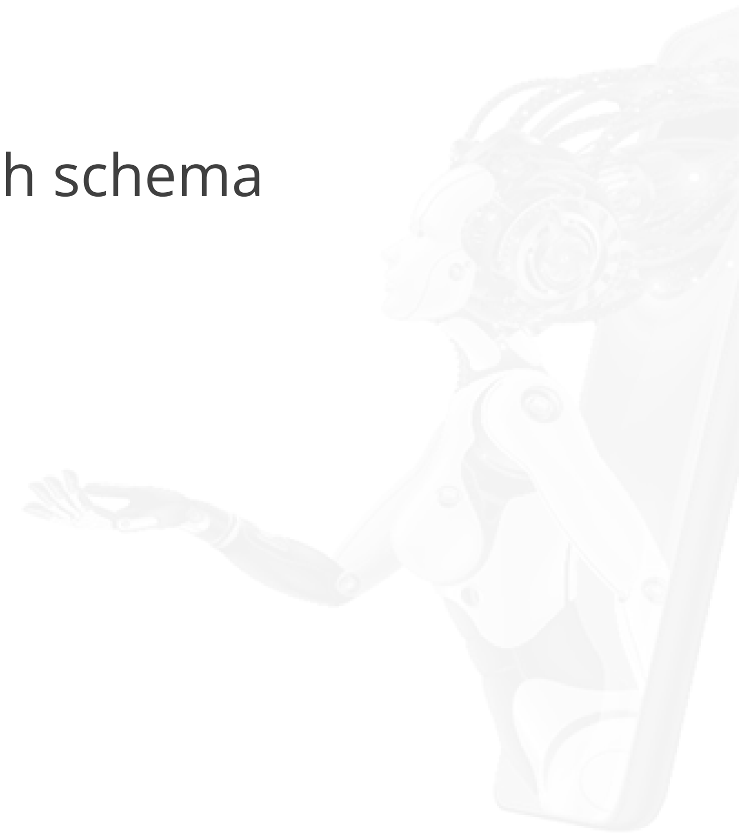


Validate all statements' syntax and semantics.

Type check with schema

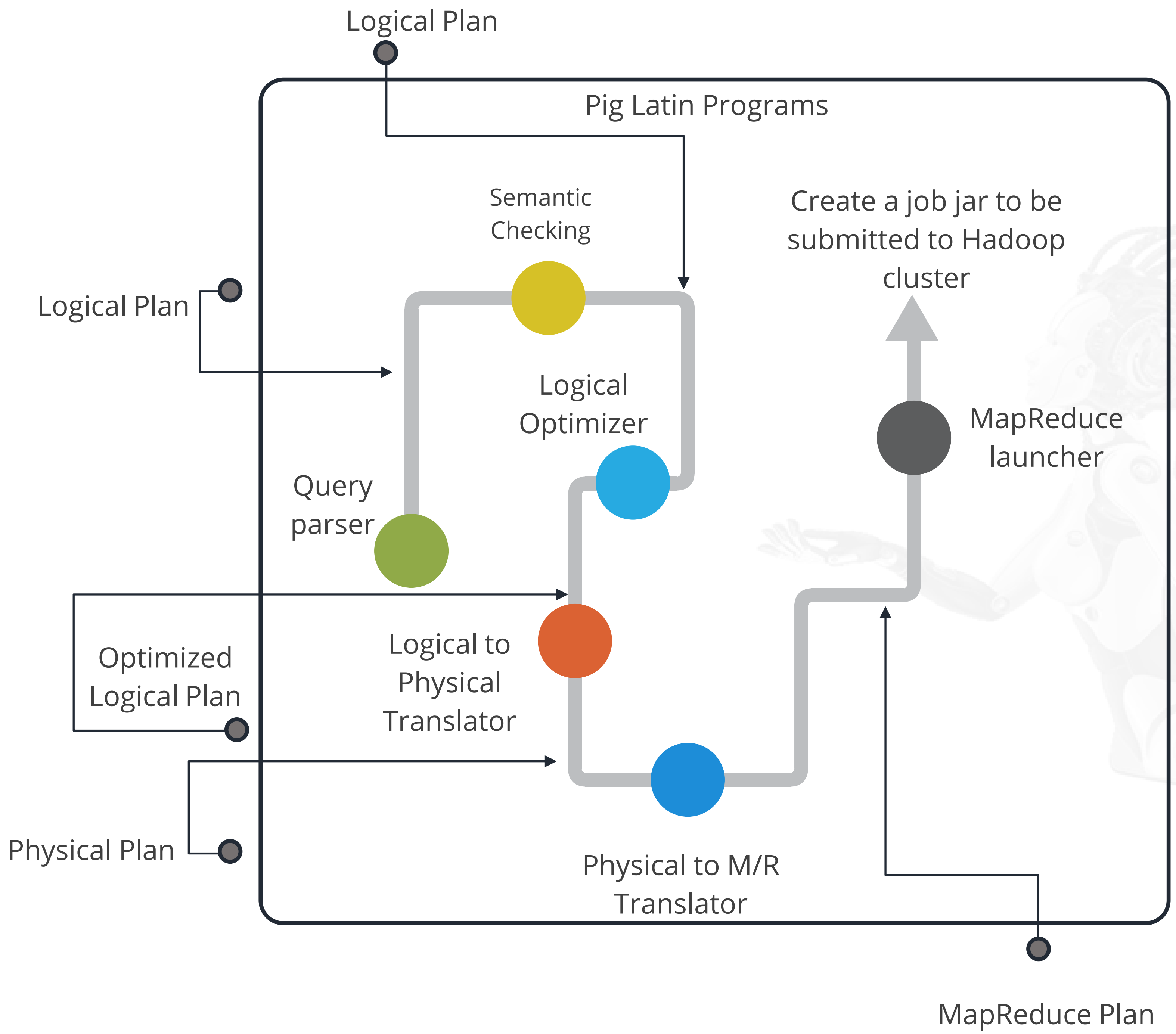
Check your references.

Execute the statement



Script Interpretation

As shown in the diagram, a Pig Latin script execution plan consists of logical, optimized logical, physical, and MapReduce plans.



Salient Features of Pig

Pig is popular among developers and analysts because it has many features.

The following are some of the features:

- Pig can operate directly over files and provide step-by-step procedural control.
- In Pig, schemas will be dynamically assigned, even if they are optional.
- User-defined functions (UDFs) and a variety of data formats are supported in Pig.

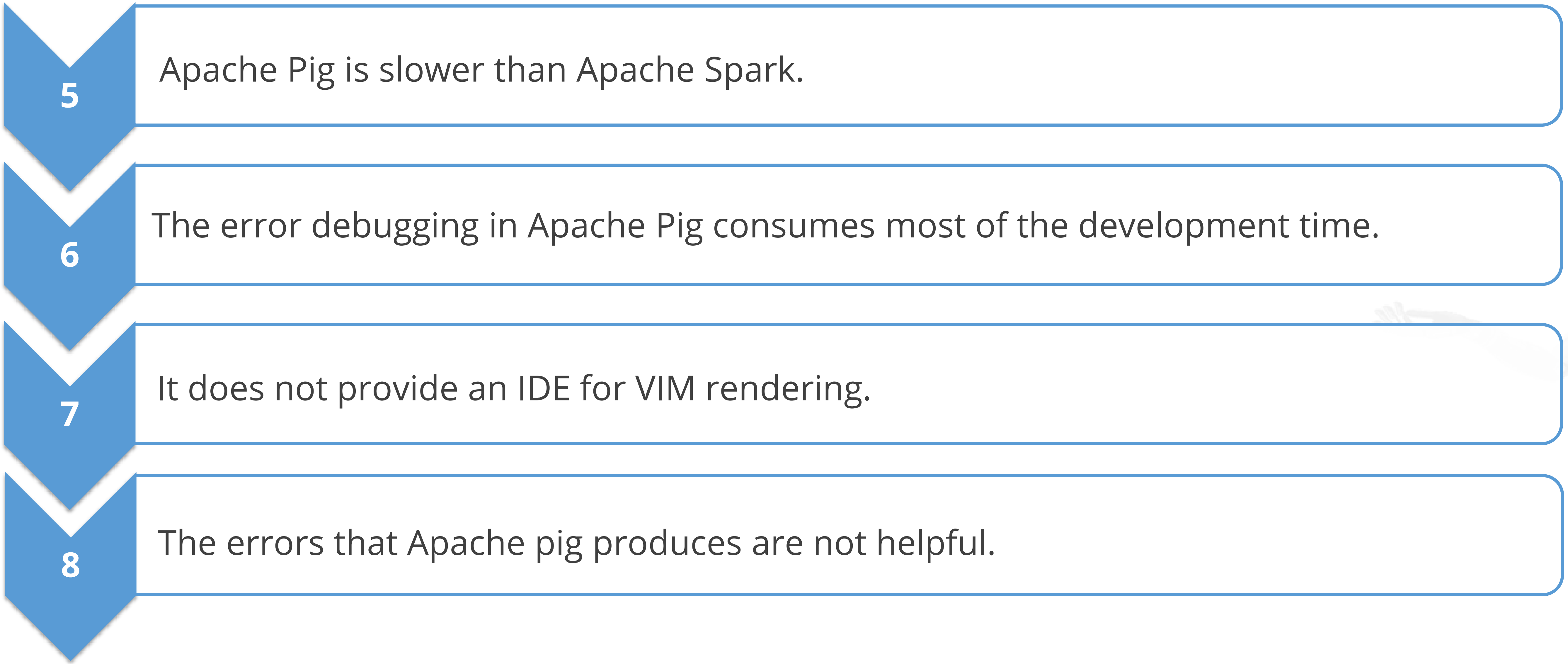
Disadvantages of Pig

The limitations of Apache Pig are mentioned below:

1. The commands do not get executed until we dump or store an intermediate or result.
2. The number of iterations between debugging and fixing the problem can be increased by dumping the result.
3. The Data Schema is enforced implicitly, not explicitly.
4. Apache Pig is designed for ETL processing; hence it is not suited for real-time processing.

Disadvantages of Pig

The limitations of Apache Pig are mentioned below:



5. Apache Pig is slower than Apache Spark.
6. The error debugging in Apache Pig consumes most of the development time.
7. It does not provide an IDE for VIM rendering.
8. The errors that Apache pig produces are not helpful.

Key Takeaways

- Apache Pig is a high-level programming language used for analyzing huge datasets.
- Apache Pig Latin uses a multi-query approach that can reduce the code length.
- Many built-in operators in Apache Pig provide data operations, such as join, filter, and ordering.
- Pig's Latin data model enables Pig to handle any type of data.

Key Takeaways

- Pig is a scripting language interacts with HDFS while SQL is a query language used to interact with databases.
- Pig is a data flow language, but MapReduce is a data processing language.
- Loading refers to the process of loading relations from files into the Pig buffer.
- Storing refers to writing outputs to the file system.

Knowledge Check

Knowledge Check

1

Which of the following commands start Pig in MapReduce mode?

A. Pig

B. Pig -x MapReduce

C. Pig -x local

D. Both Pig and Pig -x MapReduce

Knowledge Check

1

Which of the following commands start Pig in MapReduce mode?

A. Pig

B. Pig -x MapReduce

C. Pig -x local

D. Both Pig and Pig -x MapReduce

The correct answer is **D**

Pig and Pig -x MapReduce commands start Pig in MapReduce mode.

Knowledge Check

2

Which of the following commands start Pig in local mode?

A. Pig -x local

B. Pig -x MapReduce

C. Pig

D. Both b and c

Knowledge Check

2

Which of the following commands start Pig in local mode?

A. Pig -x local

B. Pig -x MapReduce

C. Pig

D. Both b and c

The correct answer is **A**

Pig -x local command starts Pig in local mode.

Knowledge Check

3

In how many ways can the Pig Latin program be written?

A. 2

B. 4

C. 3

D. 5

Knowledge Check

3

In how many ways can the Pig Latin program be written?

A. 2

B. 4

C. 3

D. 5

The correct answer is **A**

Pig Latin program can be written in two ways.

Thank You