PIG INTERVIEW QUESTIONS

Important Links:-

<https://data-flair.training/blogs/apache-pig-interview-questions/>

<https://www.dezyre.com/article/pig-interview-questions-and-answers-for-2018/244>

<https://www.edureka.co/blog/interview-questions/hadoop-interview-questions-pig/>

<https://intellipaat.com/blog/interview-question/pig-interview-questions/>

<https://mindmajix.com/apache-pig-interview-questions>

<https://www.bigdatatrunk.com/top-50-pig-interview-questions/>

<https://www.educba.com/pig-interview-questions/>

<https://www.wisdomjobs.com/e-university/apache-pig-interview-questions.html>

1. **Compare Apache Pig And Sql?**

**Answer :**

Apache Pig differs from SQL in its usage for ETL, lazy evaluation, store data at any given point of time in the pipeline, support for pipeline splits and explicit declaration of execution plans. SQL is oriented around queries which produce a single result. SQL has no in-built mechanism for splitting a data processing stream and applying different operators to each sub-stream.

Apache Pig allows user code to be included at any point in the pipeline whereas if SQL where to be used data needs to be imported to the database first and then the process of cleaning and transformation begins.

1. **Question 2. Explain The Need For Mapreduce While Programming In Apache Pig.?**

**Answer :**

Apache Pig programs are written in a query language known as Pig Latin that is similar to the SQL query language. To execute the query, there is a need for an execution engine. The Pig engine converts the queries into MapReduce jobs and thus MapReduce acts as the execution engine and is needed to run the programs.

1. **. Explain About The Bloommapfile.?**

**Answer :**

BloomMapFile is a class, that extends the MapFile class. It is used in HBase table format to provide quick membership test for the keys using dynamic bloom filters.

1. **Question 4. What Do You Mean By A Bag In Pig?**

**Answer :**

Collection of tuples is referred as a bag in Apache Pig.

**What Is The Usage Of Foreach Operation In Pig Scripts?**

**Answer :**

FOREACH operation in Apache Pig is used to apply transformation to each element in the data bag, so that respective action is performed to generate new data items.

**Syntax-** FOREACH data\_bagname GENERATE exp1, exp2.

**Explain About The Different Complex Data Types In Pig.?**

**Answer :**

**Apache Pig supports 3 complex data types:**

**Maps-**These are key, value stores joined together using #.

**Tuples-** Just similar to the row in a table, where different items are separated by a comma. Tuples can have multiple attributes.

**Bags-** Unordered collection of tuples. Bag allows multiple duplicate tuples.

**What Does Flatten Do In Pig?**

**Answer :**

Sometimes there is data in a tuple or a bag and if we want to remove the level of nesting from that data, then Flatten modifier in Pig can be used. Flatten un-nests bags and tuples. For tuples, the Flatten operator will substitute the fields of a tuple in place of a tuple, whereas un-nesting bags is a little complex because it requires creating new tuples.

1. **How Do Users Interact With The Shell In Apache Pig?**

**Answer :**

Using Grunt i.e. Apache Pig’s interactive shell, users can interact with HDFS or the local file system.

**To start Grunt, users should invoke Apache Pig with no command:**

Executing the command “pig –x local” will result in the prompt -

grunt >

This is where PigLatin scripts can be run either in local mode or in cluster mode by setting the configuration in PIG\_CLASSPATH.

To exit from grunt shell, press CTRL+D or just type exit.

1. **Question 9. What Are The Debugging Tools Used For Apache Pig Scripts?**

**Answer :**

describe and explain are the important debugging utilities in Apache Pig.

explain utility is helpful for Hadoop developers, when trying to debug error or optimize PigLatin scripts. explain can be applied on a particular alias in the script or it can be applied to the entire script in the grunt interactive shell. explain utility produces several graphs in text format which can be printed to a file.

describe debugging utility is helpful to developers when writing Pig scripts as it shows the schema of a relation in the script. For beginners who are trying to learn Apache Pig can use the describe utility to understand how each operator makes alterations to data. A pig script can have multiple describes.

1. **What Is Illustrate Used For In Apache Pig?**

**Answer :**

Executing pig scripts on large data sets, usually takes a long time. To tackle this, developers run pig scripts on sample data but there is possibility that the sample data selected, might not execute your pig script properly.

For instance, if the script has a join operator there should be at least a few records in the sample data that have the same key, otherwise the join operation will not return any results. To tackle these kind of issues, illustrate is used. illustrate takes a sample from the data and whenever it comes across operators like join or filter that remove data, it ensures that only some records pass through and some do not, by making modifications to the records such that they meet the condition. illustrate just shows the output of each stage but does not run any MapReduce task.

[Apache Solr Tutorial](https://www.wisdomjobs.com/e-university/apache-solr-tutorial-1297.html)

1. **Question 11. Explain About The Execution Plans Of A Pig Script?<br> Or<br> Differentiate Between The Logical And Physical Plan Of An Apache Pig Script?**

**Answer :**

Logical and Physical plans are created during the execution of a pig script. Pig scripts are based on interpreter checking. Logical plan is produced after semantic checking and basic parsing and no data processing takes place during the creation of a logical plan. For each line in the Pig script, syntax check is performed for operators and a logical plan is created. Whenever an error is encountered within the script, an exception is thrown and the program execution ends, else for each statement in the script has its own logical plan.

A logical plan contains collection of operators in the script but does not contain the edges between the operators.

After the logical plan is generated, the script execution moves to the physical plan where there is a description about the physical operators, Apache Pig will use, to execute the Pig script. A physical plan is more or less like a series of MapReduce jobs but then the plan does not have any reference on how it will be executed in MapReduce. During the creation of physical plan, cogroup logical operator is converted into 3 physical operators namely –Local Rearrange, Global Rearrange and Package. Load and store functions usually get resolved in the physical plan.

[Apache Storm Interview Questions](https://www.wisdomjobs.com/e-university/apache-storm-interview-questions.html)

1. **Question 12. What Do You Know About The Case Sensitivity Of Apache Pig?**

**Answer :**

It is difficult to say whether Apache Pig is case sensitive or case insensitive. For instance, user defined functions, relations and field names in pig are case sensitive i.e. the function  COUNT is not the same as function count or X=load ‘foo’ is not same as x=load ‘foo’. On the other hand, keywords in Apache Pig are case insensitive i.e. LOAD is same as load.

[Apache Tapestry Interview Questions](https://www.wisdomjobs.com/e-university/apache-hbase-practice-tests-1246-327842)

1. **Question 13. What Are Some Of The Apache Pig Use Cases You Can Think Of?**

**Answer :**

Apache Pig big data tools, is used in particular for iterative processing, research on raw data and for traditional ETL data pipelines. As Pig can operate in circumstances where the schema is not known, inconsistent or incomplete- it is widely used by researchers who want to make use of the data before it is cleaned and loaded into the data warehouse.

To build behavior prediction models, for instance, it can be used by a website to track the response of the visitors to various types of ads, images, articles, etc.

[Apache Storm Tutorial](https://www.wisdomjobs.com/e-university/apache-storm-tutorial-1298.html)

1. **Question 14. Differentiate Between Piglatin And Hiveql?**

**Answer :**

* + It is necessary to specify the schema in HiveQL, whereas it is optional in PigLatin.
  + HiveQL is a declarative language, whereas PigLatin is procedural.
  + HiveQL follows a flat relational data model, whereas PigLatin has nested relational data model.

1. **Question 15. Is Piglatin A Strongly Typed Language? If Yes, Then How Did You Come To The Conclusion?**

**Answer :**

In a strongly typed language, the user has to declare the type of all variables upfront. In Apache Pig, when you describe the schema of the data, it expects the data to come in the same format you mentioned.

However, when the schema is not known, the script will adapt to actually data types at runtime. So, it can be said that PigLatin is strongly typed in most cases but in rare cases it is gently typed, i.e. it continues to work with data that does not live up to its expectations.

[Apache Hive Interview Questions](https://www.wisdomjobs.com/e-university/apache-hive-interview-questions.html)

1. **Question 16. What Do You Understand By An Inner Bag And Outer Bag In Pig?**

**Answer :**

A relation inside a bag is referred to as inner bag and outer bag is just a relation in Pig.

[Apache Hive Tutorial](https://www.wisdomjobs.com/e-university/apache-hive-tutorial-1308.html)

1. **Question 17. Differentiate Between Group And Cogroup Operators.?**

**Answer :**

Both GROUP and COGROUP operators are identical and can work with one or more relations. GROUP operator is generally used to group the data in a single relation for better readability, whereas COGROUP can be used to group the data in 2 or more relations. COGROUP is more like a combination of GROUP and JOIN, i.e., it groups the tables based on a column and then joins them on the grouped columns. It is possible to cogroup up to 127 relations at a time.

[Apache Flume Interview Questions](https://www.wisdomjobs.com/e-university/apache-flume-interview-questions.html)

1. **Question 18. Explain The Difference Between Count\_star And Count Functions In Apache Pig?**

**Answer :**

COUNT function does not include the NULL value when counting the number of elements in a bag, whereas COUNT\_STAR (0 function includes NULL values while counting.

[Apache Cassandra Interview Questions](https://www.wisdomjobs.com/e-university/apache-cassandra-interview-questions.html)

1. **Question 19. What Are The Various Diagnostic Operators Available In Apache Pig?**

**Answer :**

* + **Dump Operator-** It is used to display the output of pig Latin statements on the screen, so that developers can debug the code.
  + **Describe Operator-**Explained in apache pig interview question no- 10
  + **Explain Operator-**Explained in apache pig interview question no -10
  + **Illustrate Operator-** Explained in apache pig interview question no -11

[Apache Pig Tutorial](https://www.wisdomjobs.com/e-university/apache-pig-tutorial-1327.html)

1. **Question 20. How Will You Merge The Contents Of Two Or More Relations And Divide A Single Relation Into Two Or More Relations?**

**Answer :**

This can be accomplished using the UNION and SPLIT operators.

[Apache Kafka Interview Questions](https://www.wisdomjobs.com/e-university/apache-kafka-interview-questions.html)

1. **Question 21. I Have A Relation R. How Can I Get The Top 10 Tuples From The Relation R.?**

**Answer :**

TOP () function returns the top N tuples from a bag of tuples or a relation. N is passed as a parameter to the function top () along with the column whose values are to be compared and the relation R.

1. **Question 22. What Are The Commonalities Between Pig And Hive?**

**Answer :**

* + HiveQL and PigLatin both convert the commands into MapReduce jobs.
  + They cannot be used for OLAP transactions as it is difficult to execute low latency queries.

[Apache Flume Tutorial](https://www.wisdomjobs.com/e-university/apache-flume-tutorial-1341.html)

1. **Question 23. What Are The Different Types Of Udf’s In Java Supported By Apache Pig?**

**Answer :**

Algebraic, Eval and Filter functions are the various types of UDF’s supported in Pig.

[Apache Ant Interview Questions](https://www.wisdomjobs.com/e-university/apache-ant-interview-questions.html)

1. **Question 24. You Have A File Employee.txt In The Hdfs Directory With 100 Records. You Want To See Only The First 10 Records From The Employee.txt File. How Will You Do This?**

**Answer :**

The first step would be to load the file employee.txt into with the relation name as Employee.

The first 10 records of the employee data can be obtained using the limit operator -

Result= limit employee 10.

[Apache Spark Interview Questions](https://www.wisdomjobs.com/e-university/apache-spark-interview-questions.html)

1. **Question 25. Explain About The Scalar Datatypes In Apache Pig.?**

**Answer :**

integer, float, double, long, bytearray and char array are the available scalar datatypes in Apache Pig.

[Apache Kafka Tutorial](https://www.wisdomjobs.com/e-university/apache-kafka-tutorial-1342.html)

1. **Question 26. How Do Users Interact With Hdfs In Apache Pig?**

**Answer :**

Using the grunt shell.

[Apache Camel Interview Questions](https://www.wisdomjobs.com/e-university/apache-camel-interview-questions.html)

1. **Question 27. What Is The Use Of Having Filters In Apache Pig?**

**Answer :**

Just like the where clause in SQL, Apache Pig has filters to extract records based on a given condition or predicate. The record is passed down the pipeline if the predicate or the condition turn to true. Predicate contains various operators like ==, <=,!=, >=.

**Example:-**

X= load ‘inputs’ as(name,address)

Y = filter X by symbol matches ‘Mr.\*’;

[Apache Solr Interview Questions](https://www.wisdomjobs.com/e-university/apache-solr-interview-questions.html)

1. **Question 28. What Is A Udf In Pig?**

**Answer :**

If the in-built operators do not provide some functions then programmers can implement those functionalities by writing user defined functions using other programming languages like Java, Python, Ruby, etc. These User Defined Functions (UDF’s) can then be embedded into a Pig Latin Script.

[Apache Ant Tutorial](https://www.wisdomjobs.com/e-university/apache-ant-tutorial-1398.html)

1. **Question 29. Can You Join Multiple Fields In Apache Pig Scripts?**

**Answer :**

Yes, it is possible to join multiple fields in PIG scripts because the join operations takes records from one input and joins them with another input. This can be achieved by specifying the keys for each input and the two rows will be joined when the keys are equal.

[Apache Tajo Interview Questions](https://www.wisdomjobs.com/e-university/apache-tajo-interview-questions.html)

1. **Question 30. Does Pig Support Multi-line Commands?**

**Answer :**

Yes.

**Define Apache Pig**

Ans. To analyze large data sets representing them as data flows, we use Apache Pig. Basically,  to provide an abstraction over MapReduce, reducing the complexities of writing a MapReduce task using [**Java programming**](https://data-flair.training/blogs/java-tutorial/), Apache Pig is designed. Moreover, using Apache Pig, we can perform data manipulation operations very easily in [**Hadoop**](https://data-flair.training/blogs/hadoop-tutorial-for-beginners/).  
However, there is much more to learn about Apache Pig, for that follow the link:[**Pig Tutorial: A Comprehensive Guide to Pig Hadoop**](https://data-flair.training/blogs/apache-pig-tutorial/)

**Que 2. Why Do We Need Apache Pig?**

Ans. At times, while performing any MapReduce tasks, programmers who are not so good at Java normally used to struggle to work with Hadoop. Hence, Pig is a boon for all such programmers. The reason is:

* Using Pig Latin, programmers can perform [**MapReduce**](https://data-flair.training/blogs/hadoop-mapreduce-tutorial/)tasks easily, without having to type complex codes in Java.
* Since Pig uses multi-query approach, it also helps in reducing the length of codes.
* It is easy to learn Pig when you are familiar with SQL. It is because Pig Latin is SQL-like language.
* In order to support data operations, it offers many built-in operators like joins, filters, ordering, and many more. And, it offers nested data types that are missing from MapReduce, for example, tuples, bags, and maps.

**Que 3. What is the difference between Pig and SQL?**

Ans. Here, are the list of major differences between Apache Pig and SQL.

* **Pig**

It is a procedural language.

* **SQL**

While it is a declarative language.

* **Pig**

Here, the schema is optional. Although, without designing a schema, we can store data. However, it stores values as $01, $02 etc.

* **SQL**

In SQL, Schema is mandatory.

* **Pig**

In Pig, data model is nested relational.

* **SQL**

In SQL, data model used is flat relational.

* **Pig**

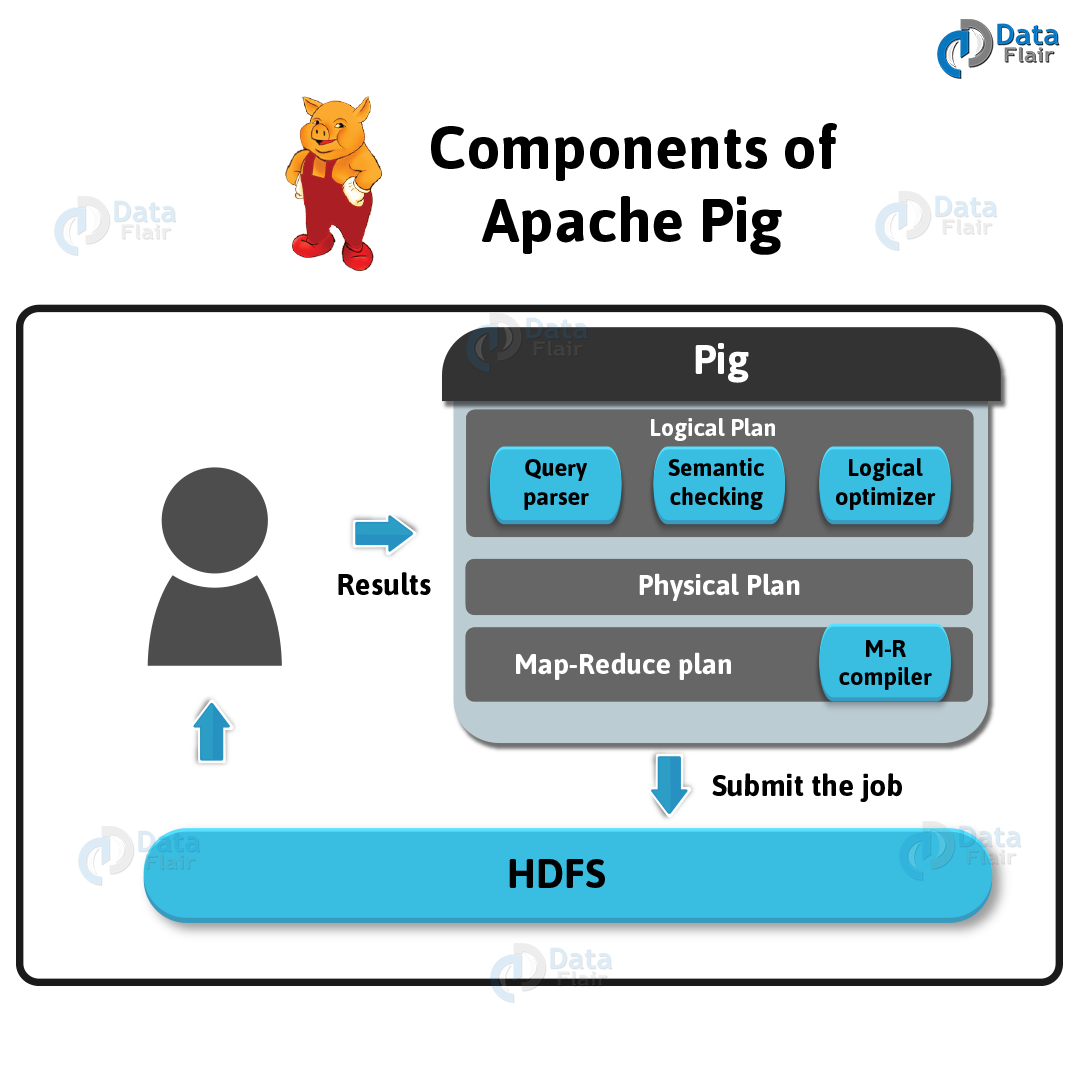
Here, we have limited opportunity for query optimization.

* **SQL**

While here we have more opportunity for query optimization.

**Que 4. Explain the architecture of Hadoop Pig.**

**Ans.**Below is the image, which shows the architecture of Apache Pig.

[](https://d2h0cx97tjks2p.cloudfront.net/blogs/wp-content/uploads/sites/2/2018/05/Components-of-Apache-Pig.png)

*Best 30 Apache Pig Interview Questions and Answers*

Now, we can see, several components in the Hadoop Pig framework. The major components are:

**1. Parser**

At first, Parser handles all the Pig Scripts. Basically, Parser checks the syntax of the script, does type checking, and other miscellaneous checks. Afterward, Parser’s output will be a DAG (directed acyclic graph). That represents the Pig Latin statements as well as logical operators.  
Basically, the logical operators of the script are represented as the nodes and the data flows are represented as edges, in the DAG (the logical plan).

**2. Optimizer**

Further, DAG is passed to the logical optimizer. That carries out the logical optimizations, like projection and push down.

**3. Compiler**

A series of MapReduce jobs have compiled from an optimized logical plan.

**4. Execution engine**

At last, these jobs are submitted to Hadoop in a sorted order. Hence, these MapReduce jobs are executed finally on Hadoop, that produces the desired results.

Learn about Pig Architecture in detail, follow the link: [**Apache Pig Architecture and Execution Modes**](https://data-flair.training/blogs/apache-pig-architecture/)

**Que 5. What is the difference between Apache Pig and Hive?**

Ans. Basically, to create MapReduce jobs, we use both Pig and Hive. Also, we can say, at times, Hive operates on[**HDFS**](https://data-flair.training/blogs/apache-hadoop-hdfs-introduction-tutorial/)as same as Pig does. So, here we are listing few significant points those set Apache Pig apart from[**Hive**](https://data-flair.training/blogs/apache-hive-tutorial/).

* **Hadoop Pig**

Pig Latin is a language, Apache Pig uses. Originally, it was created at Yahoo.

* **Hive**

HiveQL is a language, Hive uses. It was originally created at Facebook.

* **Pig**

It is a data flow language.

* **Hive**

Whereas, it is a query processing language.

* **Pig**

Moreover, it is a procedural language which fits in pipeline paradigm.

* **Hive**

It is a declarative language.

* **Apache Pig**

Also, can handle structured, unstructured, and semi-structured data.

* **Hive**

Whereas, it is mostly for structured data.

[**Follow this link, to learn more about Pig vs Hive**](https://data-flair.training/blogs/pig-vs-hive/)

**Que 6. What is the difference between Pig and MapReduce?**

Ans. Some major differences between Hadoop Pig and MapReduce, are:

* **Apache Pig**

It is a data flow language.

* **MapReduce**

However, it is a data processing paradigm.

* **Hadoop Pig**

Pig is a high-level language.

* **MapReduce**

Well, it is a low level and rigid.

* **Pig**

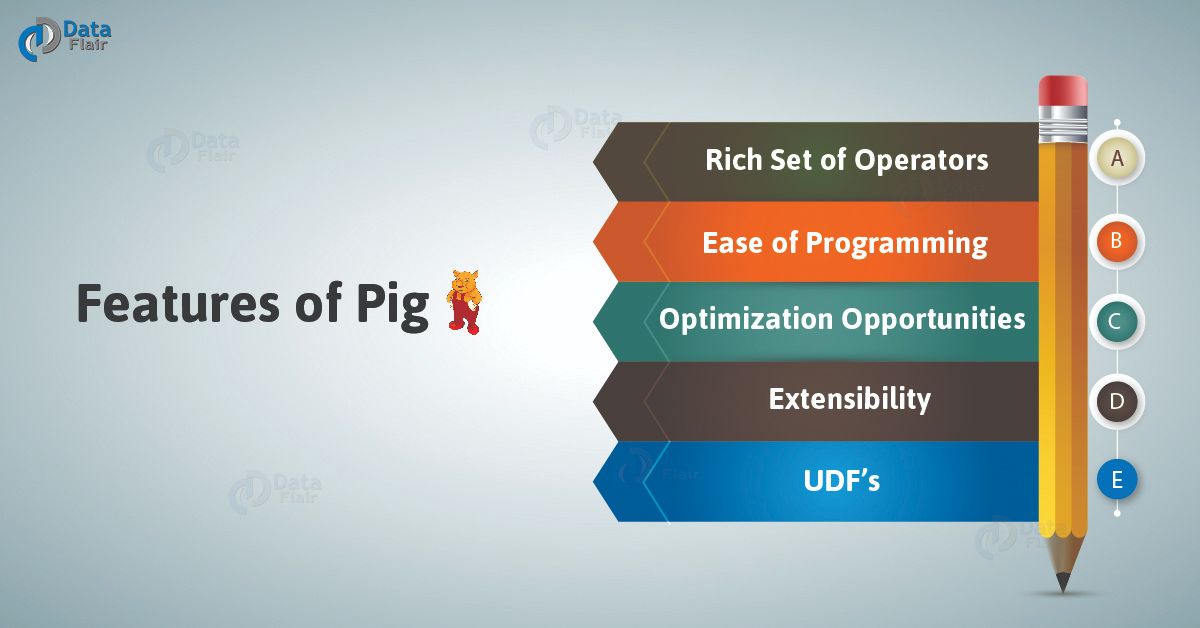
In Apache Pig, performing a join operation is pretty simple.

* **MapReduce**

But, in MapReduce, it is quite difficult to perform a join operation between datasets.

**Que 7. Explain Features of Pig.**

Ans. There are several features of Pig, such as:

[](https://d2h0cx97tjks2p.cloudfront.net/blogs/wp-content/uploads/sites/2/2018/05/Features-of-Pig-01-1.jpg)

*Features of Pig*

* **Rich set of operators**

In order to perform several operations, Pig offers many operators, for example, join, sort, filer and many more.

* **Ease of programming**

Since you are good at SQL, it is easy to write a Pig script. Because of Pig Latin as same as SQL.

* **Optimization opportunities**

In Apache Pig, all the tasks optimize their execution automatically. As a result, the programmers need to focus only on the semantics of the language.

* **Extensibility**

Through Pig, it is easy to read, process, and write data. It is possible by using the existing operators. Also, users can develop their own functions.

* **UDF’s**

By using Pig, we can create User-defined Functions in other programming languages. Like Java. Also, can invoke or embed them in Pig Scripts.  
Learn all the features in detail, follow the link: [**Top 12 Apache Pig Features you must know**](https://data-flair.training/blogs/apache-pig-features/)

**Que 8. What is Pig Storage?**

Ans. In Pig, there is a default load function, that is Pig Storage. Also, we can use pig storage, whenever we want to load data from a file system into the pig. We can also specify the delimiter of the data while loading data using pig storage (how the fields in the record are separated). Also, we can specify the schema of the data along with the type of the data.

**Que 9. While writing evaluate UDF, which method has to be overridden?**

Ans. We have to override the method exec() while writing UDF in the Pig. Whereas the base class can be different while writing filter UDF, we will have to extend FilterFunc and for evaluate UDF, we will have to extend the EvalFunc. EvaluFunc is parameterized and must provide the return type also.

**Que 10. What are the different UDF’s in Pig?**

Ans. On the basis of the number of rows, UDF can be processed. They are of two types:

* UDF that takes one record at a time, for example, Filter and Eval.
* UDFs that take multiple records at a time, for example, Avg and Sum.

Also, pig gives you the facility to write your own UDF’s for load/store the data.

[**Read more about Pif UDF’s in Detail**](https://data-flair.training/blogs/apache-pig-udf/)

**Apache Pig Interview Questions and Answers For Freshers. Q- 1,2,4,7**

**Apache Pig Interview Questions and Answers For Experience. Q- 3,5,6,8,9,10**

**Que 11. What are the Optimizations a developer can use during joins?**

Ans. We use replicated join, to perform join between a small dataset with a large dataset. Moreover, in the replicated join, the small dataset will be copied to all the machines where the mapper is running and the large dataset is divided across all the nodes. Also, it gives us the advantage of Map-side joins.

If your dataset is skewed i.e. if a particular data is repeated multiple times even if you use reduce side join, the particular reducer will be overloaded and it will take a lot of time. Pig itself, calculates skewed join and the skewed key.  
And, if you have datasets where the records are sorted in the same field, you can go for sorted join, this also happens in map phase and is very efficient and fast.

**Que 12. What is a skewed join?**

Ans. While we want to perform a join with a skewed dataset, that means a particular value will be repeated many times, is a skewed join.

**Que 13. What is Flatten?**

Ans. An operator in pig that removes the level of nesting, is Flatten. Sometimes, we have data in a bag or a tuple and we want to remove the level of nesting so that the data structured should become even, we use Flatten.

In addition, each Flatten produces a cross product of every record in the bag with all of the other expressions in the general statement.

**Que 14. What are the complex data types in pig?**

Ans. The following are the complex data types in Pig:

[](https://d2h0cx97tjks2p.cloudfront.net/blogs/wp-content/uploads/sites/2/2018/05/Data-Types-in-Pig-01.jpg)

*Data types in Pig*

* **Tuple**

An ordered set of fields is what we call a tuple.  
**For Example:** (Ankit, 32)

* **Bag**

A collection of tuples is what we call a bag.  
**For Example:** {(Ankit,32),(Neha,30)}

* **Map**

A set of key-value pairs is what we call a Map.

**For Example:** [ ‘name’#’Ankit’, ‘age’#32]

**Que 15. Why we use BloomMapFile?**

Ans. In order to extend MapFile, we use the BloomMapFile. That implies its functionality is similar to MapFile.

Also, to provide quick membership test for the keys, BloomMapFile uses dynamic Bloom filters. We use it in[**HBase**](https://data-flair.training/blogs/hadoop-hbase-tutorial/)table format.

**Que 16. How will you explain COGROUP in Pig?**

Ans. In Apache Pig, COGROUP works on tuples. On several statements, we can apply operators, which contains a few relations at least 127 relations at every time. When you make use of the operator on tables, then Pig immediately books two tables and join them through some of the columns that are grouped.

**Que 17. What is the difference between logical and physical plans?**

Ans. Pig undergoes some steps when a Pig Latin Script is converted into MapReduce jobs. After performing the basic parsing and semantic checking, it produces a logical plan. The logical plan describes the logical operators that have to be executed by Pig during execution. After this, Pig produces a physical plan. The physical plan describes the physical operators that are needed to execute the script.

**Que 18. Does ‘ILLUSTRATE’ run MR job?**

Ans. It will pull the internal data, illustrate will not pull any MR. Moreover, illustrate will not do any job, on the console. It just shows the output of each stage and not the final output.

**Que 19. Is the keyword ‘DEFINE’ as a function name?**

Ans. The keyword ‘DEFINE’ is like a function name. As soon as we have registered, we have to define it. Whatever logic you have written in Java program, we have an exported jar and also a jar registered by us. Now the compiler will check the function in the exported jar. When the function is not present in the library, it looks into our jar.

**Que 20. Is the keyword ‘FUNCTIONAL’ a User Defined Function (UDF)?**

Ans. The keyword ‘FUNCTIONAL’ is not a User Defined Function (UDF). we have to override some functions while using UDF. Certainly, we have to do our job with the help of these functions only. However, the keyword ‘FUNCTIONAL’ is a built-in function i.e a predefined function, therefore it does not work as a UDF.

**Apache Pig Interview Questions and Answers For Freshers. Q- 12,13,14,15,16,17**

**Apache Pig Interview Questions and Answers For Experience. Q- 11,18,19,20**

**Que 21. Why do we need MapReduce during Pig programming?**

Ans. Let’s understand it in this way- Pig is a high-level platform that makes many Hadoop data analysis issues easier to execute. And, we use Pig Latin for this platform. Now, a program written in Pig Latin is like a query written in SQL, where we need an execution engine to execute the query. Hence, when we write a program in Pig Latin, it was converted into MapReduce jobs by pig complier. As a result, MapReduce acts as an execution engine.

**Q 22 What are the scalar data types in Pig?**

Ans. In Apache Pig, Scalar data types are:

* int        -4bytes,
* float     -4bytes,
* double  -8bytes,
* long     -8bytes,
* char array,
* byte array

**Que 23 What are the different execution mode available in Pig?**

Ans. In Pig, there are 3 modes of execution available:

* Interactive Mode (Also known as Grunt Mode)
* Batch Mode
* Embedded Mode

**Que 24. Whether Pig Latin language is case-sensitive or not?**

Ans. We can say, Pig Latin is sometimes not a case-sensitive, for example, Load is equivalent to load.

A=load ‘b’ is not equivalent to a=load ‘b’

**Note:** UDF is also case-sensitive, here count is not equivalent to COUNT.

**Que 25. What is the purpose of ‘dump’ keyword in Pig?**

Ans. The keyword “dump” displays the output on the screen.

**For Example-** dump ‘processed’

**Que 26. Does Pig give any warning when there is a type mismatch or missing field?**

Ans. The pig will not show any warning if there is no matching field or a mismatch. However, if any mismatch occurs, it assumes a null value in Pig.

**Que 27. What is Grunt shell?**

Ans. Grunt shell is also what we call as Pig interactive shell. Basically, it offers a shell for users to interact with HDFS.

**Que 28. What co-group does in Pig?**

Ans. Basically, it joins the data set by grouping one particular data set only. Moreover, it groups the elements by their common field and then returns a set of records containing two separate bags. One bag consists of the record of the first data set with the common data set, while and other bag consists of the records of the second data set with the common data set.

**Que 29. What are relational operations in Pig latin?**

Ans. Relational operations in Pig Latin are:

* For each
* Order by
* Filters
* Group
* Distinct
* Join
* Limit

**Que 30. How is Pig Useful For?**

Ans. There are 3 possible categories for which we can use Pig. They are:

1) ETL data pipeline  
2) Research on raw data  
3) Iterative processing

**What are the different Relational Operators available in Pig language?**  
Answer: They are of following types:

* Loading and Storing
* Filtering
* Grouping and joining
* Sorting
* Combining and Splitting
* Diagnostic

**Que 3. What are the different modes available in Pig?**  
Answer: There are two modes available in the Pig:

* Local Mode (Runs on localhost file system)
* MapReduce Mode (Runs on Hadoop Cluster)

**Que 4. Can we say a COGROUP is a group of more than 1 data set?**  
Answer: Basically, it is a group of one data set. However, cogroup will group all the data sets and join them based on the common field in the case of more than one datasets. In addition, cogroup is a group of more than one data set and join of that data set as well.  
**Que 5. What does FOREACH do?**  
Answer: To apply transformations to the data and to generate new data items, we use FOREACH. According to its name, for each element of a data bag, the respective action will be performed.  
**Syntax:** FOREACH bagname GENERATE expression1, expression2, …..  
After GENERATE, all the expressions which are mentioned will apply to the current record of the data bag.  
**Que 6. Why do we use ‘filters’ Pig scripts?**  
Answer: State advantages of Apache Pig?  
Answer: Advantages of Apache Pig are:

* **Less Development time**

It consumes less time while development which is one of the major advantages.

* **Easy to learn**

However, we can say, Apache Pig’s Learn curve is not steep. So, if someone doesn’t know to write vanilla [**MapReduce**](https://data-flair.training/blogs/hadoop-mapreduce-tutorial/)or SQL they can write MapReduce jobs.

* **Procedural language**

It means Pig is a Procedural language. It is not declarative, unlike SQL. Hence, we can easily follow the commands. Also, in the transformation of data, it offers better expressiveness in every step. Moreover, on comparing it to vanilla MapReduce, it is much more like the English language which is very concise and unlike[**Java**](https://data-flair.training/blogs/java-tutorial/) but more like [**Python**](https://data-flair.training/blogs/python-tutorial-for-beginners/).

* **Dataflow**

It is a data flow language. That means where everything is about data even though we sacrifice control structures like for loop or if structures. Data transformation is a first class citizen, by this data and because of data. Also, we cannot create for loops, without data. we need to always transform and manipulate data.

* **Easy to control Execution**

We can control the execution of every step because it is procedural in nature. Also, a benefit that it is, straightforward. Hence, it is possible to write our own UDF(User Defined Function) and inject in one specific part in the pipeline.

* **UDFs**

It is possible to write our own [**UDFs**](https://data-flair.training/blogs/apache-pig-udf/).   
**Que 7. What is the bag?**  
Answer: It is one of the data models present in Pig. Moreover, it is an unordered collection of tuples with possible duplicates. Basically, to store collections while grouping, we use bags. We represent bags with “{}”.  
**Que 8. Why should we use ‘orderby’ keyword in pig scripts?**  
Answer: This statement sorts our data for us, producing a total order of our output data. In addition, the syntax of order is similar to the group. We indicate a key or set of keys by which we wish to order our data  
**For Example;**  
input2 = load ‘daily’ as (exchanges, stocks);  
grpds = order input2 by exchanges;  
**Que 9. What is Pig Latin?**  
**Answer:**It analyzes the data in Hadoop using Apache Pig. Here an interpreter layer transforms Pig Latin statements into MapReduce jobs. Then Hadoop process these jobs further. Also, we can say it is a very simple language with SQL like semantics.  
[**Follow this link to know more about Pig Latin**](https://data-flair.training/blogs/pig-latin/)  
**Que 10. What are the relational operators available related to Grouping and joining in Pig language?**  
Answer: The most powerful operators in Pig language are Grouping and Joining operators. Because in low-level MapReduce language, core MapReduce creation for grouping and joins are very typical.

* **JOIN**

To join two or more relations.

* **GROUP**

For aggregation of a single relation

* **COGROUP**

For the aggregation of multiple relations

* **CROSS**

In order to create a cartesian product of two or more relations, we use it.  
**Basic Pig Interview Questions. Q- 1,2,3,4,5,6,7,9**  
**Advanced Pig Interview Questions. Q- 8,10**  
**Que 11. What are the different String functions available in Pig?**  
Answer: STRING Pig functions are:

* UPPER
* LOWER
* TRIM
* SUBSTRING
* INDEXOF
* STRSPLIT
* LAST\_INDEX\_OF

**Que 12. What is a relation in Pig?**  
Answer: A bag of tuples, is what we call a Pig relation. It is as same as a table in a relational database, where the tuples in the bag correspond to the rows in a table. The only difference is Pig relations don’t require that every tuple contain the same number of fields or that the fields in the same position (column) have the same type.  
**Que 13. What is a tuple?**  
Answer: An ordered set of fields is what we call a tuple. Whereas, a field is a piece of data.  
**Que 14. What is the MapReduce plan in pig architecture?**  
Answer: Basically, the output of Physical plan is converted into an actual MapReduce program which then executes across the Hadoop Cluster.  
[**Let’s read more about Apache Pig Architecture**](https://data-flair.training/blogs/pig-architecture/)  
**Que 15. What relational operators can we use that are related to combining and splitting in Pig language?**  
**Answer:** We use UNION and SPLIT operators to combine and split relations in the Pig,  
**Que 16. What is UDF in Pig?**  
Answer: Apache Pig offers extensive support for User Defined Functions (UDF’s), in addition to the built-in functions. Also, it is possible to define our own functions and use them, using these UDF’s. Moreover, in six programming languages, UDF support is available. Such as[**Java**](https://data-flair.training/blogs/features-of-java/), Jython,[**Python,**](https://data-flair.training/blogs/python-features/) JavaScript, Ruby, and Groovy.  
However, we can say, complete support is only provided in Java. While in all the remaining languages limited support is provided. Also, we can write UDF’s which will include all the parts of the processing like column transformation, data load/store, and aggregation, using Java. Since Apache Pig has been written in Java, make sure the UDF’s written using Java language work efficiently compared to other languages.  
Also, we have a Java repository for UDF’s named Piggybank, in Apache Pig. Basically, we can access Java UDF’s written by other users, and contribute our own UDF’s, using Piggybank.  
[**Follow this link to know more about Apache Pig User Defined Functions**](https://data-flair.training/blogs/apache-pig-udf/)  
**Que 17. What are the primitive data types in Pig?**  
Answer: The primitive data types in Pig are:

* Int
* Long
* Float
* Double
* Char array
* Byte array

**Que 18. What is bag data type in Pig?**  
Answer: Basically, it works as a container for bags and tuple. Also, we can say it is very complex data type in Pig Latin language.  
**Que 19. Why should we use ‘distinct’ keyword in Pig scripts?**  
Answer: It is a very simple statement which removes duplicate records. Moreover, it works only on entire records, not on individual fields:  
For Example;  
input2 = load ‘daily’ as (exchanges, stocks);  
grpds = distinct exchanges;  
**Que 20. What are the different math functions available in Pig?**  
Answer: Below are most commonly used math Pig functions

* ABS
* ACOS
* EXP
* LOG
* ROUND
* CBRT
* RANDOM
* SQRT

**Basic Pig Interview Questions. Q- 11,12,13,16,17, 18,19,20**  
**Advanced Pig Interview Questions. Q- 14,15**  
**Que 21. What are the different Eval functions available in Pig?**  
Answer: Below are most commonly used Eval Pig functions

* AVG
* CONCAT
* MAX
* MIN
* SUM
* SIZE
* COUNT
* COUNT\_STAR
* DIFF
* TOKENIZE
* IsEmpty

[**Read more about Apache Pig Built-In Functions in detail**](https://data-flair.training/blogs/apache-pig-built-in-functions/)  
**Que 22. What are the relational operators available related to loading and storing in Pig language?**  
Answer: Pig uses following operators to Load data and Store it into[**HDFS,**](https://data-flair.training/blogs/apache-hadoop-hdfs-introduction-tutorial/)

* LOAD
* STORE

LOAD operator, loads the data from the file system, whereas STORE stores the data in the file system.  
**Que 23. What are different modes of execution in Apache Pig?**  
Answer: “Hadoop MapReduce (Java) Command Mode” and “Pig (Local Mode) Command Mode”. The first mode,  Local Mode needs access to only a single machine where all files are installed and executed on a local host whereas second mode MapReduce requires accessing the Hadoop cluster.  
**Que 24. Does Pig support multi-line commands?**  
Answer: Yes  
**Que 25. How would you diagnose or do exception handling in the Pig?**  
Answer. We can use following operators, for exception handling of Pig script.

* **DUMP-** “DUMP” operator displays the results on screen.
* **DESCRIBE-** “DESCRIBE” operator displays the schema of a particular relation.
* **ILLUSTRATE-** “ILLUSTRATE” operator displays step by step execution of a sequence of Pig statements.
* **EXPLAIN-** “EXPLAIN” operator displays the execution plan for pig latin statements.

**Que 26. What are Pig Execution modes?**  
Answer: There are two execution modes of Apache Pig. Although, it also depends upon where the Pig script is going to run. Also on where the data is residing. Then we can store data on a single machine or in a distributed environment like Clusters. To run Pig programs, 3 different modes are:

* Non-interactive shell or script mode, here the user has to create a file, load the code and execute the script.
* For running Apache Pig commands Grunt shell or interactive shell.
* Embedded mode, JDBC to run SQL programs from Java.

**Que 27. What are the different ways of executing Pig script?**  
Answer: To execute the Pig script, we have three ways:

* **Grunt Shell:** In order to execute all Pig Scripts, this is Pig’s interactive shell.
* **Script File:** Write all the Pig commands in a script file and execute the Pig script file. This is executed by the Pig Server.
* **Embedded Script:** If some functions are unavailable in built-in operators, it is possible to create User Defined Functions (UDF) to bring that functionality using other languages such as Java, Python, Ruby, etc.

[**Follow this link to know about Execute Pig Script**](https://data-flair.training/blogs/execute-pig-script/)  
**Que 28 What do you understand by an inner bag and outer bag in Pig?**  
Answer: Basically, Outer bag or relation is a bag of tuples. Moreover, relations are as same as relations in relational databases. For example:  
{(Taj Mahal, Agra), (India Gate, Delhi), (Qutub Minar, Delhi)}  
In addition, an inner bag contains a bag inside a tuple. For Example:  
(Delhi, {(India Gate, Delhi), (Qutub Minar, Delhi)})  
(Agra, {(Taj Mahal, Agra)})  
**Que 29. State the limitations of Apache Pig.**  
Answer: Limitations are:

* **Delay in Execution**

The commands are not executed unless either we dump or store an intermediate or final result. Basically, this increases the iteration between debug and resolve the issue.

* **Minor one**

Here, is an absence of good ide or plugin for Vim. That offers more functionality than syntax completion to write the pig scripts.

* **Errors of Pig**

Due to UDFS(Python), errors that [**Pig**](https://en.wikipedia.org/wiki/Apache_Pig) produces are not helpful at all. Because if anything goes wrong, it just gives the error like exec error in UDF even if the problem is related to syntax or the type error, let alone a logical one. It is huge.

* **Not mature**

Pig is still in the development, even if it has been around for quite some time. So, we can say it is not enough mature.

* **Support**

Generally, Google and StackOverflow do not lead good solutions for the problems.  
**Que 30. What are the Applications of Apache Pig?**  
More of its applications are:

* In order to process huge data sources. For example, weblogs.
* To perform tasks involving ad-hoc processing and quick prototyping.
* Also, to perform data processing for search platforms.
* Moreover, to process time sensitive data loads.

**Differentiate between Hadoop MapReduce and Pig**

|  |  |  |
| --- | --- | --- |
| **Characteristic** | **MapReduce** | **Pig** |
| Type of Language | Compiled Language | Scripting Language |
| Level of Abstraction | Low Level of Abstraction | Higher Level of Abstraction |
| Code | More lines of code is required. | Compatively less lines of code than Hadoop MapReduce. |
| Code Efficiency | Code efficiency is high. | Code efficiency is relatively less. |
| **Hadoop MapReduce vs Pig** | | |

**2)  Compare Apache Pig and SQL.**

* Apache Pig differs from SQL in its usage for ETL, lazy evaluation, store data at any given point of time in the pipeline, support for pipeline splits and explicit declaration of execution plans. SQL is oriented around queries which produce a single result. SQL has no in-built mechanism for splitting a data processing stream and applying different operators to each sub-stream.
* Apache Pig allows user code to be included at any point in the pipeline whereas if SQL where to be used data needs to be imported to the database first and then the process of cleaning and transformation begins.

**Explain the need for MapReduce while programming in Apache Pig.**

Apache Pig programs are written in a query language known as Pig Latin that is similar to the SQL query language. To execute the query, there is a need for an execution engine. The Pig engine converts the queries into MapReduce jobs and thus MapReduce acts as the execution engine and is needed to run the programs.

**4) Explain about the BloomMapFile.**

BloomMapFile is a class, that extends the MapFile class. It is used in HBase table format to provide quick membership test for the keys using dynamic bloom filters.

**5) What do you mean by a bag in Pig?**

Collection of tuples is referred as a bag in Apache Pig

**6) What is the usage of foreach operation in Pig scripts?**

FOREACH operation in Apache Pig is used to apply transformation to each element in the data bag, so that respective action is performed to generate new data items.

Syntax- FOREACH data\_bagname GENERATE exp1, exp2

**7) Explain about the different complex data types in Pig.**

Apache Pig supports 3 complex data types-

* Maps- These are key, value stores joined together using #.
* Tuples- Just similar to the row in a table, where different items are separated by a comma. Tuples can have multiple attributes.
* Bags- Unordered collection of tuples. Bag allows multiple duplicate tuples.

**8) What does Flatten do in Pig?**

Sometimes there is data in a tuple or a bag and if we want to remove the level of nesting from that data, then Flatten modifier in Pig can be used. Flatten un-nests bags and tuples. For tuples, the Flatten operator will substitute the fields of a tuple in place of a tuple, whereas un-nesting bags is a little complex because it requires creating new tuples.

**How do users interact with the shell in Apache Pig?**

Using Grunt i.e. Apache Pig’s interactive shell, users can interact with HDFS or the local file system. To start Grunt, users should invoke Apache Pig with no command –

Executing the command “pig –x local” will result in the prompt -

grunt >

This is where PigLatin scripts can be run either in local mode or in cluster mode by setting the configuration in PIG\_CLASSPATH.

To exit from grunt shell, press CTRL+D or just type exit.

**What are the debugging tools used for Apache Pig scripts?**

describe and explain are the important debugging utilities in Apache Pig.

* explain utility is helpful for Hadoop developers, when trying to debug error or optimize PigLatin scripts. explain can be applied on a particular alias in the script or it can be applied to the entire script in the grunt interactive shell. explain utility produces several graphs in text format which can be printed to a file.
* describe debugging utility is helpful to developers when writing Pig scripts as it shows the schema of a relation in the script. For beginners who are trying to learn Apache Pig can use the describe utility to understand how each operator makes alterations to data. A pig script can have multiple describes.

**11) What is illustrate used for in Apache Pig?**

Executing pig scripts on large data sets, usually takes a long time. To tackle this, developers run pig scripts on sample data but there is possibility that the sample data selected, might not execute your pig script properly. For instance, if the script has a join operator there should be at least a few records in the sample data that have the same key, otherwise the join operation will not return any results. To tackle these kind of issues, illustrate is used. illustrate takes a sample from the data and whenever it comes across operators like join or filter that remove data, it ensures that only some records pass through and some do not, by making modifications to the records such that they meet the condition. illustrate just shows the output of each stage but does not run any MapReduce task.

**12) Explain about the execution plans of a Pig Script**

**Or**

**Differentiate between the logical and physical plan of an Apache Pig script**

Logical and Physical plans are created during the execution of a pig script. Pig scripts are based on interpreter checking. Logical plan is produced after semantic checking and basic parsing and no data processing takes place during the creation of a logical plan. For each line in the Pig script, syntax check is performed for operators and a logical plan is created. Whenever an error is encountered within the script, an exception is thrown and the program execution ends, else for each statement in the script has its own logical plan.

A logical plan contains collection of operators in the script but does not contain the edges between the operators.

After the logical plan is generated, the script execution moves to the physical plan where there is a description about the physical operators, Apache Pig will use, to execute the Pig script. A physical plan is more or less like a series of MapReduce jobs but then the plan does not have any reference on how it will be executed in MapReduce. During the creation of physical plan, cogroup logical operator is converted into 3 physical operators namely –Local Rearrange, Global Rearrange and Package. Load and store functions usually get resolved in the physical plan.

**13) What do you know about the case sensitivity of Apache Pig?**

It is difficult to say whether Apache Pig is case sensitive or case insensitive. For instance, user defined functions, relations and field names in pig are case sensitive i.e. the function  COUNT is not the same as function count or X=load ‘foo’ is not same as x=load ‘foo’. On the other hand, keywords in Apache Pig are case insensitive i.e. LOAD is same as load.

**14) What are some of the Apache Pig use cases you can think of?**

Apache Pig big data tools, is used in particular for iterative processing, research on raw data and for traditional ETL data pipelines. As Pig can operate in circumstances where the schema is not known, inconsistent or incomplete- it is widely used by researchers who want to make use of the data before it is cleaned and loaded into the data warehouse.

To build behaviour prediction models, for instance, it can be used by a website to track the response of the visitors to various types of ads, images, articles, etc.

**15) Differentiate between PigLatin and HiveQL**

* It is necessary to specify the schema in HiveQL, whereas it is optional in PigLatin.
* HiveQL is a declarative language, whereas PigLatin is procedural.
* HiveQL follows a flat relational data model, whereas PigLatin has nested relational data model.

**Is PigLatin a strongly typed language? If yes, then how did you come to the conclusion?**

In a strongly typed language, the user has to declare the type of all variables upfront. In Apache Pig, when you describe the schema of the data, it expects the data to come in the same format you mentioned. However, when the schema is not known, the script will adapt to actually data types at runtime. So, it can be said that PigLatin is strongly typed in most cases but in rare cases it is gently typed, i.e. it continues to work with data that does not live up to its expectations.

**17) What do you understand by an inner bag and outer bag in Pig?**

A relation inside a bag is referred to as inner bag and outer bag is just a relation in Pig

**18) Differentiate between GROUP and COGROUP operators.**

Both GROUP and COGROUP operators are identical and can work with one or more relations. GROUP operator is generally used to group the data in a single relation for better readability, whereas COGROUP can be used to group the data in 2 or more relations. COGROUP is more like a combination of GROUP and JOIN, i.e., it groups the tables based on a column and then joins them on the grouped columns. It is possible to cogroup up to 127 relations at a time.

**19) Explain the difference between COUNT\_STAR and COUNT functions in Apache Pig?**

COUNT function does not include the NULL value when counting the number of elements in a bag, whereas COUNT\_STAR (0 function includes NULL values while counting.

**20) What are the various diagnostic operators available in Apache Pig?**

1. **Dump Operator-**It is used to display the output of pig Latin statements on the screen, so that developers can debug the code.
2. **Describe Operator-**Explained in apache pig interview question no- 10
3. **Explain Operator-**Explained in apache pig interview question no -10
4. **Illustrate Operator-** Explained in apache pig interview question no -11

**21) How will you merge the contents of two or more relations and divide a single relation into two or more relations?**

This can be accomplished using the UNION and SPLIT operators.

**22) I have a relation R. How can I get the top 10 tuples from the relation R.?**

TOP () function returns the top N tuples from a bag of tuples or a relation. N is passed as a parameter to the function top () along with the column whose values are to be compared and the relation R.

**23) What are the commonalities between Pig and Hive?**

* HiveQL and PigLatin both convert the commands into MapReduce jobs.
* They cannot be used for OLAP transactions as it is difficult to execute low latency queries.

**24)  What are the different types of UDF’s in Java supported by Apache Pig?**

Algebraic, Eval and Filter functions are the various types of UDF’s supported in Pig.

**25) You have a file employee.txt in the HDFS directory with 100 records. You want to see only the first 10 records from the employee.txt file. How will you do this?**

The first step would be to load the file employee.txt into with the relation name as Employee.

The first 10 records of the employee data can be obtained using the limit operator -

Result= limit employee 10.

**26) Explain about the scalar datatypes in Apache Pig.**

integer, float, double, long, bytearray and char array are the available scalar datatypes in Apache Pig.

**27) How do users interact with HDFS in Apache Pig ?**

Using the grunt shell.

**28) What is the use of having Filters in Apache Pig ?**

**Just like the where clause in SQL, Apache Pig has filters to extract records based on a given condition or predicate. The record is passed down the pipeline if the predicate or the condition turn to true. Predicate contains various operators like ==, <=,!=, >=.**

**Example -**

X= load ‘inputs’ as(name,address)

Y = filter X by symbol matches ‘Mr.\*’;

**29) What is a UDF in Pig?**

If the in-built operators do not provide some functions then programmers can implement those functionalities by writing user defined functions using other programming languages like Java, Python, Ruby, etc. These User Defined Functions (UDF’s) can then be embedded into a Pig Latin Script.

**30) Can you join multiple fields in Apache Pig Scripts ?**

Yes, it is possible to join multiple fields in PIG scripts because the join operations takes records from one input and joins them with another input. This can be achieved by specifying the keys for each input and the two rows will be joined when the keys are equal.

**31) Does Pig support multi-line commands?**

Yes

## **What are the use cases of Apache Pig?**

Apache Pig is used for analyzing and performing tasks involving ad-hoc processing. Apache Pig is used for:

* Research on large raw data sets like data processing for search platforms. For example, Yahoo uses Apache Pig to analyse data gathered from Yahoo search engines and Yahoo News Feeds.
* Processing huge data sets like Web logs, streaming online data, etc.
* In customer behavior prediction models like e-commerce websites.

## **3. What is the difference between logical and physical plans?**

♣ Tip: Approach this question by explaining when does the logical and physical plans are created.

Pig undergoes some steps when a Pig Latin Script is converted into MapReduce jobs by the compiler. Logical and Physical plans are created during the execution of a pig script.

After performing the basic parsing and semantic checking, the parser produces a logical plan and no data processing takes place during the creation of a logical plan. The logical plan describes the logical operators that have to be executed by Pig during execution. For each line in the Pig script, syntax check is performed for operators and a logical plan is created. If an error is encountered, an exception is thrown and the program execution ends.

A logical plan contains a collection of operators in the script, but does not contain the edges between the operators.

After the logical plan is generated, the script execution moves to the physical plan where there is a description about the physical operators, Apache Pig will use, to execute the Pig script. A physical plan is like a series of MapReduce jobs, but the physical plan does not have any reference on how it will be executed in MapReduce.

## **4. How Pig programming gets converted into MapReduce jobs?**

Pig is a high-level platform that makes many Hadoop data analysis issues easier to execute. A program written in Pig Latin is a data flow language, which need an execution engine to execute the query. So, when a program is written in Pig Latin, Pig compiler converts the program into MapReduce jobs.

## **5. What are the components of Pig Execution Environment?**

The components of Apache Pig Execution Environment are:

* **Pig Scripts**: Pig scripts are submitted to the Apache Pig execution environment which can be written in Pig Latin using built-in operators and UDFs can be embedded in it.
* **Parser**: The Parser does the type checking and checks the syntax of the script. The parser outputs a DAG (directed acyclic graph). DAG represents the Pig Latin statements and logical operators.
* **Optimizer:** The Optimizer performs the optimization activities like split, merge, transform, reorder operators, etc. The optimizer provides the automatic optimization feature to Apache Pig. The optimizer basically aims to reduce the amount of data in the pipeline.
* **Compiler**: The Apache Pig compiler converts the optimized code into MapReduce jobs automatically.
* **Execution Engine**: Finally, the MapReduce jobs are submitted to the execution engine. Then, the MapReduce jobs are executed and the required result is produced.

## **6. What are the different ways of executing Pig script?**

There are three ways to execute the Pig script:

* **Grunt Shell**: This is Pig’s interactive shell provided to execute all Pig Scripts.
* **Script File**: Write all the Pig commands in a script file and execute the Pig script file. This is executed by the Pig Server.
* **Embedded Script**: If some functions are unavailable in built-in operators, we can programmatically create User Defined Functions (UDF) to bring that functionality using other languages like Java, Python, Ruby, etc. and embed it in the Pig Latin Script file. Then, execute that script file.

## **7. What are the data types of Pig Latin?**

Pig Latin can handle both atomic data types like int, float, long, double etc. and complex data types like tuple, bag and map.

Atomic or scalar data types are the basic data types which are used in all the languages like string, int, float, long, double, char[], byte[]. These are also called the primitive data types.

The complex data types supported by Pig Latin are:

* **Tuple**: Tuple is an ordered set of fields which may contain different data types for each field.
* **Bag**: A bag is a collection of a set of tuples and these tuples are a subset of rows or entire rows of a table.
* **Map**: A map is key-value pairs used to represent data elements. The key must be a chararray [] and should be unique like column name, so it can be indexed and value associated with it can be accessed on the basis of the keys. The value can be of any data type.

## **What is a bag in Pig Latin?**

A bag is one of the data models present in Pig. It is an unordered collection of tuples with possible duplicates. Bags are used to store collections of tuples while grouping. The size of bag is the size of the local disk, this means that the size of the bag is limited. When the bag is full, then Pig will spill this bag into local disk and keep only some parts of the bag in memory. There is no necessity that the complete bag should fit into memory. We represent bags with “{}”.

♣ Tip:You can also explain the two types of bag in Pig Latin i.e. outer bag and inner bag, which may impress your employers.

## **9. What do you understand by an inner bag and outer bag in Pig?**

Outer bag or relation is nothing but a bag of tuples. Here relations are similar as relations in relational databases. For example:

{(Linkin Park, California), (Metallica, Los Angeles), (Mega Death, Los Angeles)}

An inner bag contains a bag inside a tuple. For Example:

(Los Angeles, {(Metallica, Los Angeles), (Mega Death, Los Angeles)})

(California, {(Linkin Park, California)})

## **10. How Apache Pig deals with the schema and schema-less data?**

♣ Tip: Apache Pig deals with both schema and schema-less data. Thus, this is an important question to focus on.

The Apache Pig handles both, schema as well as schema-less data.

* If the schema only includes the field name, the data type of field is considered as a byte array.
* If you assign a name to the field you can access the field by both, the field name and the positional notation, whereas if field name is missing we can only access it by the positional notation i.e. $ followed by the index number.
* If you perform any operation which is a combination of relations (like JOIN, COGROUP, etc.) and if any of the relation is missing schema, the resulting relation will have null schema.
* If the schema is null, Pig will consider it as a byte array and the real data type of field will be determined dynamically.

## **11. How do users interact with the shell in Apache Pig?**

Using Grunt i.e. Apache Pig’s interactive shell, users can interact with HDFS or the local file system.

To start Grunt, users should use pig –x local command . This command will prompt Grunt shell. To exit from grunt shell, press CTRL+D or just type exit.

## **12. What is UDF?**

If some functions are unavailable in built-in operators, we can programmatically create User Defined Functions (UDF) to bring that functionality using other languages like Java, Python, Ruby, etc. and embed it in the Pig Latin Script file.

♣ Tip: To understand how to create and work with UDF, go through this blog – [***creating UDF in Apache Pig***](https://www.edureka.co/blog/creating-udf-in-apache-pig/)***.***

Important points about UDF to focus on:

* LoadFunc abstract class has three main methods for loading data and for most use cases it would suffice to extend it.
* LoadPush has methods to push operations from Pig runtime into loader implementations.
* setUdfContextSignature() method will be called by Pig both in the front end and back end to pass a unique signature to the Loader.
* The load/store UDFs control how data goes into Pig and comes out of Pig.
* The meaning of getNext() is called by Pig runtime to get the next tuple in the data.
* The loader should use setLocation() method to communicate the load information to the underlying InputFormat.
* prepareToRead method enables the RecordReader associated with the InputFormat provided by the LoadFunc is passed to the LoadFunc. The RecordReader can then be used by the implementation in getNext() to return a tuple representing a record of data back to pig.
* pushProjection() method tells LoadFunc which fields are required in the Pig script. Pig will use the column index requiredField.index to communicate with the LoadFunc about the fields required by the Pig script.
* LoadCaster has methods to convert byte arrays to specific types.
* A loader implementation should implement LoadCaster() if casts (implicit or explicit) from DataByteArray fields to other types need to be supported. LoadCaster has methods to convert byte arrays to specific types.

## **13. List the diagnostic operators in Pig.**

Pig supports a number of diagnostic operators that you can use to debug Pig scripts.

* **DUMP:**Displays the contents of a relation to the screen.
* **DESCRIBE:** Return the schema of a relation.
* **EXPLAIN:** Display the logical, physical, and MapReduce execution plans.
* **ILLUSTRATE:** Gives the step-by-step execution of a sequence of statements.

♣ Tip: Go through this blog on [***diagnostic operators***](https://www.edureka.co/blog/operators-in-apache-pig-diagnostic-operators/), to understand them and see their implementations.

## **14. Does ‘ILLUSTRATE’ run a MapReduce job?**

No, illustrate will not pull any MapReduce, it will pull the internal data. On the console, illustrate will not do any job. It just shows the output of each stage and not the final output.

ILLUSTRATE operator is used to review how data is transformed through a sequence of Pig Latin statements. ILLUSTRATE command is your best friend when it comes to debugging a script. This command alone might be a good reason for choosing Pig over something else.

Syntax: illustrate relation\_name;

## **15. What does illustrate do in Apache Pig?**

Executing Pig scripts on large data sets, usually takes a long time. To tackle this, developers run Pig scripts on sample data, but there is possibility that the sample data selected, might not execute your Pig script properly. For instance, if the script has a join operator there should be at least a few records in the sample data that have the same key, otherwise the join operation will not return any results.

To tackle these kind of issues, illustrate is used. Illustrate takes a sample of the data and whenever it comes across operators like join or filter that remove data, it ensures that only some records pass through and some do not, by making modifications to the records such that they meet the condition. Illustrate just shows the output of each stage but does not run any MapReduce task.

## **16. List the relational operators in Pig.**

All Pig Latin statements operate on relations (and operators are called relational operators). Different relational operators in Pig Latin are:

* **COGROUP**: Joins two or more tables and then perform GROUP operation on the joined table result.
* **CROSS**: CROSS operator is used to compute the cross product (Cartesian product) of two or more relations.
* **DISTINCT**: Removes duplicate tuples in a relation.
* **FILTER**: Select a set of tuples from a relation based on a condition.
* **FOREACH**: Iterate the tuples of a relation, generating a data transformation.
* **GROUP**: Group the data in one or more relations.
* **JOIN**: Join two or more relations (inner or outer join).
* **LIMIT**: Limit the number of output tuples.
* **LOAD**: Load data from the file system.
* **ORDER**: Sort a relation based on one or more fields.
* **SPLIT**: Partition a relation into two or more relations.
* **STORE**: Store data in the file system.
* **UNION**: Merge the content of two relations. To perform a UNION operation on two relations, their columns and domains must be identical.

♣ Tip: Go through this blog on [***relational operators***](https://www.edureka.co/blog/operators-in-apache-pig/), to understand them and see their implementations.

## **17. Is the keyword ‘DEFINE’ like a function name?**

Yes, the keyword ‘DEFINE’ is like a function name.

DEFINE statement is used to assign a name (alias) to a UDF function or to a streaming command.

* The function has a long package name that you don’t want to include in a script, especially if you call the function several times in that script. The constructor for the function takes string parameters. If you need to use different constructor parameters for different calls to the function you will need to create multiple defines – one for each parameter set.
* The streaming command specification is complex. The streaming command specification requires additional parameters (input, output, and so on). So, assigning an alias makes it easier to access.

**18. What is the function of co-group in Pig?**

COGROUP takes members of different relations, binds them by similar fields, and creates a bag that contains a single instance of both relations where those relations have common fields. Co-group operation joins the data set by grouping one particular data set only.

It groups the elements by their common field and then returns a set of records containing two separate bags. The first bag consists of the first data set record with the common data set and the second bag consists of the second data set records with the common data set.

## **19. Can we say co-group is a group of more than 1 data set?**

Co-group is a group of data sets. More than one data set, co-group will group all the data sets and join them based on the common field. Hence, we can say that co-group is a group of more than one data set and join of that data set as well.

## **20. The difference between GROUP and COGROUP operators in Pig?**

Group and Cogroup operators are identical. For readability, GROUP is used in statements involving one relation and COGROUP is used in statements involving two or more relations. Group operator collects all records with the same key. Cogroup is a combination of group and join, it is a generalization of a group instead of collecting records of one input depends on a key, it collects records of n inputs based on a key. At a time, we can Cogroup up to 127 relations.

## **21. You have a file personal\_data.txt in the HDFS directory with 100 records. You want to see only the first 5 records from the employee.txt file. How will you do this?**

For getting only 5 records from 100 records we use limit operator.

First load the data in Pig:

personal\_data = LOAD “/personal\_data.txt” USING PigStorage(‘,’) as (parameter1, Parameter2, …);

Then Limit the data to 5 records:

limit\_data = LIMIT personal\_data 5;

## **22. What is a MapFile?**

MapFile is a class which serves file-based map from keys to values.

A map is a directory containing two files, the data file, containing all keys and values in the map, and a smaller index file, containing a fraction of the keys. The fraction is determined by MapFile.Writer.getIndexInterval().

The index file is read entirely into memory. Thus, key implementations should try to keep themselves small. Map files are created by adding entries in-order.

## **23. What is BloomMapFile used for?**

The BloomMapFile is a class that extends MapFile. So its functionality is similar to MapFile. BloomMapFile uses dynamic Bloom filters to provide quick membership test for the keys. It is used in Hbase table format.

## **24. What are the different execution modes available in Pig?**

The execution modes in Apache Pig are:

* **MapReduce Mode**: This is the default mode, which requires access to a Hadoop cluster and HDFS installation. Since, this is a default mode, it is not necessary to specify -x flag (you can execute pig OR pig -x mapreduce). The input and output in this mode are present on HDFS.
* **Local Mode:** With access to a single machine, all files are installed and run using a local host and file system. Here the local mode is specified using ‘-x flag’ (pig -x local). The input and output in this mode are present on local file system.

## **25. Is Pig script case sensitive?**

♣ Tip: Explain the both aspects of Apache Pig i.e. case-sensitive as well as case-insensitive aspect.

Pig script is both case sensitive and case insensitive.

User defined functions, the field name, and relations are case sensitive i.e. EMPLOYEE is not same as employee or M=LOAD ‘data’ is not same as M=LOAD ‘Data’.

Whereas Pig script keywords are case insensitive i.e. LOAD is same as load.

It is difficult to say whether Apache Pig is case sensitive or case insensitive. For instance, user defined functions, relations and field names in Pig are case sensitive. On the other hand, keywords in Apache Pig are case insensitive.

## **What does Flatten do in Pig?**

Sometimes there is data in a tuple or a bag and if we want to remove the level of nesting from that data, then Flatten modifier in Pig can be used. Flatten un-nests bags and tuples. For tuples, the Flatten operator will substitute the fields of a tuple in place of a tuple, whereas un-nesting bags is a little complex because it requires creating new tuples.

## **27. What is Pig Statistics? What are all stats classes in the Java API package available?**

Pig Statistics is a framework for collecting and storing script-level statistics for Pig Latin. Characteristics of Pig Latin scripts and the resulting MapReduce jobs are collected while the script is executed. These statistics are then available for Pig users and tools using Pig (such as Oozie) to retrieve after the job is completed.

The stats classes are in the package org.apache.pig.tools.pigstats:

* PigStats
* JobStats
* OutputStats
* InputStats.

## **28. What are the limitations of the Pig?**

Limitations of the Apache Pig are:

1. As the Pig platform is designed for ETL-type use cases, it’s not a better choice for real-time scenarios.
2. Apache Pig is not a good choice for pinpointing a single record in huge data sets.
3. Apache Pig is built on top of MapReduce, which is batch processing oriented.

**Compare Pig and Hive.**

|  |  |  |
| --- | --- | --- |
| **Criteria** | **Pig** | **Hive** |
| Language | Pig Latin | SQL-like |
| Application | Programming purposes | Report creation |
| Operation | Client Side | Server side |
| Data support | Semi-structured | Structured |
| Connectivity | Can be called by other applications | JDBC & BI tool integration |

**2. Does Pig differ from MapReduce? If yes, how?**

Yes, Pig differs from MapReduce because, in MapReduce, the group by operation is performed at reducer side and filter, and also in the map phase the projection is implemented. Pig Latin provides the operations that are similar to MapReduce, such as groupby, orderby, and filters. We can analyze the Pig script and data flow to find the error checking. Pig Latin is lower in cost to write and maintain compared to MapReduce Java code.

Go through this Apache [Pig Tutorial](https://intellipaat.com/tutorial/hadoop-tutorial/apache-pig/) to get a better understanding of the concepts.

**3. Explain the uses of Map Reduce in Pig.**

* Apache Pig programs are written in Pig Latin query language which is similar to the SQL query language. To execute this queries, there requires an execution engine. The Pig engine enables to convert the queries into MapReduce jobs and thus MapReduce acts as the execution engine and is designed to run the programs as per the requirements.
* Pigs’ operators are using Hadoops’ API depending upon the configurations the job is executed in local mode or Hadoop cluster. Pig is never passes any outputs to Hadoop instead set the inputs and data locations for map-reduce.
* Pig Latin provides a set of standard Data-processing operations, such as join, filter, group by, order by, union, etc which are mapped to do the map-reduce tasks. A Pig Latin script describes a (DAG) directed acyclic graph, where the edges are data flows and the nodes are operators that process the data.

**4. Explain the uses of PIG.**

We can use Pig in three categories, they are

1. ETL data pipeline : It helps to populate our data warehouse. Pig can pipeline the data to an external application, it will wait until it’s finished, so that it has receive the processed data and continue from there. It is the most common use case for Pig.
2. Research on raw data.
3. Iterative processing.

Know on [What is Apache Pig?](https://intellipaat.com/blog/what-is-pig/) in this riveting blog post .

**5. Name the scalar data type and complex data types in Pig.**

The scalar data types in pig are int, float, double, long, chararray, and bytearray.  
The complex data types in Pig are map, tuple, and bag.  
Map: The data element with the data type chararray where element has pig data type include complex data type

Example- [city’#’bang’,’pin’#560001]

In this city and pin are data element mapping to values.

**Tuple :** It is a collection of data types and it has fixed length. Tuple is having multiple fields and these are ordered.

**Bag :** It is a collection of tuples, but it is unordered, tuples in the bag are separated by comma

Example: {(‘Bangalore’, 560001),(‘Mysore’,570001),(‘Mumbai’,400001)

***Interested in learning Apache Pig? Well, we have the***[**Big Data Training Course**](https://intellipaat.com/big-data-hadoop-training/)***to give you a head start in your career.***

**6. State the usage of ‘filters’, ‘group’ , ‘orderBy’, ‘distinct’ keywords in pig scripts.**

**Filters :**Filters has the similar functionality as where clause in SQL. Filters contain predicate and if it evaluates true for a given record, then that record will be passed down the pipeline. Otherwise, it will not predicate the results and thus contains different operators like ==,>=, <=,!=.so,== and != which is been applied in creating maps and tuples.

A= load ‘inputs’ as (name,address)

B=filter A by symbol matches ‘CM.\*';

**GroupBy :**The group statement collects various records with the same key. In SQL database GroupBy creates a group which feeds directly to one or more aggregate functions. But in Pig Latin has no direct connection between group and aggregate functions.

Input 2 = load ‘daily’ as(exchanges,stocks);

grpds = group input2 by stocks;

**Order :** The Order statement sorts the data producing a total order of output data. The Order syntax is similar to Group. Give a key or set of keys to order your data as per requirement. The following are the examples for the same:

Input 2 = load ‘daily’ as(exchanges,stocks);

grpds = order input2 by exchanges;

**Distinct :** The distinct statement is very simple to understand and implement. It removes duplicate records and the original data will be secured. It is implemented only on entire records, not on individual fields. Consider the below examples which explains the same:

Input 2 = load ‘daily’ as(exchanges,stocks);

grpds = distinct exchanges;

**7. Explain the LOAD keyword in Pig script.**

Load helps to load data from the file system. It is a relational operator  
In the first step in data-flow language we need to mention the input, which is completed by using ‘load’ keyword.  
**The LOAD syntax is**

LOAD ‘mydata’ [USING function] [AS schema];

Example- A = LOAD ‘intellipaat.txt’;

A = LOAD ‘intellipaat.txt’ USINGPigStorage(‘\t’);

**8. What are the relation operations in Pig? Explain any two with examples.**

The relational operations in Pig:

foreach, order by, filters, group, distinct, join, limit.foreach: It takes a set of expressions and applies them to all records in the data pipeline to the next operator.A =LOAD ‘input’ as (emp\_name :charrarray, emp\_id : long, emp\_add : chararray, phone : chararray, preferences : map [] );B = foreach A generate emp\_name, emp\_id;Filters: It contains a predicate and it allows us to select which records will be retained in our data pipeline.

Syntax: alias = FILTER alias BY expression;

Alias indicates the name of the relation, By indicates required keyword and the expression has Boolean.

Example: M = FILTER N BY F5 == 4;

**Take this**[**Hadoop online training**](https://intellipaat.com/big-data-hadoop-training/)**to learn about Pig, Pig Latin and everything to deploy Pig in real world scenario.**

**9. Does Pig support multi-line commands?**

Yes, pig supports both single line and multi-line commands. In single line command it executes the data, but it doesn’t store in the file system, but in multiple lines commands it stores the data into ‘/output’;/\* , so it can store the data in HDFS.

**10. Explain different execution modes available in Pig.**

Three different execution modes available in Pig they are,

1. Interactive mode or Grunt mode.
2. Batch mode or Script mode.
3. Embedded mode  
   Interactive mode or grunt mode: Pig’s interactive shell is known as grunt shell. If no file is specified to run in Pig it will start.
4. grunt> run scriptfile.pig

grunt> exec scriptfile.pig

**Batch mode or Script mode :** Pig executes the specified commands in the script file.

**Embedded mode :** We can embed Pig programs in Java and we can run the programs from Java.

Get a [clear understanding of Apache Pig](https://intellipaat.com/blog/what-is-pig/) in this riveting blog now.

**11. What are the exception handling operators in Pig script?**

Following operators are used for handling the exception in pig script.

**DUMP :** It helps to display the results on screen.

**DESCRIBE :** It helps to display the schema of aparticular relation.

**ILLUSTRATE :** It helps to display step by step execution of a sequence of pig statements

**EXPLAIN :** It helps to display the execution plan for Pig Latin statements.

**12. Differentiate between the physical plan and logical plan in Pig script.**

Both plans are created while to execute the pig script.

**Physical plan :** It is a series of MapReduce jobs while creating the physical plan.It’s divided into three physical operators such as Local Rearrange, Global Rearrange, and package. It illustrates the physical operators Pig will use to execute the script without referring to how they will execute in MapReduce Loading and storing functions are resolved in physical plan.

Example- A: Load(/emp:PigStorage(‘ ‘))

**Logical plan :** The Logical plan is a plan which is created for each line in the Pig scripts. It is produced after semantic checking and basic parsing. With every line, the logical plan for that particular program becomes extended and larger because each and every statement has its own logical plan.Loading and storing function are not resolved in logical plan.

Example: X: (Name: LOLoad schema: emp\_id#36:bytearray,emp\_name#37:bytearray,city#38:bytearray,salary#39:bytearray)Required Fields:null

**13. Is Pig script case sensitive?**

Pig script is both case sensitive and case insensitive. For example, in user defined functions, the field name, and relations are case sensitive ,i.e., INTELLIPAAT is not same as intellipaat or M=load ‘test’ is not same as m=load ‘test’. And Pig script keywords are case insensitive i.e., LOAD is same as a load.

***Give your career a big boost by learning Apache Pig through our***[***Comprehensive Training Course***](https://intellipaat.com/big-data-hadoop-training/#curriculum)***now!***

**14. Highlight the difference between group and Cogroup operators in Pig.**

Both the operators can work with one or more relations. Group and Cogroup operators are identical. Group operator collects all records with the same key. Cogroup is a combination of group and join, it is a generalization of a group instead of collecting records of one input depends on a key, it collects records of n inputs based on a key. At a time we can Cogroup upto 127 relations.

**15. What is the function of UNION and SPLIT operators? Give examples.**

**Union operator helps to merge the contents of two or more relations.**

Syntax: grunt> Relation\_name3 = UNION Relation\_name1, Relation\_name2

Example: grunt> INTELLIPAAT = UNION intellipaat\_data1.txt intellipaat\_data2.txt

**SPLIT operator helps to divide the contents of two or more relations.**

Syntax: grunt> SPLIT Relationa1\_name INTO Relationa2\_name IF (condition1), Relation2\_name (condition2);

Example: SPLIT student\_details into student\_details1 if marks<35, student\_details2 if (8590);

**16. How can we see only top 15 records from the student.txt out of 100 records in the HDFS directory?**

We should change the name student.txt into STUDENT it is the relation name. We can see the top 15 records in using limit operator  
Result = limit student 15.

**17. What is the use of BloomMapFile?**

It is an extended class of MapFile. Its functionality is similar to MapFile. It is used in the Hbase table format, Bloom Map File uses dynamic Bloom filters to provide rapid membership test for the keys.

**18. How does the Pig platform handle relational systems data?**

There are two ways Pig can work with relational datasets.

1. Load relational data directly into the Hadoop framework, where Pig can access it.
2. Using database connectors, Pig can load data directly from a relational database system and we can access it.

**19. What are the drawbacks of Pig?**

Some of the drawbacks of Pig are:

1. Pig is not really a convenient option for real-time use cases.
2. Pig does not prove to be useful when you need to fetch single record from a huge dataset.
3. Since it works on MapReduce, it works in batches.

**20. Mention the common features in Pig and Hive.**

The common features in Both Hive and Pig are

1. Internally both are converted the commands into MapReduce.
2. Both the technologies provide high-level abstractions.
3. Both do not support low-latency queries.
4. Both do not support OLAP or OLTP.

**21. Differentiate between Pig Latin and Pig Engine.**

Pig Latin is scripting language like Perl for searching huge data sets and it is made up of a series of transformations and operations that are applied to the input data to produce data.  
Pig engine is an environment to execute the Pig Latin programs. It converts Pig Latin operators into a series of MapReduce jobs.

**22. Explain the terms in the below syntax.EXPLAIN [-script pigscript] [-out path] [-brief] [-dot] [-paramparam\_name = param\_value] [-param\_filefile\_name] alias;**

script: It is used to specify a Pig script

* **out :** Used to specify the output path (directory)
* **brief :**Does not expand nested plans
* **dot :** outputs a format that can be passed to the dot utility for graphical display – will generate a directed-acyclic-graph (DAG) of the plans in any supported format (.gif, .jpg …).
* **Alias :** name of a relation.
* **paramparam\_name = param\_value :** used to see the parameters According to IBM, [processing your data is simple with Apache Pig](https://www.ibm.com/developerworks/library/l-apachepigdataquery/).

**23. What are all stats classes in the org.apache.pig.tools.pigstats package?**

Stat classes are in the package

* PigStats
* JobStats
* OutputStats
* InputStats.

##### What do we understand by PIG?

Pig, it is an Apache open-source project, which operates on Hadoop, providing the engine for the parallel data flow. It contains the language referred as pig Latin, which expresses the data flow. It consists of various operations like sort, joins, filter, etc. & is capable of scripting UDF (User Define Functions) for reading, writing, & processing. Pig uses Map Reduce & HDFS  for storing & the entire task for processing.

### 2. What is the difference in Pig and SQL?

\* Pig Latin shifts from SQL in a declarative style of encoding whereas Hive's query language is similar to SQL.  
\* Pig is above Hadoop and runs on principle, which can sit on top of Dryad too.  
\* Hive & Pig, both their commands collect to [MapReduce](https://mindmajix.com/hadoop/mapreduce-architecture-overview) jobs.

#### 3. Explain the requirement of MapReduce while we program in Apache Pig.

The programs of Apache Pig are written in a language referred as Pig Latin, which is analogous to SQL language. To carry out the query, we require an engine for execution. Pig engine alters all the queries to MapReduce tasks. Thus MapReduce operates as the primary execution engine needed to execute the programs.

##### 4. Explain BloomMapFile.

[BloomMapFile](https://hadoop.apache.org/docs/r1.0.4/api/org/apache/hadoop/io/BloomMapFile.Reader.html) is categorized as the class, which broadens MapFile class, and generally used for HBase table arrangement to speed up the relationship test for keys, which uses the filters of dynamic bloom.

##### 5. What is a bag in Pig?

A compilation of tuples is known as the bag, in Apache Pig.

##### 6. Why do we need the for each operation in Pig scripts?

The operation FOREACH in Apache Pig is required to apply to each component in data bag, for which the respective action can be performed to create data items.

##### 7. Explain the different data types in Pig.

Following are the three complex data types that is supported by Apache Pig:

\* Map, which is the key, value store, connected mutually using #.  
\* Tuples, similar to the row in the table, where a comma separates various items. Tuples may possess multiple attributes.  
\* Bags are a collection of tuples, in a unsynchronized manner, which allows many duplicate tuples.

##### 8. What is the function of Flatten in Pig?

Many times there are data in one of the tuple or bag which on removal, lead to next level of nesting for that data. In those cases, Flatten, a modifier, embedded in Pig is used. Flatten uninstalls bags & tuples and replaces all the areas in tuple, whereas the un-nesting bags are more complex of its need in creating a new tuple.

##### 9. What are describe & explain in Apache Pig scripts?

Explain & Describe are important utilities for debugging in [Apache Pig](https://mindmajix.com/hadoop/introduction-to-apache-pig).

Describe is helpful to all developers when scripting Pig because it displays the schema of the relation in a script. For developers, who are freshers & are learning Apache Pig use this utility to recognize the process of these operator making the modification to this data. Pig script has many describe.  
Explain utility is extremely helpful to developers of Hadoop, when they are trying to optimize Pig Latin scripts or debug error. Explain is applied on a specific alias in scripts or is applied on the entire script in the interactive shell of grunt. Explain utility creates many text graphs, which are printed to files.

##### 10. How does the user communicate with shell in Apache Pig?

Users interact with HDFS or any local file system through Grunt, which is the Apache Pig’s communicative shell. To initiate Grunt, users need to invoke the Apache Pig with a no command as follows:

\* Executing command “pig –x local” will prompt - grunt >  
\* Pig Latin scripts can run either in local mode or the cluster mode by setting up the configuration in PIG\_CLASSPATH.  
\* For exiting from grunt shell, users need to press CTRL+D or just key in the exit.

##### 11. What is a function of illustrate in Apache Pig?

Illustrate is used for implementing the scripts of Pig on vast sets of data, which generally is time-consuming. That’s why developers execute the scripts of pig on a sample data where it’s possible that the selected sample data, may not execute the script correctly. E.g., if the script consists of join operator then there must be few records in sample data which has the same key, or else join operation may not return the results. For managing these issues, developers use the function, illustrate, which takes a data from the sample and whenever it faces operators like the filter or join, which removes the data, it makes sure that some records go through whereas some are restricted, by modifying records in such so that they follow the condition set. Illustrate shows output of every step but does not execute MapReduce jobs.

##### 12. What do we know about case sensitivity of Pig?

Firstly, it is hard to find whether Pig is case sensitive or insensitive. E.g., in user-specific functions, field names, and relations in pig those are case sensitive. The function COUNT is not similar to the functions of count or X=load ‘foo’ is not similar to  x=load ‘foo.' Additionally, keywords in Pig are obviously case insensitive. E.g.  LOAD is similar to load.

###### 13. Distinguish between physical & logical plans in an Apache Pig script.

Physical & logical plans are generated while executing a pig script. Pig is based on the function of interpreter checking. The Logical plan is generated after the semantic verification & parsing while the processing of no data takes place in the generation of any logical plan. A consistent plan consists of a compilation of operators but does not consist of edges involving the operators. After generation of the logical plan, the execution of the script goes to physical plan. Physical plan is the explanation of physical operators, which Pig will use, for the execution of the script. It is more or less similar to a sequence of MapReduce works, but the plans don’t have any such reference of its execution in MapReduce. While the generation of any physical plan, the logical operator cogroup is transformed into physical operators, which are – Global Rearrange, Local Rearrange, and Package.

###### 14. Is Co-group is a group of more than 1 data set?

A group of data sets is referred to as Co-group. In any case, of more than one data set, co-group, groups all the data sets and then joins them based on a common field. That is why; we can say that co-group is obviously a group of more than one data set.

##### 15. Differentiate between HiveQL & PigLatin.

\* PigLatin is procedural language, whereas HiveQL is declarative.  
\* In HiveQL it is necessary to specify the schema, whereas in PigLatin it is optional.  
\* PigLatin has a nested relative data model, whereas HiveQL has a flat data model.

##### What are the uses of Apache Pig?

Pig big data tools, is specifically used for processing iteratively, for traditional ETL data pipelines & research on raw data. Pig operates in situations where the schema is unknown, incomplete, or inconsistent; it is used by all developers who want to use the data before being loaded into the data warehouse. For building prediction models for behavior, it is used by the website to detect the reply of visitors to a variety of images, ads, articles, etc.

##### 17. Is PigLatin strongly typed language?

Strongly typed language, is characterized where the user should state all the type of variables openly, whereas in Pig, the description of the data, it anticipates the data to approach in the mentioned format. If the schema is unknown, the script adapts to the actual data types at the runtime. That’s why it is stated that PigLatin might be strongly typed in many scenarios, but in some situations, it is otherwise gently typed. It keeps on working with the data, which may not be up to the expectations.

###### 18. Distinguish between COGROUP & GROUP operators.

A GROUP & COGROUP operator is same & work within one or many relations. Operator GROUP is usually used for grouping the data in any one single relation, for enhanced readability, while COGROUP is for gathering the data for 2 or higher relations. COGROUP is a mixture of JOIN & GROUP, i.e., it can group the tables, which are based on columns and joins them on grouped pieces. At any given time, cogroup can feature up to 127 relations.

###### 19. What do we understand by the outer bag and inner bag in Pig?

The outer bag is just any relation in Pig whereas sny relation within a bag is known as the inner bag.

###### 20. Differentiate between COUNT and COUNT\_STAR functions in Pig.

The Function COUNT\_STAR (0) comprises of NULL values as it counts, whereas COUNT function doesn’t include the NULL value when counting the number of elements in a bag.

###### 21. Do Pig support multi-line commands?

Pig supports single & multi-line commands, both. In the single line command, it carries out the data but doesn’t store the file in the system, but in multiple lines commands it stores the data in HDFS.

###### 22. If I have a relation R then how can I get top 10 tuples from the relation R?

Function TOP () returns the top (N) tuples from a relation or a bag of tuples. (N) Is passed as a constraint to function top () with the column, where the values are supposed to be evaluated in comparison to the relation R.

###### 23. How can we combine the contents of two or more relations & then divide them into a single relation into two or more relations?

The operation can be easily done by using the SPLIT and UNION operators.

##### 24. What are the various types of UDF’s in Java supported by Apache Pig?

Types of User Defined Functions supported in Pig are, Eval Algebraic and Filter functions are.

##### 25. What are the standard functionalities between Pig and Hive?

PigLatin and HiveQL both alter the commands to MapReduce work & cannot be used for transactions in OLAP as it is extremely difficult in executing queries of low latency.

##### 26. If we have a file employee.txt in the Hadoop Data File System directory with minimum 100 records, & want to see the first 25 records only from the employee.txt file. How can we do this?

Firstly we need to load the file employee.txt with the relation name as Employee. Then we can pull the first ten records of the data from the employee file by using the limit operator – Result = limit employee 25.

##### 27. What are the limitations of Pig Script?

Following are some of the Limitations of the Apache Pig a:

\*  Apache Pig isn’t preferable for analytics of a single record in huge data sets.  
\* Pig platform is specifically designed for ETL-type use cases, it’s not a good choice for synchronized or real time scenarios.  
\* Apache Pig is built on top of MapReduce, which is itself batch processing oriented.

###### 28. Can we join multiple fields in Apache Pig Scripts?

We can join multiple fields in PIG by the join operator, which extracts the records from any one input & joins them with the other specified input. This is done by specifying the keys for each input & both the rows will join as soon as the keys are equal.

##### 29. Why do we use Filters in Apache Pig ?

As the clause in SQL, Apache Pig has to filter for extraction of the records, which are based on predicate or specified condition. The records are then passed through the pipeline if the condition turns to true. Predicate surrounds a variety of operators like ==, <=,!=, >=. For instance - Y = filter X by symbol matches ‘Mr.\*’; X= load ‘inputs’ as(name,address)

##### 30. What is UDF in Pig?

If the Built in operators does not provide some of the basic functions, then developers can apply those functions by writing the user defined functions by using programming languages like Python, Java, Ruby, etc. (UDF’s) better known as User Defined Functions are then rooted into the Pig Latin Script.

**How to write Java UDF?**

UDFs can be developed by extending EvalFunc class and overriding execution method.  
Example: This UDF replaces a given string with another string

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27  28 | Package kelly.training.pig.udf;  Import java.io.IOException;  Import org.apache.hadoop.conf.configuration;  Import org.apache.pig.EvalFunc;  Import org.apache.pig.data Tuple;  Import org.apache.pig.impl.util.UDFContext;  Public classTransform extends EvalFunc{        Public string exec(Tuple input) throws IOException {             if(input == null || input.size[] == 0) {                      Return null;  }  Configuration conf=UDFContext.getUDFContext().getJobConf();  String from = conf.get(“replace.string”);  if(from == null){  Throw new IOException (“replace.string should not be null”);  }  String to = conf.get(“replace.by.string”);  if(to==null){  Throw new IOException (“replace.by.string should not be null”);  }  Try{  String str = (string) input.get(0);  Return str.replace(from, to);  } catch (exception e){  Throw new IOException(“caught exception processing input row”,e);  }  }  } |

**32. What is Grunt Shell?**

Grunt Shell is an interactive based shell. Which means where exactly we will get the output than and their itself. Whether it is success (or) fail.

**33. What is Pigstorage?**

Loads or stores relations using field delimited text format.  
Each line is broken into fields using a configurable field delimiter (defaults to a tab character) to be stored in the tuples fields. It is the default storage when none is specified.

**34. Where Does Pig Live?**

1. Pig is installed on user machine.  
2. No need to install anything on the hadoop cluster  
3. Pig and Hadoop versions must be compatible.  
4. Pig submits and executes jobs to the hadoop cluster

**35. Hive used for types of applications?**

1. Summarization  
Ex:- Daily/Weekly aggregations of impression/click counts  
2. Complex measure of user engagement  
3. Ad Hoc Analysis  
Ex:- How many group admins broken down by state/country  
4. Data Mining (Assembling Training Data)  
Ex:- User engagement as a function of user attributes.  
5. Spam Detection  
6. Anomalous patterns for site integrity.   
7. Application API usage patterns  
8. Ad Optimizations  
9. Document indexing   
10. Customer facing business intelligence (Ex: Google analytics) Predictive modeling, hypothesis testing

**36. What is Hive QL?**

1. Support SQL like query Language called HiveQL for select, join, aggregate, union all and subquery in the from clause.   
2. Support DDL statement such as CREATE table with serialization format, partitioning and bucketing columns.  
3. Command to load data from external sources and INSERT into HIVE tables.  
4. Do not support UPDATE and DELETE.  
5. Support multi table INSERT  
6. Support user defined column transformation (UDF) and aggregation (UDAF) function written in Java.

**37. What is the Difference Between Pig & SQL?**

|  |  |
| --- | --- |
| **Pig** | **SQL** |
| Pig is procedural | SQL is declarative |
| Nested relational data model | Flat relational data model |
| Schema is optional | Schema is required |
| Scan Centric analytytic workloads | OLTP + OLAA workloads |
| Limited query optimization | Significant opportunity for query optimization |

**38. What is the Difference Between Hive & Pig?**

|  |  |
| --- | --- |
| **Hive** | **Pig** |
| Language is SQL | Language is Pig Latin |
| Schema: Table definitions that are stored in a metastore | A schema is optionally defined at runtime |
| Hive programmatically access is JDBC, ODBC | Pig access is pigserver |
| The hive have partitions | There is no partitions |
| Server is optional | No server |
| Custom Serializer/ Deserializer | Custom Serializer/ Deserializer |
| DFS direct access at run time | DFS direct access at default |
| Join/order/set is possible | Join/order/set is possible |
| Shell command interface is possible | Shell command interface have |
| Streaming is supported | Streaming is supported |
| Web interface is possible | There is no web interface |

**39. What is the Difference Between Mapreduce & Pig?**

|  |  |
| --- | --- |
| **Mapreduce** | **Pig** |
| Mapreduce expects the programming language skills for writing the business logic | Pig there is no much of programming skills. As we are writing whole logic will making use of pig transformation (or) operations. |
| If we can do any change in the Mapreduce reduce program, we need to certain problems we can change the process entire. Compiling the program Executing the program Packing up the program Deploying the same cluster environment | In the pig, we can completes dealing with simple scripting we can avoid other transaction process. 5 % of the Mapreduce code 5% of the Mapreduce development time Increases programmer productivity 25% of the Mapreduce execution time |
| As a general saying of Hadoop Mapreduce program write 200 lines of mapreduce code. | In pig we can that type of Mapreduce program, we can write 10 lines of code. |
| Mapreduce requires multiple stages, Leading to long development life cycles Rapid prototyping increase productivity. | Pig provides the log analysis Ad Hoc queries across various large data sets. |