1. **List out Execution Modes of Pig**

Pig has two execution modes or exectypes:

* **Local Mode**

In this mode, all the files are installed and run from your local host and local file system. There is no need of Hadoop or HDFS. This mode is generally used for testing purpose.

$ pig –x local

* **MapReduce Mode**

MapReduce mode is where we load or process the data that exists in the Hadoop File System (HDFS) using Apache Pig. In this mode, whenever we execute the Pig Latin statements to process the data, a MapReduce job is invoked in the back-end to perform a particular operation on the data that exists in the HDFS.

$ pig -x mapreduce (or) pig

1. **List out the Features and Limitations of Hive.**

## Features of Hive

* It stores schema in a database and processed data into HDFS.
* It is designed for OLAP.
* It provides SQL type language for querying called HiveQL or HQL.
* It is familiar, fast, scalable, and extensible.
* Helps in processing unstructured data
* Fault Tolerance

**Limitations:**

|  |  |
| --- | --- |
| Does not support OLAP | Apache Hive doesn’t support online transaction processing (OLTP) but Online Analytical Processing(OLAP) is supported. |
| No updation and Deletion | Hive does not support update and delete operation on tables. |
| Doesn’t support subqueries | Subqueries are not supported. |
| Latency | The latency in the apache hive query is very high. |
| Only non-real or cold data is supported | Hive is not used for real-time data querying since it takes a while to produce a result. |

1. **Compare Hbase with RDBMS**

## HBase and RDBMS

|  |  |
| --- | --- |
| **HBase** | **RDBMS** |
| HBase is schema-less, it doesn't have the concept of fixed columns schema; defines only column families. | An RDBMS is governed by its schema, which describes the whole structure of tables. |
| It is built for wide tables. HBase is horizontally scalable. | It is thin and built for small tables. Hard to scale. |
| No transactions are there in HBase. | RDBMS is transactional. |
| It has de-normalized data. | It will have normalized data. |
| It is good for semi-structured as well as structured data. | It is good for structured data. |

1. **Define list and DataFrame in R**

## Lists

A list in R can contain many different data types inside it. A list is a collection of data which is ordered and changeable.

To create a list, use the list() function:

### **Example**

# List of strings  
thislist <- list("apple", "banana", "cherry")  
  
# Print the list  
print(thislist)

DATAFRAME:

**DataFrames** are generic data objects of R which are used to store the tabular data.

Data frames can also be taught as mattresses where each column of a matrix can be of the different data types. DataFrame are made up of three principal components, the data, rows, and columns.

1. **Define ordered and unordered factors**

An unordered factor is one where there is no inherent order to the levels, examples:

Color of car

Race

Nationality

Sex

State/Country of birth

Etc.

In the above, the order of the levels could be changed without it really changing the meaning (think of the order of bars in a bar chart). We may want to print/plot in some specific order such as alphabetic for easy lookup or based on the summary values of another vector for nice looking plots, but there is no overriding reason why we would order color as blue/green/red vs. green/red/blue, etc.

Ordered factors have some natural order, for example maybe you are studying a drug and have doses labeled as Low, Medium, and High. It makes the most sense to print and plot in that order rather than alphabetically (High, Low, Medium). Any continuous variable that has been cut into categories (best not to do this, but if done) has a natural order. Survey questions where you response can range from strongly disagree to strongly agree are usually ordered (but there may be disagreement on what the correct ordering is).

Set2:

1) **Define Macro in pig latin**

A macro is a block of code to do a particular task .

It can appear anywhere in a Pig script. We can develop more reusable scripts in Pig Latin Using Macros also. Macro is a kind of function written in Pig Latin.

2)**compare hbase with rdbms**

|  |  |
| --- | --- |
| **HBase** | **RDBMS** |
| HBase is schema-less, it doesn't have the concept of fixed columns schema; defines only column families. | An RDBMS is governed by its schema, which describes the whole structure of tables. |
| It is built for wide tables. HBase is horizontally scalable. | It is thin and built for small tables. Hard to scale. |
| No transactions are there in HBase. | RDBMS is transactional. |
| It has de-normalized data. | It will have normalized data. |
| It is good for semi-structured as well as structured data. | It is good for structured data. |

3) **Explain about the different types of partitioning in Hive?**

* Static Partitioning

In static or manual partitioning, it is required to pass the values of partitioned columns manually while loading the data into the table. Hence, the data file doesn't contain the partitioned columns.

* Dynamic Partitioning

In dynamic partitioning, the values of partitioned columns exist within the table. So, it is not required to pass the values of partitioned columns manually.

**4) Define clustering? Give applications**

[Clustering](https://www.geeksforgeeks.org/clustering-in-machine-learning/) in [R](https://www.geeksforgeeks.org/introduction-to-r-programming-language/) is an unsupervised learning technique in which the data set is partitioned into several groups called as clusters based on their similarity. Several clusters of data are produced after the segmentation of data. All the objects in a cluster share common characteristics. During data mining and analysis, clustering is used to find the similar datasets.

#### Applications of Clustering in R

**Marketing**

Medical Science

Gaming

Internet

**5) List out Regression types**

**\* Linear Regression**

\* **Logistic Regression**

\* **Multinomial Logistic Regression**

\* **Ordinal Logistic Regression**