1. Why Map-reduce program is needed in Pig Programming?

ans:- *PIG is a data flow language, the key focus of Pig is manage the flow of data from input source to output store.*

*As part of managing this data flow it moves data feeding it to process1, taking output and feeding it to process2.*

*The core features are preventing execution of subsequent stages if previous stage fails, manages temporary storage of data and most*

*importantly compresses and rearranges processing steps for faster processing. While this can be done for any kind of processing tasks*

*Pig is written specifically for managing data flow of Map reduce type of jobs. Most if not all jobs in a Pig are map reduce jobs or data movement jobs.*

*Pig allows for custom functions to be added which can be used for processing in Pig, some default ones are like ordering, grouping, distinct, count etc.*

*Map reduce on the other hand is a data processing paradigm, it is a framework for application developers to write code in so that its easily scaled*

*to PB of tasks, this creates a separation between the developer that writes the application vs the developer that scales the application.*

*Not all applications can be migrated to Map reduce but good few can be including complex ones like k-means to simple ones like counting uniques in a dataset.*

2. What are advantages of pig over MapReduce?

ans:- In simple terms Map Reduce is low level of programming and Pig is a high-level language for expressing data analysis programs which internally create sequence of Map Reduce Programs.

Pig is simple to learn and use as compared to Map Reduce.

Pig data flow language i.e pig Latin. For MapReduce, Java is by default supported programming language. However support for other language is also available.

Pig provides inbuilt optimization for MR jobs whereas in map reduce developer needs to take care of optimization.

Here are detail definition from official documentation:

Hadoop MapReduce is a software framework for easily writing applications which process vast amounts of data (multi-terabyte data-sets) in-parallel on large clusters (thousands of nodes) of commodity hardware in a reliable, fault-tolerant manner.

Pig's infrastructure layer consists of a compiler that produces sequences of Map-Reduce programs, for which large-scale parallel implementations already exist (e.g., the Hadoop subproject)

3. What is pig engine and what is its importance?

Ans:- Pig engine acts as interpreter between Pig latin script and mapreduce jobs.It creating environment to execute pig scripts into series of mapreduce jobs in parallel manner.

4. What are the modes of Pig execution?

Ans :- 1.MapReduce/Hadoop Mode: Here Pig jobs run as a series of MapReduce jobs picking the input and output paths from HDFS. Note: Input file has to be copied in HDFS in case of Map reduce mode. Type the command **pig or pig –x mapreduce** to run Pig in MapReduce Mode. While running pig in MapReduce mode, make sure job history server is running. mr-jobhistory-daemon.sh start historyserver Job history server helps us to view previous MR job details

2. Local Mode: Here the entire Pig job runs as a single JVM picking the local Unix path for execution. Note: Input file has to be kept in local file system in case of local mode. Type the command **pig or pig -x local to run Pig in Local Mode.**

5. What is grunt shell in Pig?

Ans :- It is the mode where pig scripts runs.

6. What are the features of Pig Latin language?

Ans :- They are easy to understand.

Debugging is easy as compared to mapreduce jobs .

Less complex

7. Is Pig latin commands case sensitive?

Ans :- no pig latin are not case sensitive ,but PigStorage is case-sensitive.

8. What is a data flow language?

Ans:- In computer programming, **dataflow** programming is a programming paradigm that models a program as a directed graph of the **data flowing** between operations, thus implementing **dataflow** principles and architecture.