**Pig Interview Tips:**

**Can you give some examples as how Hadoop is being used in the realtime environment?**  
 - Hadoop is being used by Walmart to study customer behavior  
 - Telecom companies are using hadoop to understand the churn behavior by analyzing the CDR records  
 - Wind companies(VESTAS) are using Hadoop technologies to define locations  
 - Healthcare companies are using hadoop to analyze data from varied sources  
  
**What is PIG?**  
Pig is platform for analyzing large data set. It allows user to analyze large unstructured dataset by transforming, and applying functions on the dataset. Pig interpreter will be converted into mapreduce internally by the jobtracker.There are four stages of Pig Query (Pig Latin)

Load

Transform (Group/Filter)

Dump

Store

A sample Pig query looks like this.;

**Does ILLUSTRATE run MR job?**  
No-Illustrate doesn't run the mapreduce job. It just shows that how the data would look when map reduce job is run.Illustrate will not pull any MR, it will pull the internal data. On the console, illustrate will not do any job. It just showsoutput of each stage and not the final output.  
  
  
**Is the keyword DEFINE like a function name?**

Yes, the keyword ‘DEFINE’ is like a function name. Once you have registered, you have to define it. Whatever logic you have written in Java program, you have an exported jar and also a jar registered by you. Now the compiler will check the function in exported jar. When the function is not present in the library, it looks into your jar.  
  
**Is the keyword “FUNCTIONAL” a user defined function?**

No, the keyword ‘FUNCTIONAL’ is not a User Defined Function (UDF). While using UDF, we have to override some functions. Certainly you have to do your job with the help of these functions only. But the keyword ‘FUNCTIONAL’ is a built-in function i.e a pre-defined function, therefore it does not work as a UDF.  
  
**Why do we need MapReduce during a PIG programming?**  
Pig is a high-level platform that makes many Hadoop data analysis issues easier to execute. The language we use for this platform is: Pig Latin. A program written in Pig Latin is like a query written in SQL, where we need an execution engine to execute the query. So, when a program is written in Pig Latin, Pig compiler will convert the program into MapReduce jobs. Here, MapReduce acts as the execution engine.

**Are there any problem which can be solved only by using MapReduce job and cannot be solved by PIG? In what kind of scenario MR Job will be more useful than PIG?**  
**There is no custom partitioner in pig.**Let us take a scenario where we want to count the population in two cities. I have a data set and sensor list of different cities. I want to count the population by using one mapreduce for two cities. Let us assume that one is Bangalore and the other is Noida. So I need to consider key of Bangalore city similar to Noida through which I can bring the population data of these two cities to one reducer. The idea behind this is some how I have to instruct map reducer program – whenever you find city with the name ‘Bangalore‘ and city with the name ‘Noida’, you create the alias name which will be the common name for these two cities so that you create a common key for both the cities and it get passed to the same reducer. For this, we have to write custom partitioner.  
  
In mapreduce when you create a ‘key’ for city, you have to consider ’city’ as the key. So, whenever the framework comes across a different city, it considers it as a different key. Hence, we need to use customized partitioner. There is a provision in mapreduce only, where you can write your custom partitioner and mention if city = bangalore or noida then pass similar hashcode. However, we cannot create custom partitioner in Pig. As Pig is not a framework, we cannot direct execution engine to customize the partitioner. In such scenarios, MapReduce works better than Pig.  
**Does PIG give any warning when there is type mismatch or missing field?**  
No, Pig will not show any warning if there is no matching field or a mismatch. If you assume that Pig gives such a warning, then it is difficult to find in log file. If any mismatch is found, it assumes a null value in Pig.  
**What co-group does in Pig?**  
Co-group 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 record of the first data set with the common data set and the second bag consists of the records of the second data set with the common data set.  
 **Can we say that co-group is a group of more than one dataset?**  
Cogroup is a group of one data set. But in the case of more than one data sets, cogroup will group all the data sets and join them based on the common field. Hence, we can say that cogroup is a group of more than one data set and join of that data set as well.  
**What does FOREACH do?**  
FOREACH is used to apply transformations to the data and to generate new data items. The name itself is indicatingthat for each element of a data bag, the respective action will be performed.  
  
Syntax : FOREACH bagname GENERATE expression1, expression2, …..  
The meaning of this statement is that the expressions mentioned after GENERATE will be applied to the current record of the data bag.  
**What is bag?**  
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 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 “{}”.

How to run pig in local mode?

In local mode , Pig runs in a single JVM and accesses the local filesystem . this mode is suitable only for small datasets .

% pig -x local

How to run pig in MapReduce mode?

% pig -x mapreduce

Or

% pig

In this mode , Pig translates queries into MapReduce jobs and runs them on a Hadoop cluster. The cluster may be a pseudo or fully distributed cluster.

Write some example for loading data from local file system?

grunt> Test = load ‘/home/training/climate.txt’ as (year:chararray , temp:int, city:chararray);

Write some example for loading data from HDFS?

grunt> Test = load ‘/user/training/climate.txt’ as (year:chararray, temp:int , city:chararray);

How to display schema of bag?

grunt > describe <bagname> ;

How to display list of tuples from a bag?

grunt > Dump <bagname>;