Pig Interview Questions

**1)what is pig?**  
Pig is a Apache open source project which is run on hadoop,provides engine for data flow in parallel on hadoop.It includes language called pig latin,which is  for expressing these data flow.It includes different operations like joins,sort,filter ..etc and also ability to write UserDefine Functions(UDF) for proceesing and reaing and writing.pig uses both HDFS and MapReduce i,e storing and processing.

**2)what is differnce between pig and sql?**  
Pig latin is procedural version of SQl.pig has certainly similarities,more difference from sql.sql is a query language for user asking question in query form.sql makes answer for given but dont tell how to answer the given question.suppose ,if user want to do multiple operations on tables,we have write maultiple queries and also use temporary table for storing,sql is support for subqueries but intermediate we have to use temporary tables,SQL users find subqueries confusing and difficult to form properly.using sub-queries creates an inside-out design where the first step in the data pipeline is the innermost query .pig is designed with a long series of data operations in mind, so there is no need to write the data pipeline in an inverted set of subqueries or to worry about storing data in temporary tables.

**3)How Pig differs from MapReduce**  
In mapreduce,groupby operation performed at reducer side and filter,projection can be implemented in the map phase.pig latin also provides standard-operation similar to mapreduce like orderby and filters,groupby..etc.we can analyze pig script and know data flows ans also early to find the error checking.pig Latin is much lower cost to write and maintain thanJava code for MapReduce.

pig interview questions

**4)How is Pig Useful For?**  
In three categories,we can use pig .they are 1)ETL data pipline 2)Research on raw data 3)Iterative processing   
Most common usecase for pig is data pipeline.Let us take one example, web based compaines gets the weblogs,so before storing data into warehouse,they do some operations on data like cleaning and aggeration operations..etc.i,e transformations on data.

**5)What are the scalar datatypes in pig?**  
scalar datatype  
int    -4bytes,  
float  -4bytes,  
double -8bytes,  
long   -8bytes,  
chararray,  
bytearray

**6)What are the complex datatypes in pig?**  
**map:**  
map in pig is chararray to data element mapping where element have pig data type including complex data type.  
example of map  [‘city’#’hyd’,’pin’#500086]  
the above example city and pin are data elements(key) mapping to values  
**tuple:**  
tuple have fixed length and it have collection datatypes.tuple containing multiple fields and also tuples are ordered.  
example, (hyd,500086) which containing two fields.  
**bag:**  
A bag containing collection of tuples which are unordered,Bag constants are constructed using braces, with tuples in the bag separated by com-  
mas. For example, {(‘hyd’, 500086), (‘chennai’, 510071), (‘bombay’, 500185)}

**7)Whether pig latin language is  case-sensitive or not?**  
pig latin is some times not a case sensitive.let us see example,Load is equivalent to load.  
A=load ‘b’ is not equivalent to a=load ‘b’  
UDF are also case sensitive,count is not equivalent to COUNT.

**8)How should ‘load’ keyword is useful in pig scripts?**  
first step in dataflow language we need to specify the input,which is done by using ‘load’ keyword.load looks for your data on HDFS in a tab-delimited file using the default load function ‘PigStorage’.suppose if we want to load data from hbase,we would use the loader for hbase  
‘HBaseStorage’.  
example of pigstorage loader  
A = LOAD ‘/home/ravi/work/flight.tsv’ using PigStorage (‘t’) AS (origincode:chararray, destinationcode:chararray, origincity:chararray, destinationcity:chararray, passengers:int, seats:int, flights:int, distance:int, year:int, originpopulation:int, destpopulation:int);  
example of hbasestorage loader  
x= load ‘a’ using HBaseStorage();  
if dont specify any loader function,it will takes built in function is ‘PigStorage’   
the ‘load’ statement can also have ‘as’ keyword for creating schema,which allows you to specify the schema of the data you are loading.  
PigStorage and TextLoader, the two built-in Pig load functions that operate on HDFS files.

**9)How should ‘store’ keyword is useful in pig scripts?**  
After we have completed process,then result should  write into somewhere,Pig provides the store statement for this purpose  
store processed into ‘/data/ex/process';  
If you do not specify a store function, PigStorage will be used. You can specify a different store function with a using clause:  
store processed into ‘?processed’ using HBaseStorage();  
we can also pass argument to store function,example,store processed into ‘processed’ using PigStorage(‘,’);

**10)What is the purpose of ‘dump’ keyword in pig?**  
dump diaplay the output on the screen  
dump ‘processed’

**11)what are relational operations in pig latin?**  
they are  
a)for each  
b)order by  
c)filters  
d)group  
e)distinct  
f)join  
g)limit

**12)How to use ‘foreach’ operation in pig scripts?**  
foreach takes a set of expressions and applies them to every record in the data pipeline  
A = load ‘input’ as (user:chararray, id:long, address:chararray, phone:chararray,preferences:map[]);  
B = foreach A generate user, id;  
positional references are preceded by a $ (dollar sign) and start from 0:  
c= load d generate $2-$1

**13)How to write ‘foreach’ statement for map datatype in pig scripts?**  
for map we can use hash(‘#’)  
bball = load ‘baseball’ as (name:chararray, team:chararray,position:bag{t:(p:chararray)}, bat:map[]);  
avg = foreach bball generate bat#’batting\_average';

**14)How to write ‘foreach’ statement for tuple datatype in pig scripts?**  
for tuple we can use dot(‘.’)  
A = load ‘input’ as (t:tuple(x:int, y:int));  
B = foreach A generate t.x, t.$1;

**15)How to write ‘foreach’ statement for bag datatype in pig scripts?**  
when you project fields in a bag, you are creating a new bag with only those fields:  
A = load ‘input’ as (b:bag{t:(x:int, y:int)});  
B = foreach A generate b.x;  
we can also project multiple field in bag  
A = load ‘input’ as (b:bag{t:(x:int, y:int)});  
B = foreach A generate b.(x, y);

**16)why should we use ‘filters’ in pig scripts?**  
Filters are similar to where clause in SQL.filter which contain predicate.If that predicate evaluates to true for a given record, that record will be passed down the pipeline. Otherwise, it will not.predicate contain different operators like ==,>=,<=,!=.so,== and != can be applied to maps and tuples.  
A= load ‘inputs’ as(name,address)  
B=filter A by symbol matches ‘CM.\*';

**17)why should we use ‘group’ keyword in pig scripts?**  
The group statement collects together records with the same key.In SQL the group by clause creates a group that must feed directly into one or more aggregate functions. In Pig Latin there is no direct connection between group and aggregate functions.  
input2 = load ‘daily’ as (exchanges, stocks);  
grpds = group input2 by stocks;

**18)why should we use ‘orderby’ keyword in pig scripts?**  
The order statement sorts your data for you, producing a total order of your output data.The syntax of order is similar to group. You indicate a key or set of keys by which you wish to order your data  
input2 = load ‘daily’ as (exchanges, stocks);  
grpds = order input2 by exchanges;

**19)why should we use ‘distinct’ keyword in pig scripts?**  
The distinct statement is very simple. It removes duplicate records. It works only on entire records, not on individual fields:  
input2 = load ‘daily’ as (exchanges, stocks);  
grpds = distinct exchanges;

**20)is it posible to join multiple fields in pig scripts?**  
yes,Join select records from one input and join with another input.This is done by indicating keys for each input. When those keys are equal, the two rows are joined.   
input2 = load ‘daily’ as (exchanges, stocks);  
input3 = load ‘week’ as (exchanges, stocks);  
grpds = join input2 by stocks,input3 by stocks;

we can also join multiple keys  
example:  
input2 = load ‘daily’ as (exchanges, stocks);  
input3 = load ‘week’ as (exchanges, stocks);  
grpds = join input2 by (exchanges,stocks),input3 by (exchanges,stocks);

**21)is it possible to display the limited no of results?**  
yes,Sometimes you want to see only a limited number of results. ‘limit’ allows you do this:  
input2 = load ‘daily’ as (exchanges, stocks);  
first10 = limit input2 10;