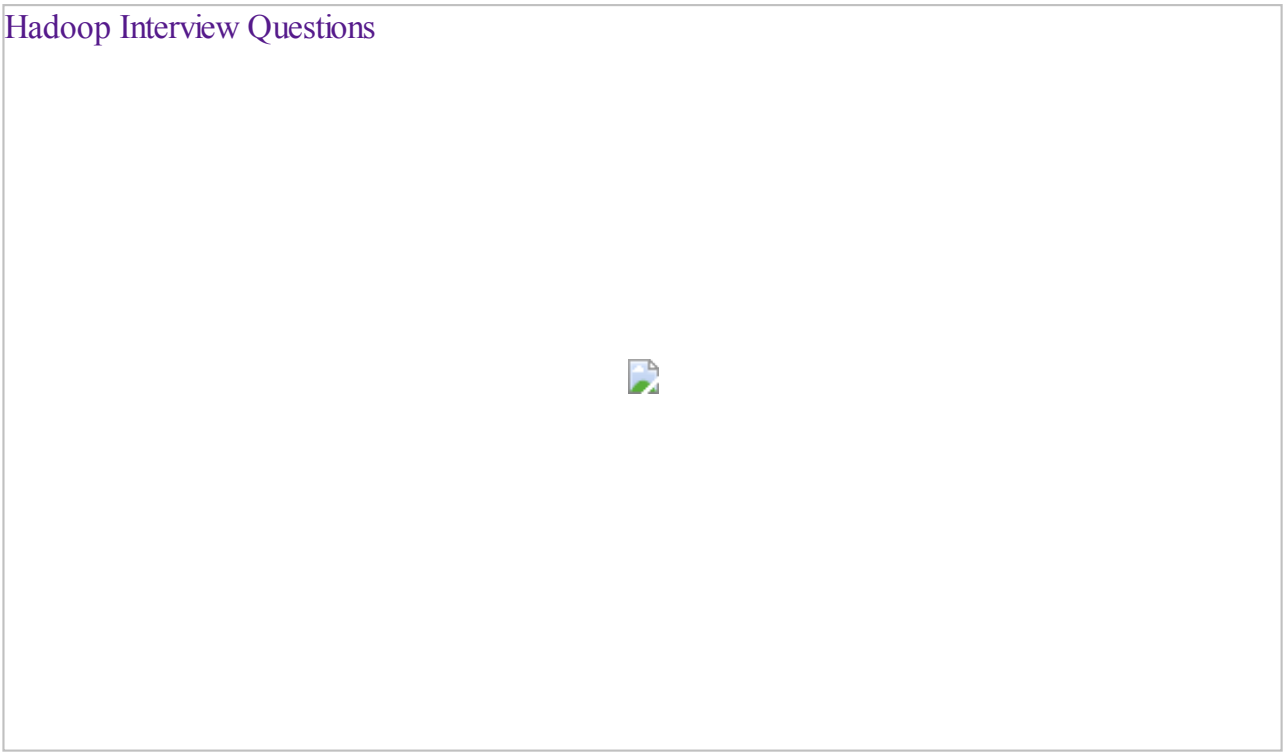
edureka!

Hadoop Interview Questions – PIG

April 25, 2013 | Big Data and Hadoop, Interview Questions



Looking out for Hadoop Interview Questions that are frequently asked by employers?
Here is the fourth list of Hadoop Interview Questions which covers PIG…

Can you give us some examples how Hadoop is used in real time environment?

Let us assume that the we have an exam consisting of 10 Multiple-choice questions and 20 students appear for that exam. Every student will attempt each question. For each question and each answer option, a key will be generated. So we have a set of *key-value pairs* for all the questions and all the answer options for every student. Based on the options that the students have selected, you have to analyze and find out how many students have answered correctly. This isn't an easy task. Here Hadoop comes into picture! Hadoop helps you in solving these problems quickly and without much effort. You may also take the case of how many students have wrongly attempted a particular question.

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.

What is PIG?

PIG is a platform for analyzing large data sets that consist of high level language for expressing data analysis programs, coupled with infrastructure for evaluating these programs. PIG's infrastructure layer consists of a compiler that produces sequence of MapReduce Programs.

What is the difference between logical and physical plans?

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.

Does 'ILLUSTRATE' run MR job?

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

word 'DEFINE' like a function name?

yword 'DEFINE' is like a function name. Once you have registered, you have to define it. gic you have written in Java program, you have an exported jar and also a jar registered by you. mpiler will check the function in exported jar. When the function is not present in the library, it our jar.

word 'FUNCTIONAL' a User Defined Function (UDF)?

ord 'FUNCTIONAL' is not a User Defined Function (UDF). While using UDF, we have to override some rtainly you have to do your job with the help of these functions only. But the keyword 'FUNCTIONAL' is a on i.e a pre-defined function, therefore it does not work as a UDF.

need MapReduce during Pig programming?

-level platform that makes many Hadoop data analysis issues easier to execute. The language we platform is: *Pig Latin*. A program written in *Pig Latin* is like a query written in SQL, where we cution engine to execute the query. So, when a program is written in Pig Latin, Pig compiler will program into MapReduce jobs. Here, *MapReduce* acts as the execution engine.

Are there any problems which can only be solved by MapReduce and cannot be solved by PIG? In which kind of scenarios MR jobs will be more useful than 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 a 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 cogroup is a group of more than 1 data set?

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 indicating that 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 "{}".

To refer to the first list, click Hadoop Interview Questions – HDFS!
To refer to the second list, click Hadoop Interview Questions – Setting up Hadoop Cluster!
To refer to the third list, click Hadoop Interview Questions – MapReduce!

- **Search**

Recent Posts

- Free Webinar on 'Introduction to MongoDB'
- Importance of Data Science With Cassandra
- How to become a Hadoop Administrator?
- Why Big Data Professionals Need to Learn MongoDB?
- Why Should you go for Hadoop Administration Course?

Categories

- Android
- Apache Cassandra
- Apache Storm
- Big Data and Hadoop
- Business Analytics With R
- Cloud Computing
- Data Science
- Hadoop Administration
- Interview Questions
- Java
- MongoDB
- PMI-ACP
- PMP Exam Preparation
- Python for Big Data Analytics
- Resources & Misc.
- Uncategorized
- Webinars

Subscribe

RSS Feed