coreservlets.com – Hadoop Course

Pig Advanced

In this exercise, you will have a chance to develop several Pig scripts. Your Pig scripts will join the data as well as utilized User Defined Function (UDF) that you will get a chance to implement.

Approx. Time: 60 minutes

Perform

1. Implement Pig script that will generate a file on HDFS that will contain a report of books sold and their buyers. Consider the following inputs:

   ```
   /training/exercises/pig/book-purchases.txt
   /training/exercises/pig/books.txt
   ```

   Store the resulting file on HDFS in the following location

   ```
   /training/exercises/pig/bookPurchases
   ```

   Please note that input files use different column separator, the first uses tab (‘\t’) and the second a comma.

   See if you can implement this solution using different operators. What is the major difference?

2. Implement a Pig script that will display all the book records whose title contains word ‘Hadoop’. To accomplish this task first implement User Defined Function (UDF) Filter which will allow you to select values that contain word ‘Hadoop’. Then implement and test pig script that utilizes your UDF. Use the following file as an input:

   ```
   /training/exercises/pig/books.txt
   ```

   Print the result to screen.

Solution

1. You can implement the report using JOIN or COGROUP operators. Both solutions reside in the Solutions project:

   JOIN solution: `src/main/resources/pig/BookJoin.pig`

   COGROUP solution: `src/main/resources/pig/BookCogroup.pig`

   You will find that the code is very similar but the main difference is the output produced. JOIN will produce rows with tuples from both inputs without preserving any structure. COGROUP will preserve the structure of each input in its own bag. JOIN creates a row for each match; in our case, this is 1-to-many join where 1 book matches to multiple purchases. In the case of join for each match there will be a new row therefore JOIN will produce a row for each record in "many" set. COGROUP produces a row for each unique key with 2 bags, 1 for each input. In our case, COGROUP will produce a record for each record in books.txts

   The choice of an operator will depend on the use case.

2. The solution can be found in the Solutions project

   Filter UDF: `pig.advanced.IsHadoop.java`

   Pig Script: `src/main/resources/pig/IsHadoopFilter.pig`

   To run the pig script build the project ‘mvn package’; this will create a jar file and copy into Pig’s scripts directory. You can then execute the script by:

   ```
   $ cd $PLAY_AREA/pig/scripts-solutions
   $ pig IsHadoopFilter.pig
   ```