Lesson 4 Hands-On

Directions

This Hands-On will **not** be graded, but you are encouraged to complete it. The best way to become a great data scientist is to practice. Once you have submitted your project, you will be able to access the solution on the next page. Note that the solution will be slightly different from yours, but should look similar.

Caution!

Do not submit your project until you have completed all requirements, as you will not be able to resubmit.

Description

Using the Book data files, write a Pig Query that will find all the books that have a 1-star rating. Include a file with your Pig Query and a screenshot of your results.

Alternative Assignment if You Can't Run Hadoop and/or Ambari

If your computer refuses to run Hadoop and/or Ambari, **here** is an alternative exam to test your understanding of the material. Please attach it instead.

Lesson 4 Hands-On Solution

```
ratings = LOAD '/user/maria_dev/books_data/books.csv' USING
PigStorage(',')
     AS (bookID:int, authors:chararray, average_rating:float,
isbn:chararray, isbn13:chararray, language_code: chararray,
num_pages:int, ratings_count: int, text_reviews_count:int);

metadata = LOAD '/user/maria_dev/books_data/bookIDs.csv' USING
PigStorage(',')
     AS (bookID: int, title: chararray);

nameLookup = FOREACH metadata GENERATE bookID, title;

groupedRating = GROUP ratings by bookID;

averageRatings = FOREACH groupedRating GENERATE group AS bookID,
     AVG(ratings.average_rating) AS avgRating,
COUNT(ratings.average_rating) AS numRatings;

badBooks = FILTER averageRatings BY avgRating < 2.0;

namedBadBooks = JOIN badBooks BY bookID, nameLookup BY bookID;

finalResults = FOREACH namedBadBooks GENERATE nameLookup::title
AS bookName,
     badBooks::avgRating as avgRating, badBooks::numRatings as
numRatings;

finalResultsSorted = ORDER finalResults By numRatings DESC;

DUMP finalResultsSorted;
```

Lesson 4 Hands-on Solution - Alternative Assignment

This exam serves as the assessment for those students who cannot utilize the Hadoop system and/or Ambari GUI. Answers are shown in bold.

1. Explain the concept of TEZ and how it works.
   TEZ makes jobs on Hadoop go faster. It runs off of directed acyclic graphs, or DAGs, which find the most efficient path for your work to be conducted. It removes unnecessary steps and figures out what can be run in parallel. These efficiencies save both time and money!
2. What is Pig's version of the for loop? a. Group b. Filter c. Join **d. For each / generate**
3. How is HBase structured? Describe or draw – whatever makes the most sense to you.
   HBase uses key-value pairs, with each row having a unique key. But, it also makes use of column families, so that you can have a little more granularity. A cell is the intersection of a column and a row, and you can store multiple versions of cell data and even timestamp it!