JAMES COOK UNIVERSITY
AUSTRALIA

MA5831 Advanced Data Processing and Analysis using SAS

Assignment 4 Content

Name:	Managing data with Hive and Pig in Hadoop
Type:	Case study
Issued:	8:00 PM AEST Monday of Week 1
Due:	11.59 PM AEST Wednesday of Week 7
Weight:	20%
Length:	1500 words +/-10%

Submission method: PDF or MS Word submission to LearnJCU

Overview

This case study reinforces the concepts that you learned in Weeks 4, 5 and 6, and the practical work you carried out in the SAS course ‘Big Data Programming and Loading’. Specifically, it reinforces working with Hive tables, and writing and executing Pig scripts on a Hadoop environment for big data.

The data used in this case study comes from a retail bank’s customer call centre or through a retail bank’s customer website. These complaints come from across the United States regarding the bank’s various financial product offerings. Your aim is to generate the appropriate queries and scripts to answer the tasks posed under the tasks section with supporting evidence.

The case study, along with a video tutorial on data setup, can be found on the SAS Academy for Data Science in the Case Studies section of Module 2: Big Data Programming and Loading:

- Case Study: Hadoop Data Management with Hive, Pig and SAS (pdf)

Learning outcomes

This assessment evaluates your ability to:

1. Load the data from the Windows location into HDFS to support the HiveQL and Pig queries
2. Create and combine Hive Tables
3. Sort and Report on Hive Tables
4. Write and execute Pig scripts in a Hadoop environment.

Work-based skills

This assessment relates to the following work-based skills:

1. Ability to work with big unstructured data e.g. (text) to read and process it for analysis
2. Critically appreciate the use of a Hadoop in a real world scenario.

The case study

The problem

A national bank associated with the College of Engineering has a Consumer Complaints Department. Every month, the department is asked to generate reports of consumer complaints. The reports are based on the bank's financial products, such as consumer loans, debt collection, credit reporting, mortgage, and so on, for the various states in which it operates. The reports are analysed and reviewed by the bank to find where improvements can be made to the services they provide. To support the report templates, the consumer-complaints data needs to be divided into subsets of data and reside in the Hadoop Distributed File System (HDFS).

The data

There is a step-by-step video guide to connecting to the data in the overview section of the case study in the SAS Academy for Data Science. You should thoroughly review the consumer complaints text file before you begin.

The information collected consists of the fields that are specified in the following data dictionary table. Each row is a unique consumer complaint with an assigned complaint ID.

Field	Description
Complaint ID	Automatically generated complaint ID number that is used to track each consumer complaint
Product	Bank's primary financial-product offering category, such as consumer loan, mortgage, debt collection, and so on
Sub-product	Bank's secondary financial-product offering category, which is used to subset the type of financial product, such as vehicle loan, medical, other (phone, health club), and so on
Issue	Primary reason that the customer filed the complaint, such as checking (cheque) account charges
Sub-issue	Secondary reason that the customer filed the complaint, such as problems caused by low funds in a checking (cheque) account
State	Consumer's two-byte state code as validated by the U.S. Postal Service
ZIP Code	Consumer's postal code, per the bank's consumer address database
Submitted Via	How the complaint was entered, such as web, phone, referral
Date Received	When the consumer complaint was received by the bank
Date Sent to Company	When the consumer complaint was sent to the bank
Company Response	Status of the consumer complaint at the bank, such as in progress, closed with explanation, and so on
Timely Response?	Did the bank respond to the consumer complaint in a timely manner, according to the customer? Yes or No?
Consumer Disputed?	Was the bank's response to the consumer complaint disputed by the customer? Yes or No?

The tasks

Technical preparation and setup

The case study assignment will be completed using the SAS Academy for Data Science with the data loaded in the Hadoop File System location from Week 5 (Self Learning Practical : Hadoop and SAS Chapter 1: The Apache Hadoop Project > Hadoop Essentials > Exercise: Loading Data onto the Hadoop Name Node).

Open an mRemoteNG session as the student to student.ClientNode

Type **ls /home/student/DIHPS/data/consumer_complaints.txt** to verify if the data exists.

If data does not exist, using WinSCP UPLOAD **D:/workshop/DIHPS/data/consumer_complaints.txt** to **/home/student/DIHPS/data**

Type **hdfs dfs –mkdir –p /user/student/DIHPS/consumer** to create the appropriate folders, if they do not already exist.

Type **hdfs dfs –put /home/student/DIHPS/data/consumer_complaints.txt /user/student/DIHPS/consumer** to copy the data to hdfs.

For further details about the SAS Virtual Lab, please refer to 'Accessing SAS® Software Using the Virtual Lab Reservation System' (pdf) in the SAS Academy for Data Science.

Data processing tasks

Using HiveQL, you query the consumer-complaint table to respond to the tasks posed in the Reporting tasks section. Then, using Pig Latin, you generate the files that are necessary to support the consumer-complaint reports. You can use the Hive and Pig command-line tools to submit your code, or you can use the Hive and Pig Query Editors in Hue.

Specifically, you need to do the following tasks:

1. Review the contents of consumer_complaints.txt and create a Hive table schema that is appropriate for the table properties to support subsequent query operations using HiveQL.
2. Create HiveQL queries that answer the tasks in the Reporting tasks section and review the information for accuracy.
3. Use Pig Latin scripts to generate the actual tables in the /user/student/BBDA/output HDFS location to support the consumer-complaint reports templates.

Reporting and coding tasks

Your report needs to include the following tasks:

1. Using a Hive query, determine the 10 states with the maximum number of complaints.
2. Using a Hive query, determine how many complaints are associated with the “Medical” sub-product offering.
3. Using a Hive query, determine the five ZIP codes with the smallest number of complaints.
4. Using a Hive query, determine how many complaints, grouped by Product and State, are associated with the word *fraud* in the issue description.
5. Using a Hive query, create a new table that summarises the total number of complaints by Product, State, and Submitted_Via fields. Group and display the columns for Product, State, and Submitted_Via fields.
6. Using a Pig Latin script, create two tables, *web_results* and *other_results*, that separate the consumer complaints that were entered via the web submission form from other submission forms. Verify that the correct number of records are written to the appropriate tables.
7. Using Pig Latin, create a table, *max_complaints*, that lists the 10 states with the maximum number of complaints.
8. Using a Pig Latin script, create a table, *medical_complaints_list*, that lists complaints that are associated with the “Medical” sub-product offering. Then, create a table, *medical_complaints_total*, with the total count of medical complaints received to determine how many consumer complaints are associated with the ‘Medical’ sub-product offering.
9. Using a Pig Latin script, create a table, *least_complaints*, that lists the five ZIP codes with the smallest number of complaints.
10. Using a Pig Latin script, create a table, *id_theft_complaints*, grouped by Product and State fields, that lists complaints associated with identity theft issues.
11. Using a Pig Latin script, create a summary table, *complaint_summary*, grouped by Product, Sub-product, and State fields, that lists all complaints that were received per state.

Assessment criteria

Written responses, with supporting code and output evidence, for the reporting tasks, submitted to LearnJCU. See marking rubric for further details (20%).

Examples of supporting evidence, including:

- Hive queries
- Pig Latin script files.

Submission guidelines

Your submission for Assessment 4 should be uploaded to LearnJCU as three (3) separate files:

1. The task cover sheet
2. Completed answers for the tasks in the Reporting tasks section and saved in the following format A4_firstname_lastname (PDF or DOCX)
    - Length: 1500 words +/-10%
3. Completed Pig Latin scripts or Hive queries used in the Data processing tasks section.

Marking Criteria: MA5831 Assessment 4 – Case study: Managing data with Hive and Pig in Hadoop

Criteria	High Distinction	Distinction	Credit	Pass	Fail
Criterion 1: Student has an understanding of the overall scripting and querying of big data with the level completeness, readability and execution in their work.	Student demonstrates an excellent understanding of the overall scripting and querying of big data with an excellent level of completeness, readability and execution in their work.	Student demonstrates a strong understanding of the overall scripting and querying of big data with a strong level of completeness, readability and execution in their work.	Student demonstrates a developing understanding of the overall scripting and querying of big data with a developing level of completeness, readability and execution in their work.	Student demonstrates a basic understanding of the overall scripting and querying of big data with a basic level of completeness, readability and execution in their work.	Student fails to demonstrate an understanding of the overall scripting and querying of big data with poor level of completeness, readability and execution in their work.
Criterion 2: Student applies and integrates Hive and Pig in their case study answers.	Completes and explains clearly the code and results of the required tasks correctly with Pig Latin scripts or Hive queries where required.	Completes all the code and results of the required tasks correctly with Pig Latin scripts or Hive queries where required.	Completes more than 50% of the code and results of the required tasks correctly with Pig Latin scripts or Hive queries where required.	Completes less than 50% of the code and results of the required tasks correctly with Pig Latin scripts or Hive queries where required.	No evidence of code to support results.
Criterion 3: Student demonstrates quality analytical writing skills using appropriate conventions.	Demonstrates sophisticated analytical writing and synthesis skills in answers to the case study tasks, which are logically structured and written with clarity. No grammatical errors, good use of paragraphs to structure answers and document.	Demonstrates competent analytical writing in answers to the case study tasks, which are logically structured and written with clarity. Few grammatical errors, good use of paragraphs to structure answers and document.	Evidence of developing analytical writing in answers to the case study tasks with the beginnings of a logically structured flow and written with clarity. Some grammatical errors, poor sentence structure and poor use of paragraphs to structure document.	Very limited evidence of analytical writing in answers to the case study tasks with poor structure and flow. Some grammatical errors, poor sentence structure and poor use of paragraphs to structure document.	No evidence of analytical writing in answers to the case study tasks which include grammatical errors, poor sentence structure and poor use of paragraphs to structure document. Answers are poorly structured, do not flow easily and do not address the case study tasks. Does not exhibit good writing skills.