ISOM 671: Managing Big Data (Individual Assignment 3)

Name
Email

There are 3 numbered questions.* Please submit your assignment as a single PDF or Word file by uploading it to course canvas page. *You should provide: all commands, results of any commands, and answers to questions, if any.

1. (10 points) Discuss the differences between Hive, Impala, and Spark.

1.1. Load the NY Taxi tripdata.csv from canvas to your S3 bucket and from there into Hive and Spark.

1.2. Write a code in Hive and Spark to find total records in the data and find number of records with rate codes: 1, 3, and 5.

1.3. Look at the log file in YARN timeline server and Spark history server, and take a snapshot of the entry that is returning the number of records requested by your code.

2. (10 points) Read the article Doing Data Science (https://hdsr.mitpress.mit.edu/pub/hnptx6lq/release/8).

2.1. Assuming you are the CIO at Coke, what database systems (Hive, Impala, and Spark) would you implement for different functional groups (e.g. marketing, operations, and finance) within the organization. Please try to align your answer to the data science framework (fig 1) as presented in the article.

Note: the points will be allocated not on the correctness of the answer but the reasoning and analysis behind the answer.

3. (10 points) Load multiple Shakespeare plays in your S3 bucket and load all of those in Spark. (data source: https://github.com/Pseudomanifold/Shakespeare/tree/master/Plays/comedies)

3.1. Write a PySpark code that calculates word frequency across all documents

3.2. Write a PySpark code that find the frequency for word “love” in each document and results are presented in descending order of count