ISOM 671: Managing Big Data (Assignment 2)

Name
Email

There are 3 numbered questions (1-point each).* Please submit your assignment as a single PDF or Word file by uploading it to course canvas page. *You should provide: SQL statements, results of any SQL statement (typically copy first 10 rows), and answers to questions, if any.

The following two questions are based on the NoSQL and Hadoop content of course.

1. (10 points) Briefly discuss the role of following Hadoop ecosystem technologies:
1.1. Yarn
1.2. Zookeeper
1.3. Oozie
1.4. Sqoop
1.5. Hue

2. (10 points) Submit the list of pig script commands in text file (e.g. nyse.pig) file
2.1. Upload NYSE stock data (daily and dividends) from canvas to your S3 bucket
2.2. Using Pig shell (grunt):
2.2.1. Load files into Pig
2.2.2. Join daily and dividends on stock and date
2.2.3. Calculate dividend/close_price
2.2.4. Find stock ticker and date for minimum and maximum value of dividend/close_price

3. (10 points) Submit the above hive queries in text file (e.g. nytaxi.hql) file:
3.1. Upload tripdata.csv from canvas to a folder (hive) in your S3 bucket
3.2. Create external table nyTaxi with:
3.2.1. Column definitions (refer to data for column names and data types)
3.2.2. Fields terminated by “,”
3.2.3. Lines terminated by “\n”
3.2.4. Stored as textfile at location “YOUR S3 Bucket Folder”
3.3. Get distinct rate_code_id from the table
3.4. Show all rows/columns where rate_code_id = 1