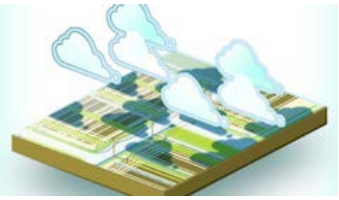

Quick Questions 15

Announcements

Syllabus

Orientation (Mandatory; Start Here)

AWS Credit Opt-in Consent Form

Responses to Commonly Asked Questions

TASKS

Project Overview ←

Task 1

Task 1 Queries

Task 1 Example Solutions

Task 2

Task 2 Queries

Task 2 Example Solutions

COURSE COMMUNITY

All Forums

Know Your Classmates

Help Center

Task 1: Data Extraction, Batch Processing with Hadoop

Help Center

Instructions

Update 1/25/16: You may use DynamoDB instead of Cassandra.

In general, the goals of this task are to perform the following:

1. Extract and clean the transportation dataset, and then store the result in HDFS.
2. Answer **two questions from Group 1**, **three questions from Group 2**, and **both questions from Group 3** using Hadoop. Store the results for questions from Group 2 and Question 3.2 in Cassandra or DynamoDB.

Part 1

Your first task is to clean and store the dataset. To accomplish this, you must first retrieve and extract the dataset from the EBS volume snapshot. Afterwards, you should explore the data set and decide on how you wish to clean it. The exact methodology used to clean the dataset is left to you. The cleaned data must be stored on HDFS.

*Note:*The dataset contains a large amount of information that will not be useful for this task. Before you start, you should explore the database directory and decide on what you want to keep and discard. Consider removing or combining redundant fields and storing the useful data in a format that makes it easier for you to answer your chosen questions. Again, it is worthwhile to consider how many AWS credits you need to perform the tasks, and optimize your usage.

Part 2

Your second task is to answer your chosen questions using Hadoop. As noted above, the results for questions from Group 2 and Question 3.2 should be stored in Cassandra or DynamoDB. The exact approach you use to answer these questions is again left to you. Whatever approaches you choose, make sure you briefly explain and justify them in your report. See Task 1 Queries for specific queries.

Submission

PDF Report

You must submit your report in PDF format. Your report should be no longer than **4 pages, 11 point font**. Your report should include the following:

1. Give a brief overview of how you extracted and cleaned the data.
2. Give a brief overview of how you integrated each system.
3. What approaches and algorithms did you use to answer each question?
4. What are the results of each question? Use only the provided subset for questions from Group 2 and Question 3.2.
5. What system- or application-level optimizations (if any) did you employ?
6. Your opinion about whether the results make sense and are useful in any way.

Video Demonstration Link

In your report, you will also need to submit a link to a video demonstration of your approach. Your video should be **no more than 5 minutes long**. Your video should include the following:

1. Ingesting and analyzing data for each question.
2. Displaying/querying the results for each question.

Following is a list of suggested websites to upload your video.

- YouTube
- Vimeo
- Youku
- Vidme
- Sendvid

Submit Task 1

Evaluation

Your peers will evaluate your submission based on the **Task 1 Rubric**. This assignment is worth 50 points. The evaluation period will begin immediately after the submission deadline. You must evaluate **five** of your peers' submissions or your own submission score will be penalized by 20%.

Evaluate Task 1

Deadlines

See the Syllabus for detailed information about deadlines for this task.

Getting Help

You can discuss Task 1 with your peers in the Task 1 Discussion forum.

Created Tue 25 Aug 2015 4:48 PM MSK

Last Modified Sat 30 Jan 2016 1:19 AM MSK