UDACITY

‹ Return to Classroom

DISCUSS ON STUDENT HUB

Data Engineering Capstone

REVIEW

CODE REVIEW

HISTORY

Requires Changes

1 specification requires changes

Hello Udacity Learner

This was a brilliant submission. You did a great job and should be proud of yourself.
After reviewing this submission, I am impressed with the effort and understanding put into this project.
Please, you still have very few modifications to do, follow the instructions and I bet that you will pass in your next submission.
I encourage you to exploit your ability to keep learning new things, you will be amazed at how great a problem solver you'll become.
This is a huge step towards a bright career and the Udacity team wishes you success all the way.
All efforts are appreciated, please keep the learning flame burning. Have a nice wonderful day!

Further Reading

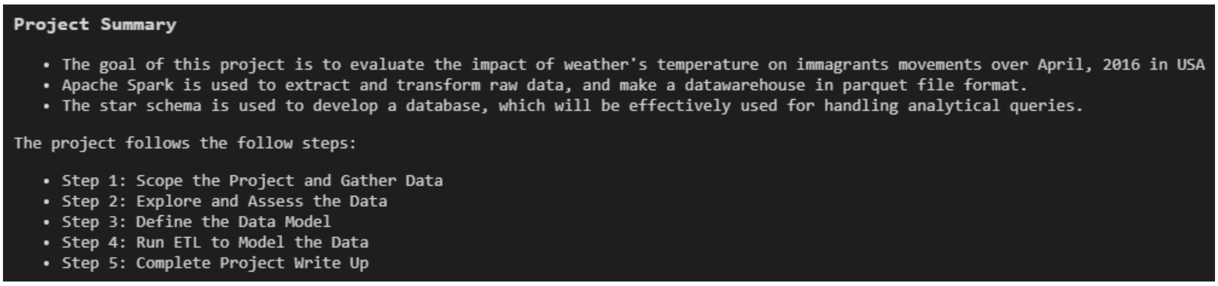
- The Rise of Data Engineering: Common Skills and Tools: You know that in recent years data engineering has greatly evolved and this goes with the tools too. Here in this article, you will explore the most commonly used and important data engineering tools in use today, how best you can integrate them into your projects.
- Data Engineer vs Data Scientist: There are a lot of people that mix out these two disciplines, here you will have an in-depth analysis of the difference between them and also have an explanation of how similar

they are.

STAY HEALTHY 🙏

Write Up

The write up includes an outline of the steps taken in the project.
The purpose of the final data model is made explicit.



Project Summary

- The goal of this project is to evaluate the impact of weather's temperature on immagrants movements over April, 2016 in USA
- Apache Spark is used to extract and transform raw data, and make a datawarehouse in parquet file format.
- The star schema is used to develop a database, which will be effectively used for handling analytical queries.

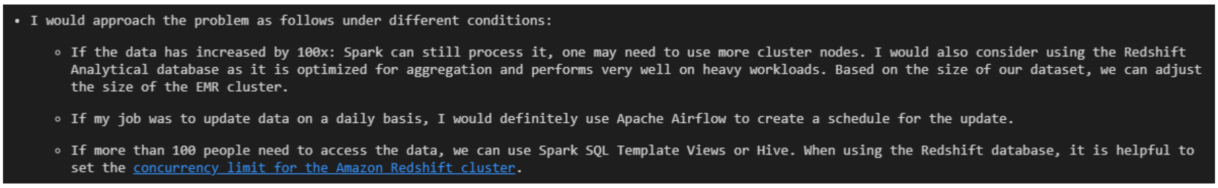
The project follows the follow steps:

- Step 1: Scope the Project and Gather Data
- Step 2: Explore and Assess the Data
- Step 3: Define the Data Model
- Step 4: Run ETL to Model the Data
- Step 5: Complete Project Write Up

Very good work at the beginning of this project, your purpose is clear and the steps taken in this project are clearly documented. What a nice start!

The write up describes a logical approach to this project under the following scenarios:

- The data was increased by 100x.
- The pipelines would be run on a daily basis by 7 am every day.
- The database needed to be accessed by 100+ people.



- I would approach the problem as follows under different conditions:
    - If the data has increased by 100x: Spark can still process it, one may need to use more cluster nodes. I would also consider using the Redshift Analytical database as it is optimized for aggregation and performs very well on heavy workloads. Based on the size of our dataset, we can adjust the size of the EMR cluster.
    - If my job was to update data on a daily basis, I would definitely use Apache Airflow to create a schedule for the update.
    - If more than 100 people need to access the data, we can use Spark SQL Template Views or Hive. When using the Redshift database, it is helpful to set the concurrency limit for the Amazon Redshift cluster.

Your approaches are all logical and good but I will propose this for the last scenario because it is more focused on the great number of users (The database needed to be accessed by 100+ people.) I would have rather said that
The more people accessing the database the more CPU resources you need to get a fast experience. By using a distributed database you can improve your replications and partitioning to get faster query results for each user.

It's important not to rely much on AWS.

Extra Materials

- Optimize Amazon S3 for High Concurrency in Distributed Workloads

The choice of tools, technologies, and data model are justified well.

The data model is well described/justified also you gave justifications on the tools and technologies that were used in this project. In this project, your data model description was not the best so I have included some resources, the link will give you insights on how to choose a data model for a particular purpose.

Data Models

- Picking the right data model: You will see how you can analyze different parameters in order to choose a data model to achieve a particular goal.

Tools and Technologies

- Who uses Spark and why?: Spark helps data scientists by supporting the entire data science workflow, from data access and integration to machine learning and visualization using the language of choice—which is typically Python. It also provides a growing library of machine-learning algorithms through its machine-learning library (MLlib).

Execution

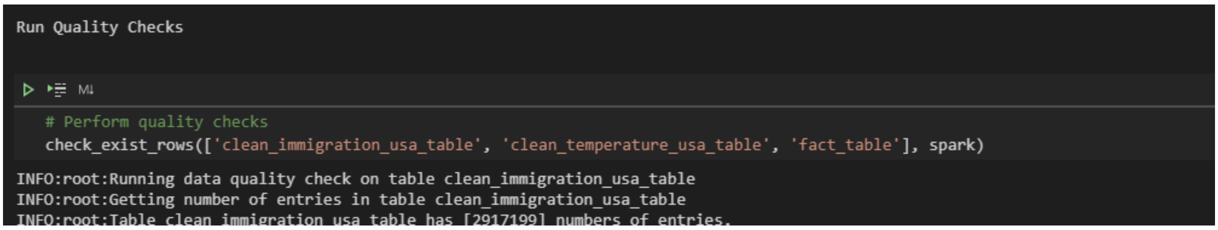
All coding scripts have an intuitive, easy-to-follow structure with code separated into logical functions. Naming for variables and functions follows the PEP8 style guidelines. The code should run without errors.

Your code is neat and you respected the PEP8 guidelines. However I think there is still room for improvement, I noticed that you don't comment much on your code, please consider going through this document as it helps a lot with basic good programming practices, it would still be enriching and sharpen your skills.

Extra Material

- Toward Developing Good Programming Style: There are good programming styles that can suit you based on the context in which you are found. This resource will provide you with instructions on working in a notebook and can help you increase your skills notably in best programming practices.

The project includes at least two data quality checks.



Run Quality Checks

```
# Perform quality checks
check_exist_rows(['clean_immigration_usa_table', 'clean_temperature_usa_table', 'fact_table'], spark)
```

```
INFO:root:Running data quality check on table clean_immigration_usa_table
INFO:root:Getting number of entries in table clean_immigration_usa_table
INFO:root:Table clean_immigration_usa_table has [2917199] numbers of entries.
```

You have a very good data quality check, but this rubric requires that you have a least **two** data quality checks. Please write another test for this project. Please note that you are to write a test different from the `checking_rows()` you have a different test from the list below:

`checking rows()`, you have a different test from the list below,

- Integrity constraints on the relational database (e.g., unique key, data type, etc.)
- Unit tests for the scripts to ensure they are doing the right thing
- Source/Count checks to ensure completeness

- **The ETL processes result in the data model outlined in the write-up.**
- **A data dictionary for the final data model is included.**
- **The data model is appropriate for the identified purpose.**

You gave a detailed outline of the ETL processes in this write up and you gave valid reasons for using the star schema. A data dictionary for the final data model was included. 👍

ETL Processes

- 3 Ways to Build An ETL Process with Examples: This will give you very important tips to build an efficient ETL pipeline for a specific purpose.

Data Dictionary

- Create a data dictionary: Explains how to create different sorts of data dictionaries.

The project includes:

- **At least 2 data sources**
- **More than 1 million lines of data.**
- **At least two data sources/formats (csv, api, json)**

You used data from different sources. In this rubric, we want to see your capability to handle data of a certain volume and you nailed it. 😄

☑ RESUBMIT

DOWNLOAD PROJECT

Learn the best practices for revising and resubmitting your project.

RETURN TO PATH

Rate this review

START