Hey,

Congratulation, the project is very detailed and nicely presented. The project is meeting all expectation and passes all the rubrics.

The detailed write up is an absolute delight to read, the presentation is very clear. Hallmark of any great project is the ability of the project to be able to reproduce by anyone else, and the write up like yours definitely helps.

All the best

**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.**

Great start.

* The project notebook is very professionally setup with proper steps and detailed notes.
* The write is very very clear with project motivation and steps undertaken detailed out.
* The data model present is also very clear and the way it is going to be used is also detailed.

Nice work

**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.**

The data was increased by 100x.

Great suggestion. Apache Spark is linearly scalable, which means you may simply add the number of clusters to increase the performance. With Spark distribution mechanism using clustering mulitple nodes can be invoked for processing of the files which will enhance the performance

The pipelines would be run on a daily basis by 7 am every day.

Airflow are most obvious choice. Excellent

The database needed to be accessed by 100+ people.

Redshift database system allows for greater accessibility by trading off some write flexibility.

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

Overall great work.

The use of python with S3 is a good approach, the spark is much better as it handles large data set with ease. postgres can be easily scaled with redshift on Amazon AWS.

The tables created are also justified with the final conclusion the project is deriving

**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.**

The Code is PEP8 compliant, there are optimal comments and docstring are utilized appropriately.

Nicely done  
You can read more about

* [PEP](https://www.python.org/dev/peps/pep-0008/)

**The project includes at least two data quality checks.**

Excellent.

You have now two good data quality check in the notebook. The data quality step is a very important part of any data engineering project.

* **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.**

I liked the way the write up is presented, I can see there is very detailed step by step ETL process, the data dictionary is very comprehensive and more than serves the purpose

**The project includes:**

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

No project is complete without marking the reference to data sources or any web link that was referred while doing the project. This is very important to provide to credit to original authors