Meets Specifications

I appreciate and commend the efforts and hard work put into this piece. Congratulations  for making it pass this stage of learning with us and I wish that this spirit is carried forward in subsequent projects. You should be proud of yourself because success is no accident. It is hard work, perseverance, learning, studying, sacrifice and most of all, love of what you are doing or learning to do. Please keep practicing on these projects and I wish you 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.**

**Good job!**

Good work in purpose and outline of the project 👍

You did a marvelous work with the README. We need a readme file so that we can follow up and better appreciate your work, without readme it will be a bit difficult for an individual to understand your thought process.

**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.**

**Nicely done!**

Good work with your explanation 👍

**Comment**

* One could also use operating system's task scheduler for scheduling jobs.
* One could also keep a separate user table for keeping track of users logging in and out and whenever there is a spike in user activity, additional clusters can be initiated.

**Tip:**

* If you are interested in knowing more about the heavy-read system design, I would suggest that you read this [tutorial](https://medium.com/@narengowda/netflix-system-design-dbec30fede8d) about Netflix system.
* I would suggest that you read this [experience](https://medium.com/velotio-perspectives/lessons-learnt-while-building-an-etl-pipeline-for-mongodb-amazon-redshift-using-apache-airflow-543bb0b75017) sharing about the lesson learned when implementing data pipeline.

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

Good job on providing your reasonings about each tool you use and justifying the data model well👍

Python: flexible and easy to use  
Pandas: i am comfortable with python and pandas is a python library optimized for data analysis and manipulation  
Spark: Has a python api called pyspark. Also has a dataframe like pandas and allows sql queries. Suitable for data with over 1 million rows  
Local Storage: Ideally AWS S3 and Redshift would be prefered choices for files and data storage. But i used local storage because of the limited amount of data.

**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.**

Nice work with this one!  
Good usage of mark down cells and comments to document the work. These made the code to have a logical structure and easy-to-follow. The data wrangling steps were all applied.

**Learning Notes**

* [Why use functions in programming?](http://www.cs.utah.edu/~germain/PPS/Topics/functions.html)
* [Why should we comment code?](https://www.cs.utah.edu/~germain/PPS/Topics/commenting.html)
* [Top 10 tips of writing cleaner code](https://www.makeuseof.com/tag/10-tips-writing-cleaner-better-code/)

**Suggestions**

You may find this link helpful to Check your code for PEP8 requirements.

* [PEP 8: Style Guide for Python Code](https://www.python.org/dev/peps/pep-0008/)
* [How to Write Beautiful Python Code With PEP 8](https://realpython.com/python-pep8/)
* [PEP-8 Tutorial: Code Standards in Python](https://www.datacamp.com/community/tutorials/pep8-tutorial-python-code)

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

Nice work with all the data quality checks in project. 👍

**Learning Notes**

A useful resource to know how to classify correctly each type of issue:  
[Standard data quality and tidiness](https://ryanwingate.com/purpose/tidy-data/)

* **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.**

Good work in this section 👍

**The project includes:**

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

Good work here 👍 Since the data quality is good, this section is marked as meets specifications.