Data Engineer

This is an interview project created by Kranio, the project should be uploaded at a public repository, at any git provider of your choice.

Problem description

Imagine we work for a movie studio companny, and we're trying to decide what will be our next great production. For that, you should do some analytics based on past releases database found on kaggle movies dataset and present them for executives of the companny.

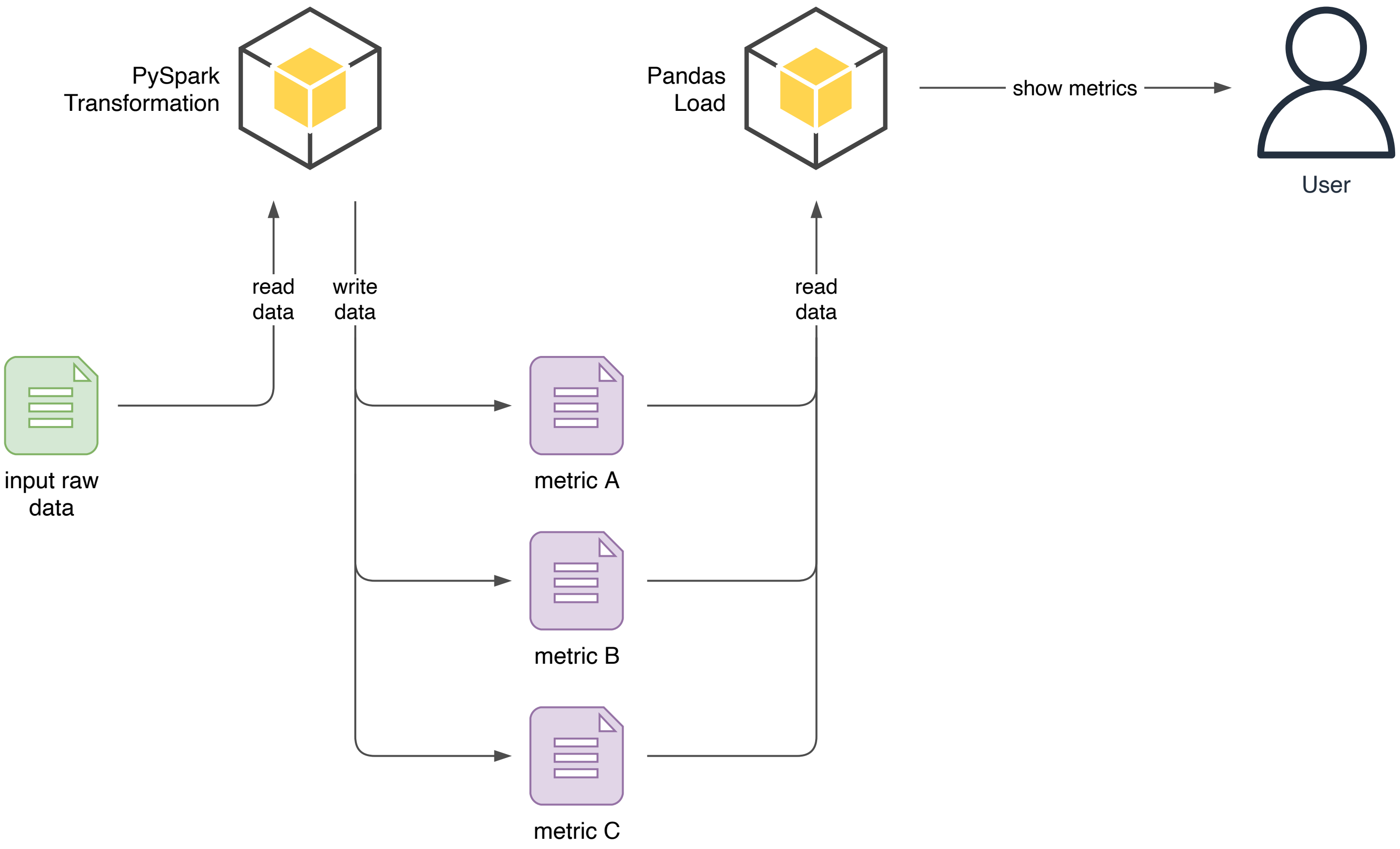
The analysis should be achieved through a small pipeline with two phases, transformation and load, transformation should be done with PySpark (metrics will be described bellow), and for the second phase, you should load the transformed data with Pandas and show their results with any charts library of your choice, some examples are:

- plotly
- matplotlib
- seaborn

The final report should present the following metrics:

- Most profitable movies by semester
- Most profitable genres by year
- Top 10 profitable movies by genres (maximum of 5 genres)
- Movies popularity by month (to know the best release date)
- Total of releases for each genre for the last 5 years

Architecture



Requirements & Observations

- You can create the solution with Jupyter Notebooks if you preffer.
- Add a a description how to test your solution.
- Always try to follow the most the best practices and clean code, you can use PyLint to check this.
- The metrics should be stored as Parquet format for better read speed.
- Solution running on Docker containers would be desirable
- Orchestrate the pipeline with Airflow would be desirable

Achieve what you can, don't be afraid of change the requirements if there is some tecnology you don't manage so well.

Good lucky and Have Fun!