**ETL Project Proposal**

**Team Members:**

1. Raphael Serrano
2. Swobabika Jena
3. Thomas Maina
4. Diana Madonko

**Project overview**

Our question of interest is to explore published books around the world and related information. This data will be able to assist people with book choices as it would provide more information on ratings, affordability and other areas that influence their choices.

**EXTRACT - Proposed data sources (minimum two data sources)**

* *Where is the data located?*

*Goodreads database*

* *What are the data set formats?*

*JSON*

* *How will you get this data? (e.g. API, scraped data, download data)*

*API and Scraped data*

* *What are your assumptions about the data? What does it cover? What is it missing?*

*We assume our data will cover the following:*

*~ Book Name*

*~ Author*

*~ ISBN*

*~ Category*

*~ Price*

*~ Format*

**TRANSFORM - Proposed clean-up and analysis**

* *What are the transformations you will apply to the data? (e.g. filtering, aggregation, derived columns)*

*Filtering and derived columns*

* *What steps will you take to clean the data and ensure its validity (e.g. messy data, duplicated data, incorrectly formatted data)*

*Deduplication, format revision and cleaning values.*

* *How will you identify potential issues with your data sources? (e.g. exploratory data analysis, simple statistics etc)*

*Exploratory data analysis*

* *How will the data be integrated? (e.g. joins, merges)*

*Merges*

* *How will you apply these transformations (e.g. jupyter notebook, python script)*

*Jupyter Notebook*

* *IMPORTANT → Why did you apply these transformations? How did this enrich your data?*

*We have chosen these transformations as they best suit the data we have selected.*

**LOAD - Data storage**

* *What type of database (relational, document) will you store the data?*

*Relational database*

* *Why did you choose this database over another database?*

*We chose this database as it will allow us to easily apply analytical functions and derived data*

* *What are your expected tables / documents and relationships between tables / documents in your database?*

*Our tables will be Book data, Review Ratings, Authors*

**Potential limitations**

* *What are the potential limitations of your above proposed steps?*

*~ The potential limitations we can run into is the dataset having limited data than expected*

*~ There could be missing fields that we need*

* *How can you control these potential issues?*

*~ We will explore different sources of data.*

#########################################################################

**ETL DIAGRAM**

Diagram

Description automatically generated