The product to be developed, which is a data cleaning tool, is to be developed with the goal of creating a robust and user-friendly tool that streamlines the process of cleaning, transforming, and analysing datasets. The system will aim to address a host of data cleaning and data quality challenges.

Reasons why this system exists?

Data sets are often marred with diverse quality problems like duplicates, missing values, violations of integrity constraints, and outliers. Identifying and rectifying these issues requires a systematic approach that a data cleaning tool like this provides (Chu, Ilyas, Krishnan & Wang, 2016). The initial phase of error detection involves spotting anomalies or mistakes in a database. Given the intricacies of qualitative error detection, the tool needs to address what types of errors to look for, how to detect them, and where in the system these errors are most likely to occur (Chu et al., 2016). As errors are identified, the next step is to repair them. The tool must decide what to repair (data or constraints), how to execute the repair, and where the repair needs to happen (Chu et al., 2016).

Data often undergoes numerous transformations due to structural changes, schema evolution, system migration, or integration of multiple data sources. Each scenario presents unique challenges, such as naming conflicts or differing data representations, requiring systematic cleaning processes (Rahm & Do, 2000). Data originating from a single source can have quality issues, especially if there are limited constraints on what can be entered and stored. These issues can be due to the lack of schema, inadequate data model restrictions, or poor schema design, leading to errors and inconsistencies at both the schema and instance levels (Rahm & Do, 2000). These are more pronounced when integrating multiple data sources. Independent development and maintenance of these sources lead to heterogeneity, resulting in conflicts and inconsistencies at both schema and data levels. Problems include naming conflicts, structural conflicts, differing value representations, and various data conflict types at the instance level (Rahm & Do, 2000). A significant challenge in cleaning data imported from multiple sources is identifying overlapping data, specifically matching records that refer to the same real-world entity (Rahm & Do, 2000. A data cleaning tool is needed to eliminate duplicate information and consolidate complementing information for a more consistent view on instances referring to a similar entity (Rahm & Do, 2000). Data cleaning need not only be reactive but could also be proactive, requiring tools to help in designing appropriate database schemas and integrity constraints. This measure usually helps to prevent the entry of ‘dirty data’ and hence reduces future cleaning requirements on the dataset (s). (Rahm & Do, 2000).

Can you review other existing systems or products that address this problem? (how do they meet or fail to meet the needs of your target users)

There are a plethora of data cleaning tools that could be deemed as our competitors. However, for the sake of brevity, two of the more well-known data cleaning application will be discussed with their merits (how they meet user needs) and demerits (how they fall short of user needs highlighted)

Oracle Analytics, in reference to its desktop application, is a tool that allows the processing of datasets from various sources to allow its users make decisions and evaluate predictions based on data. It can be found here <https://www.oracle.com/ie/business-analytics/#:~:text=Offering%20valuable%2C%20actionable%20insights%20from,Explore%20Analytics%20products>

Pros of the Oracle Analytics desktop app:

* Its Data Visualization features allows it users to create rich visual representations of data from various sources. It offers a wide range of charts, graphs, and tables to help users understand their data.
* It considers business users and provides them with the ability to analyse data without relying heavily on IT. Users can drag-and-drop data, create calculations, and produce insights.
* It’s data Integration allows the app to connect to a range of data sources, from local Excel spreadsheets to databases and cloud applications. It can blend and transform data from various sources to provide a unified view.
* Its advanced analytics allows the app go basic visualization and offer machine learning tools to enable features such clustering, forecasting and enabling deeper insights.
* Its users can create their analyses or reports which they can share s with others aiding decision making.

Cons of the Oracle Analytics desktop app:

* Its comprehensive nature makes for a steeper learning curve for features, especially when compared to more intuitive, plug-and-play analytics tools. Particularly since it uses a lot of SQL syntax to make transformations to columns it is not very beginner friendly.
* The mac desktop version is loaded with a myriad of technical deficiencies and couldn’t import CSV files, which serves as a significant concern since many datasets are in CSV’s.
* Oracle solutions are known to be on the pricier side, especially for small to medium-sized businesses. The licensing and maintenance costs can be a barrier for some organizations particularly the cloud suite of the analytics tool.
* Some users have reported challenges when connecting to non-Oracle sources or third-party tools.

Open refine (formerly google refine) , in reference to its desktop application, is a tool that can help transform datasets into other formats or carry out transformations on the dataset itself depending on the user’s choice. It can be found here <https://openrefine.org/>

Pros of the Open refine desktop app:

* Open Refine records all changes applied to the data, allowing users track and copy changes to another dataset if needed.
* It can handle large amounts of data instances efficiently, allowing users to process and clean data that would be cumbersome in traditional spreadsheet software.
* It can allow a variety of transformations and conversions from one data format to another.

Cons of the Open refine desktop app:

* Legacy UI feel as its kept much of its original UI from the original google refine application.
* It does not support advanced data visualizations.
* It is not built to perform complex statistical analysis.
* It does not integrate directly with data storage systems (databases)

**References**

Chu, X., Ilyas, I. F., Krishnan, S., & Wang, J. (2016, June). Data cleaning: Overview and emerging challenges. In *Proceedings of the 2016 international conference on management of data* (pp. 2201-2206).

Rahm, E., & Do, H. H. (2000). Data cleaning: Problems and current approaches. *IEEE Data Eng. Bull.*, *23*(4), 3-13.

**Appendix**

<https://www.oracle.com/ie/business-analytics/#:~:text=Offering%20valuable%2C%20actionable%20insights%20from,Explore%20Analytics%20products>

<https://openrefine.org/>