**Martin Hmira‘s Data Engineering project**

**Business Concept**

Our business goal is to get a better understanding of open-source technology trends. We want to understand which are the "hot" projects and see how the monitored projects/companies are performing compared to others. Our goal is to better understand the development of the open source tech market - we want to answer these questions using several data sources.

**Datasets:**

We will use the following freely available datasets provided by BigQuery: ● Github ● Hackernews ● StackOverflow

**Configuration:**

We will get the configuration data from a Google Sheets. The target is to create a common data repository for these datasets, load them into a central database, apply the necessary transformations and create useful insights from the available data, that is presentable with Data Visualizations. Example Question: Check if there is a correlation between a trending Hackernews post about a tool and the Github stars statistic.

**Architecture**

**Database**

We will use Google BigQuery to build our Data Warehouse. More information on the free tier is here: <https://cloud.google.com/free>

**Ingestion**

To ingest the data from the source system we will use Open Source Ingestion tool called Airbyte. Airbyte is easy to learn, it has an intuitive interface, and it can be run as a docker container Airbyte documentation: <https://docs.airbyte.com/>

**Transformation**

To transform/digest the raw dataset we will use dbt which is a Python-based Open Source SQL templating framework. The advantage of this is that it is easy to develop, and software development best practices can be applied, for example: ● All code lives on Github ● You can write DRY code (Jinja-based SQL templating) ● Out-of-the-box data quality testing. ● CI pipeline Dbt is an open core project, so the fully featured CLI is available for free, but there is SaaS integrated environment/scheduler that is free for one developer seat – but not necessary for this project Docs - <https://docs.getdbt.com/> Get started - <https://courses.getdbt.com/courses/fundamentals>

**Visualization**

Our choice of tool is Power BI for the data visualization part of the project. The desktop version of Power BI is free to use. Documentation: <https://docs.microsoft.com/en-us/power-bi/>