QS – data loading advise

Niels van der Burg

# Introduction

The quantified student project focusses on quantifying student to provide students with ways to improve their performance. The project is currently building a method to reliably collect and save data. Based of talks with the product owner (Eric Slaats), the data will be combined with open-data and other sources. This research focusses on providing a method to save this data.

# Research Question

How can Quantified Student create a method to create dashboards based of our own data and external sources?

To answer this, several sub questions will be researched to provide an overview of the issue:

1. What data will be collected?
2. How is data collection from several sources done in the field?
3. What are the requirements for collecting the data?

To answer these sub questions, the dot framework is used. The table below will describe the research methods applied for this advise and analysis:

| Subquestions | Method | Actions |
| --- | --- | --- |
| What data will be collected? | Interview | Interviewing product owner and the current project lead |
| How is data collection from several sources done in the field | Expert interview | Interviewing a data engineer or architect for insights into methods of data collection |
| What are the requirements for collecting the data? | Requirements prioritization | Documenting the requirements for this system and prioritizing them |

# What data will be collected?

To establish what data the stakeholders and product owner wanted to collected. An interview was held to establish what the collected data is and what the exact goal is of the project. The people interviewed were Eric Slaats (product owner) and Jelle Maas (current project lead).

Out of the interview with Eric Slaats I was able to establish the following:

* The data will be collected from a set of static sources, these include:
  + Weather (most likely KNMI API).
  + Watches.
  + Canvas.
  + Possible future canvas based plugins.
  + CBS or other data sources.
  + User-based data such as spotify.
* There needs to be an analysis of which data can be collected.
* There needs to be an analysis of the GDPR to prevent legal issues.

Out of the interview with Jelle Maas came the following conclusions:

* It is unclear which data is needed to collect from the smartwatches.
* There is, as of writing this, no method for saving the data.
* The data will be collected from several watches, the current team is looking into implementing open-source software to extend the list of supported watches.
* What data will be collected?

**Conclusion**

The data sources will be collected from publicly available datasets. Thus warranting the need for collection via the usage of HTTP or via an API. There is, as of writing, no database for saving the collected watch data.

The data will be collected from a combination of API (HTTP and HTTPS) and databases within FHICT (smart watch).

# How is data collection from several sources done in the field

To ensure that this research is focussing on a proven technique. An interview will be held with experts to get insights into possible solutions based of their knowledge and experience. For this an interview was held with (XXXX) and (XXX).

Out of the interview with (XXXX):

# What are the requirements for collecting the data?

To get provide an advise that best fits this isseu. A requirements prioritization will be made. This will allow for a concrete selection of solution and provides a best fit based of the prioritizations.

|  | **LakeLMS** | **Apache Druid** | **Apache Doris** | **Graphana** | **Apache airflow** | **In-house** |
| --- | --- | --- | --- | --- | --- | --- |
| **Load several sources** | Y | Y | Y | N |  | Y |
| **Load from HTTP** | Y | Y | Y | N |  | Y |
| **Load from HTTP(S)** | Y | Y | Y | N |  | Y |
| **Load from SQL** | Y | Y | Y | Y |  | Y |
| **Allow for custom queries** | Y | Y | Y | Y |  | Y |
| **Has an API** | Y | Y | Y | Y |  | Y |
| **Can be produced within 2 semesters** | Y | Y | Y | Y |  | N |
| **New data sources without coding\*** | Y | Y | Y | Y |  | ? |
| **Scheduled loading of data** | N | Y | Y | Y |  | Y |
| **Join data automatically** | N | Y | Y | N |  | Y |
| **Enforces a structured format** | N | Y | Y | Y |  | Y |
| **Maintenance not completely on delta\*** | Y | Y | Y | Y |  | N |
| **Well documented** | Y | Y | Y/N | Y |  | Y/N |

*New data sources without coding\*  
To prevent that the application relies on students maintaining it.*

*Maintenance not completely on delta\*  
To prevent that the application is difficult to maintain or just not maintained. It is preferred that development is not reliant on Delta. Since we cannot expect students to maintain a project when they are done studying.*