A

dHARTI PATEL

8807575

ZARNA GOHIL

8800060

BASVaRAJ JALIMINCHE

8800149

April 23, 2022

GROUP PROJECT

**TABLE OF CONTENT**

1. **Introduction**
2. **Snapshots**
3. **Difficulties**
4. **Curate**
5. **Outcome**
6. **Reference**

**INTRODUCTION**

**What is MarkLogic?**

MarkLogic is a database designed from the ground up to make massive quantities of heterogenous data easily accessible through search. The design philosophy behind the evolution of MarkLogic is that storing data is only part of the solution. The data must also be quickly and easily retrieved and presented in a way that makes sense to different types of users. Additionally, the data must be reliably maintained by an enterprise grade, scalable software solution that runs on commodity hardware. The purpose of this guide is to describe the mechanisms in MarkLogic that are used to achieve these objectives.

**Why and where is MarkLogic used?**

MarkLogic fuses together database internals, search-style indexing, and application server behaviors into a unified system. It uses XML and JSON documents as its data model, and stores the documents within a transactional repository. It indexes the words and values from each of the loaded documents, as well as the document structure. And, because of its unique Universal Index, MarkLogic does not require advance knowledge of the document structure and adherence to a particular schema. Through its application server capabilities, it is programmable and extensible.

MarkLogic clusters on commodity hardware using a shared-nothing architecture and supports massive scale, high-availability, and very high performance. Customer deployments have scaled to hundreds of terabytes of source data while maintaining sub-second query response time.

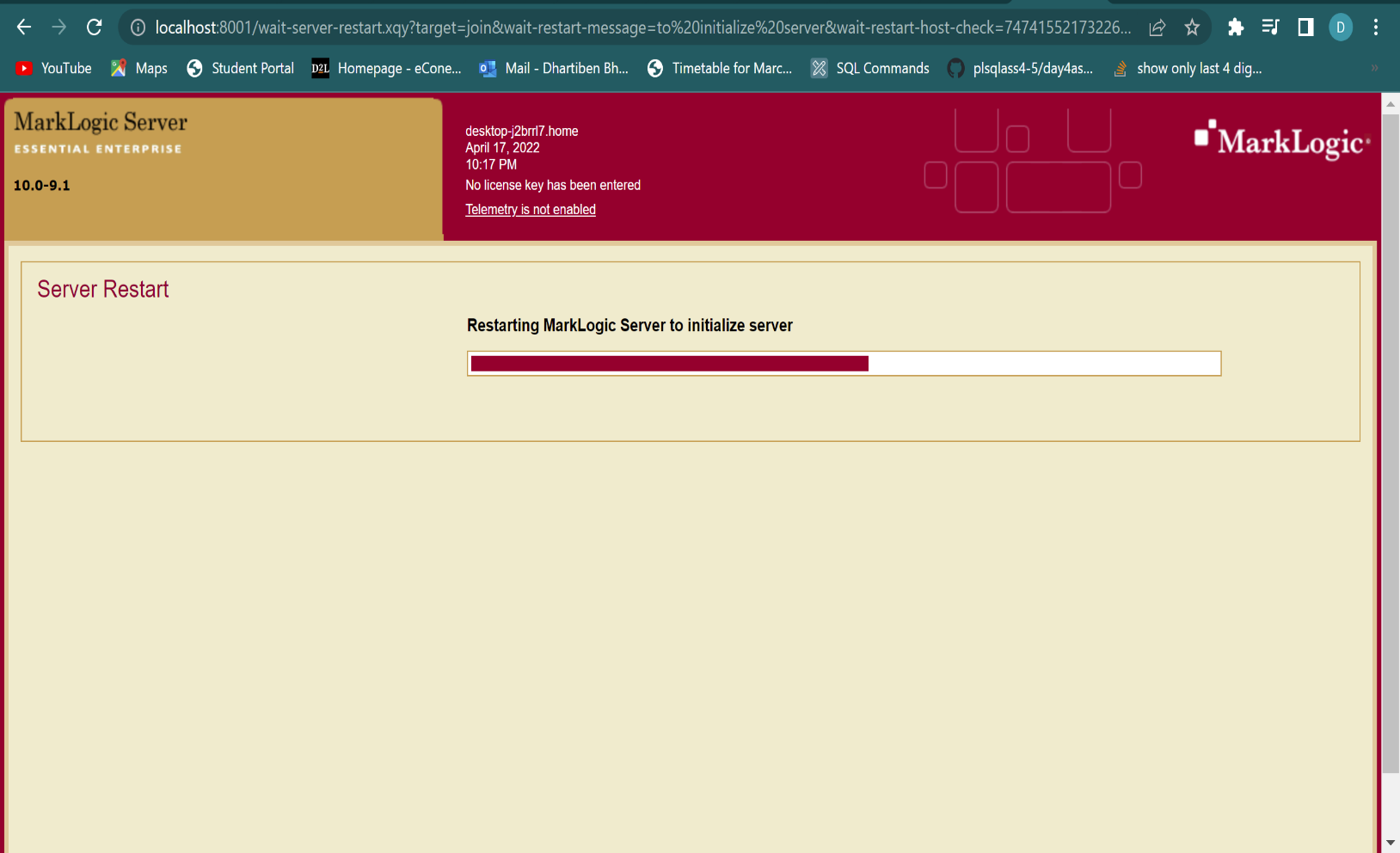
MarkLogic Server currently operates in a variety of industries. Though the data stored in and extracted from MarkLogic is different in each type of industry, many customers have similar data-management challenges.

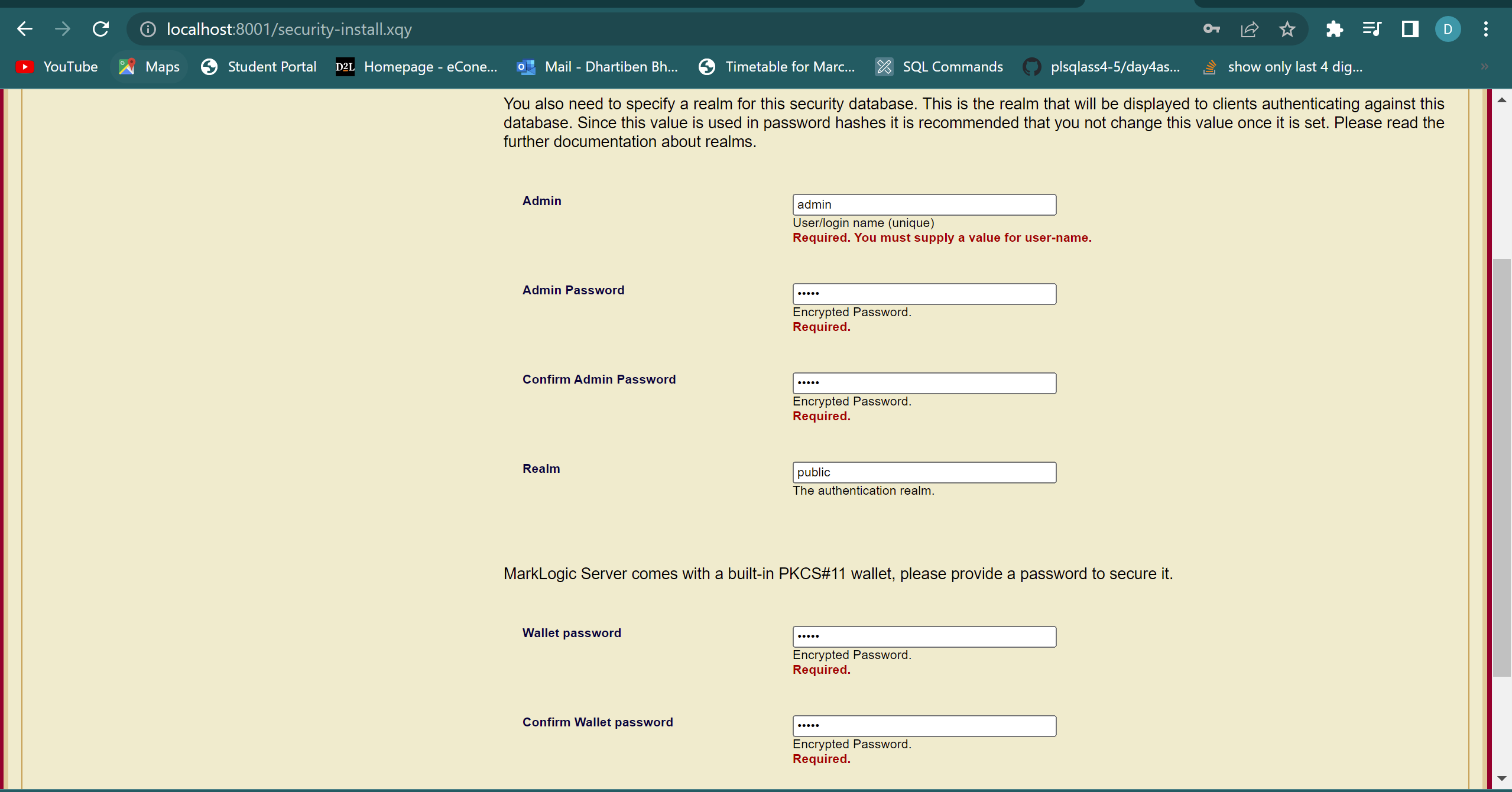
Common themes include:

* Rapid application development and deployment
* Ability to store heterogenous data from multiple sources in a single repository and make it immediately available for search
* Accurate and efficient search
* Enterprise-grade features
* Low cost

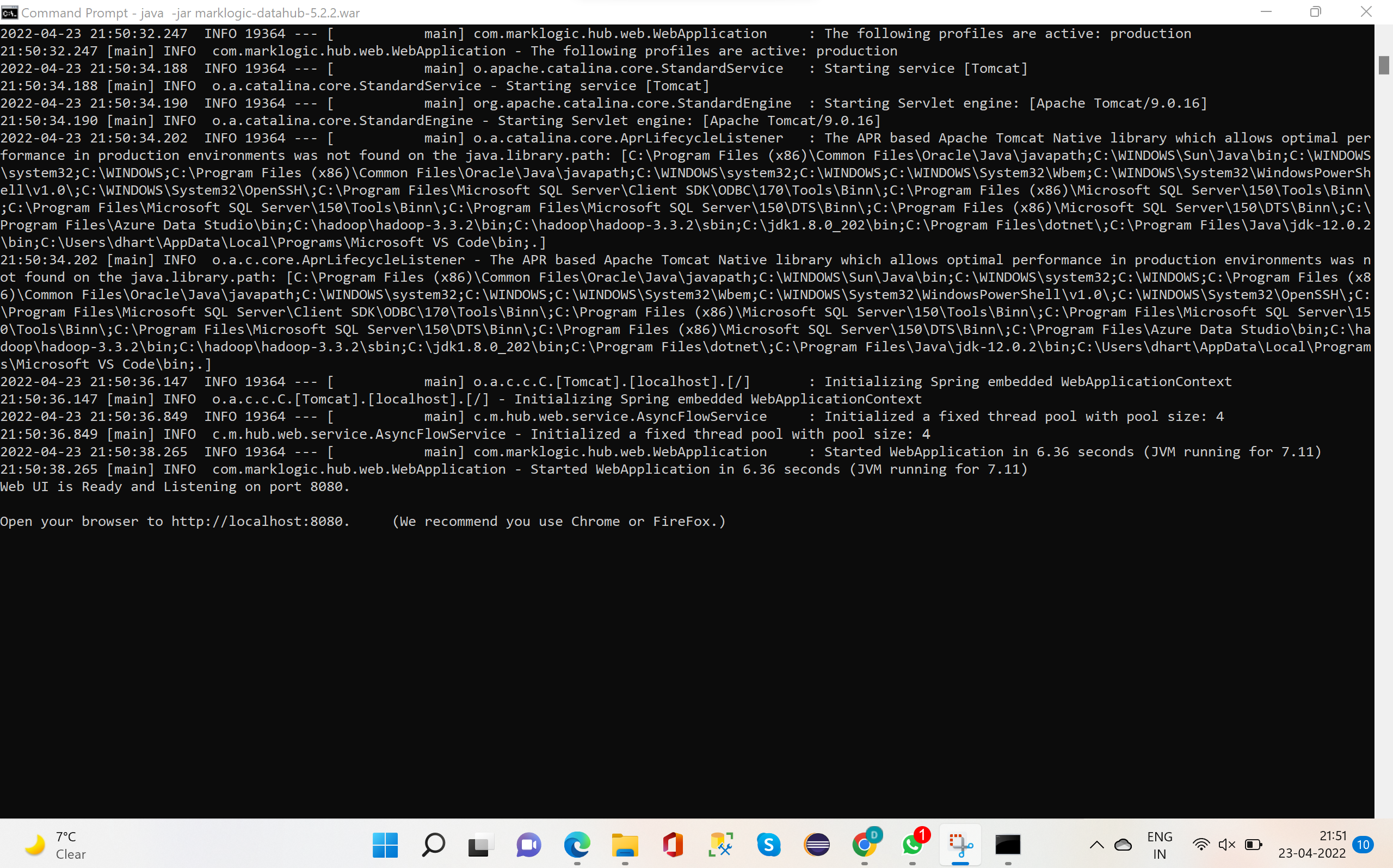
**SNAPSHOTS**

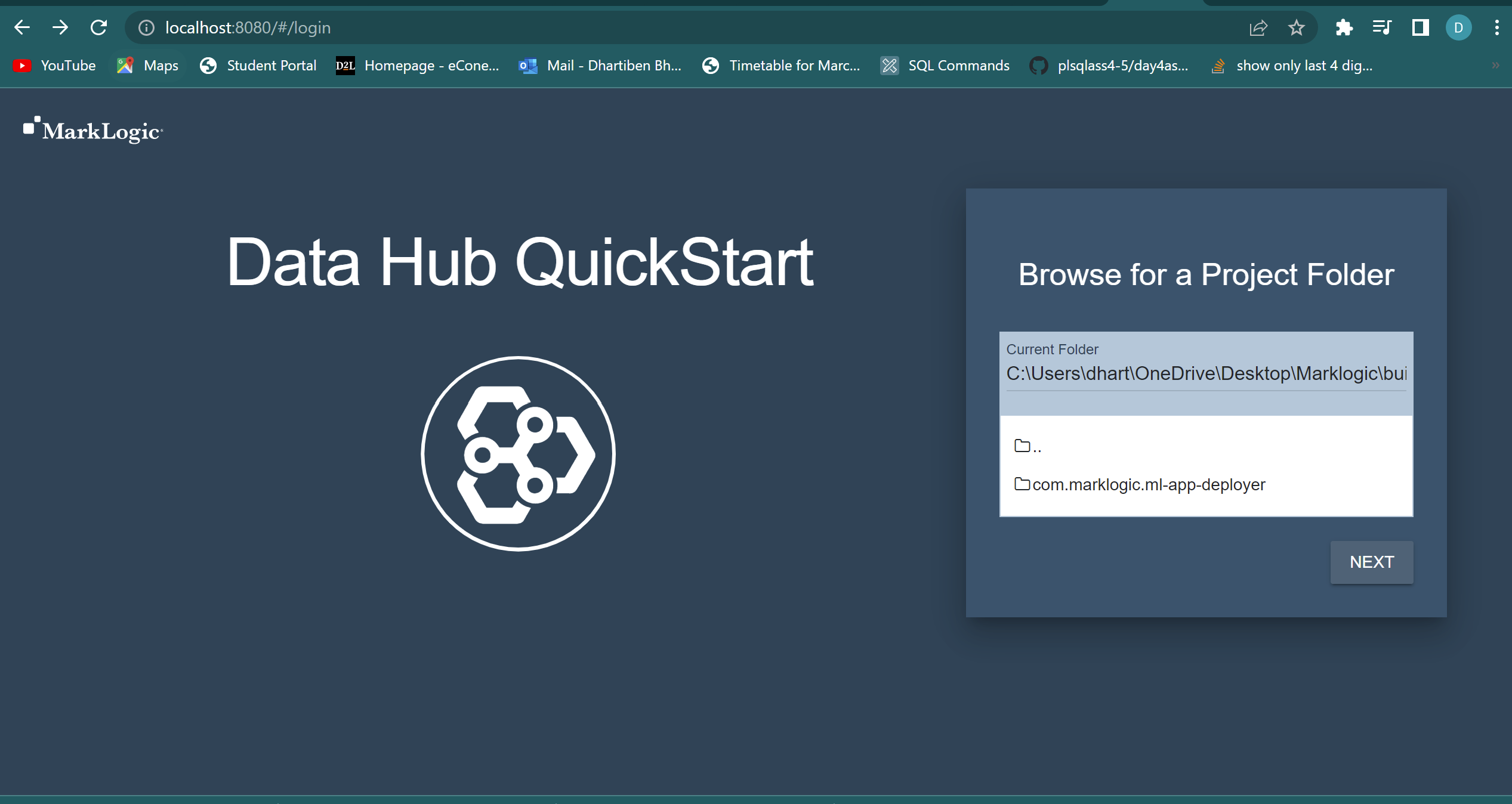
STEP 1: Downloaded Marklogic latest version

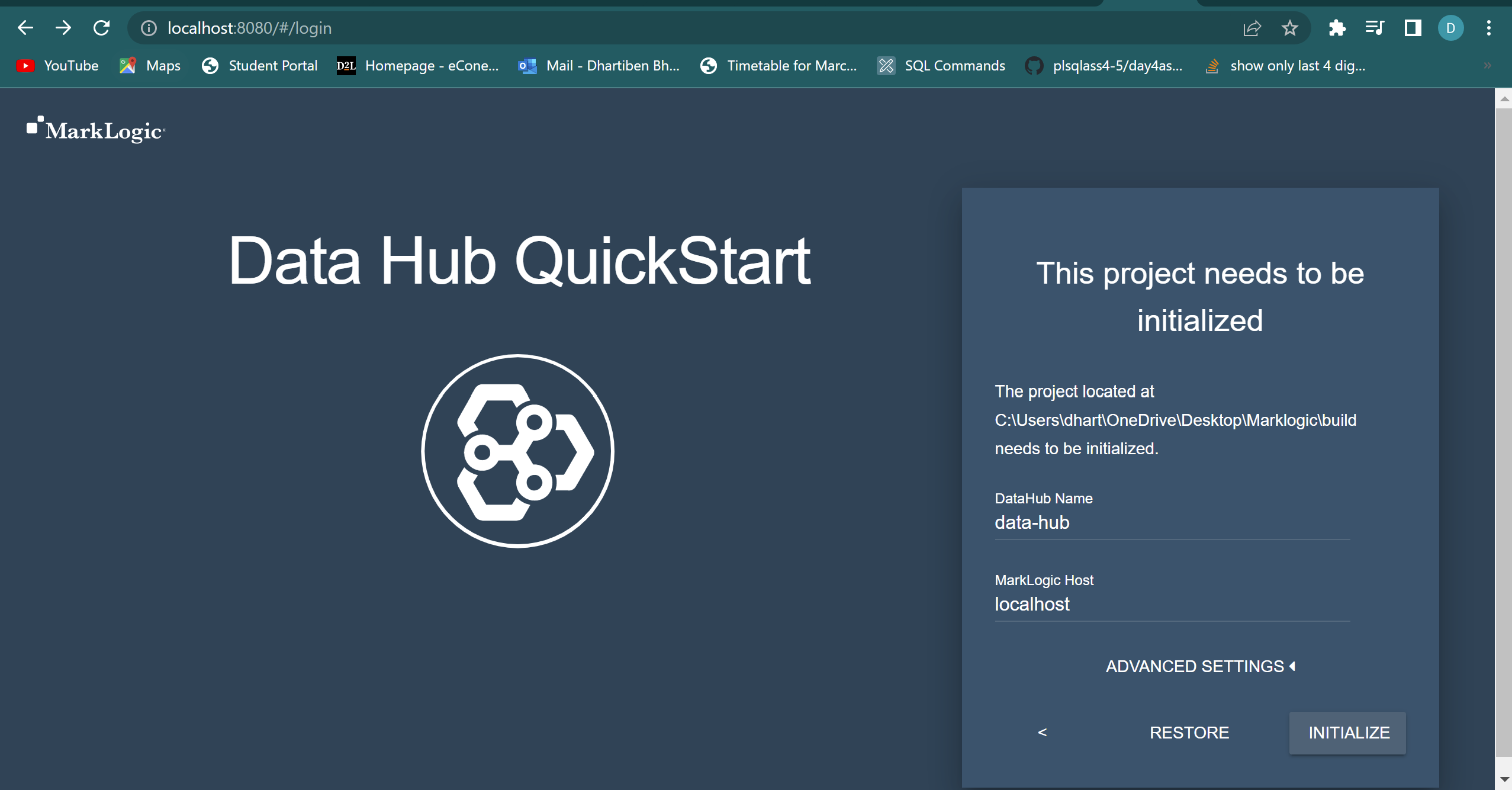


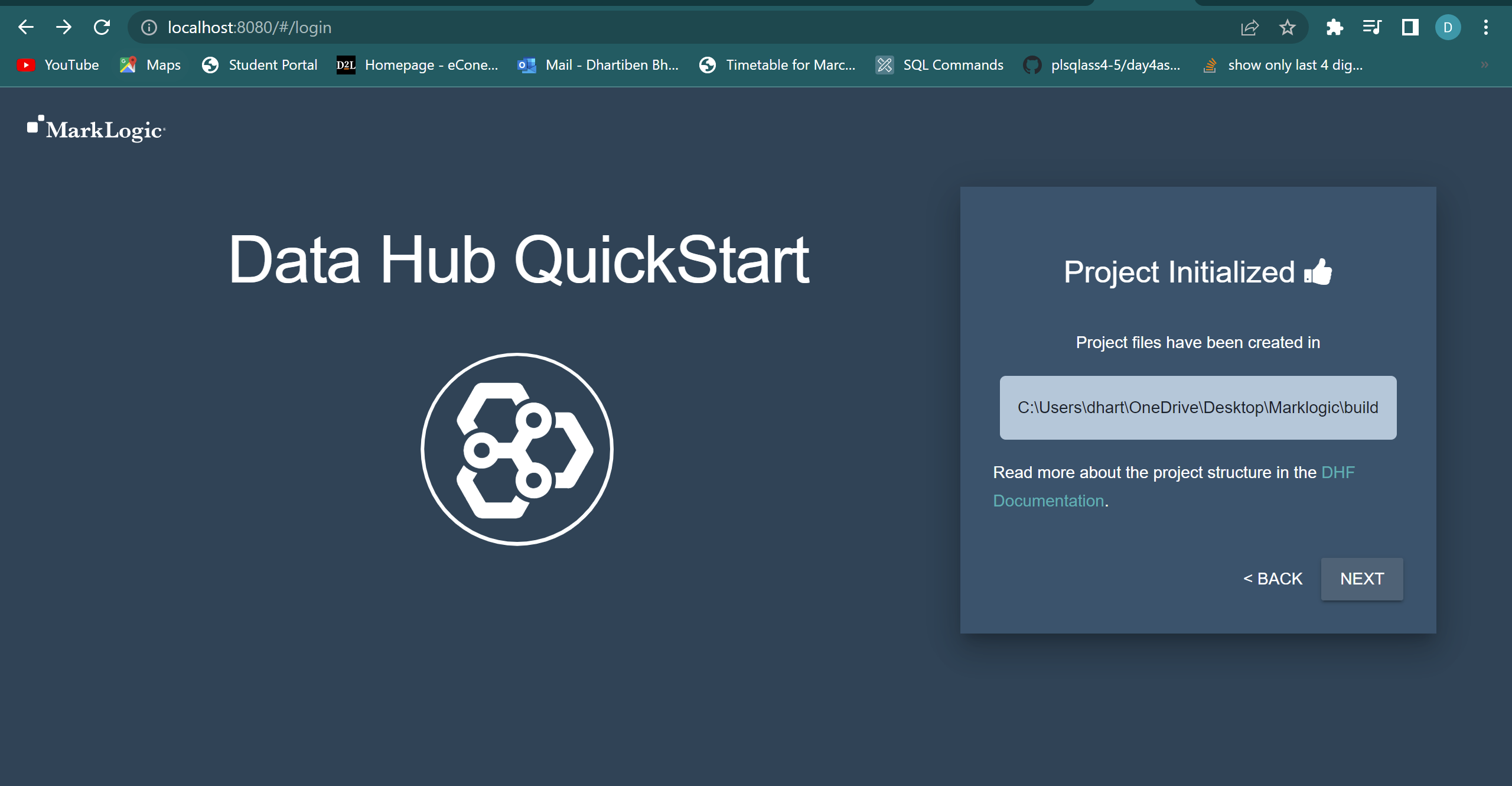
STEP 2: Setting up password and admin

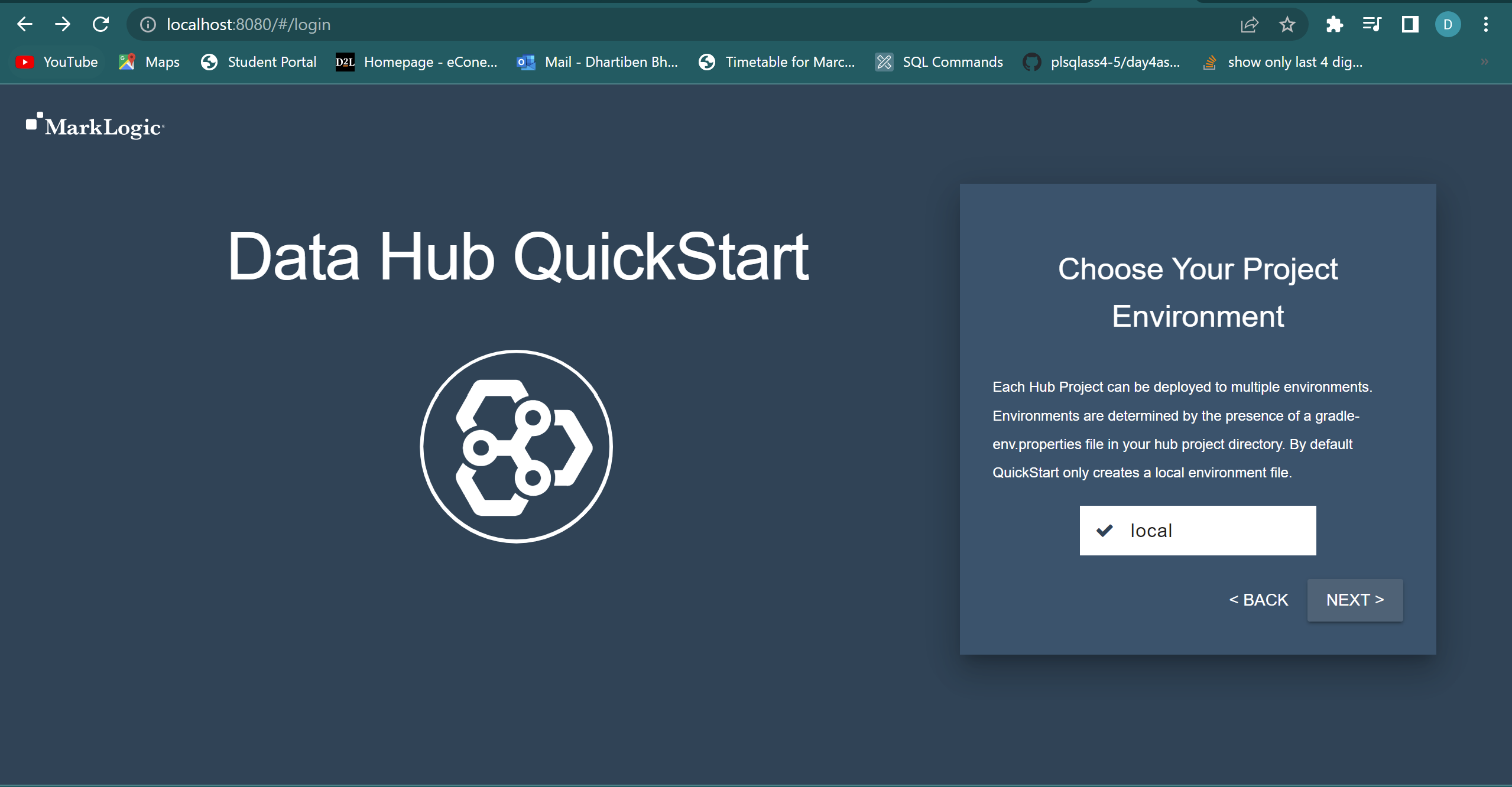
STEP 3: Setting up Datahub by running java -jar marklogic-datahub.5.2.2.war

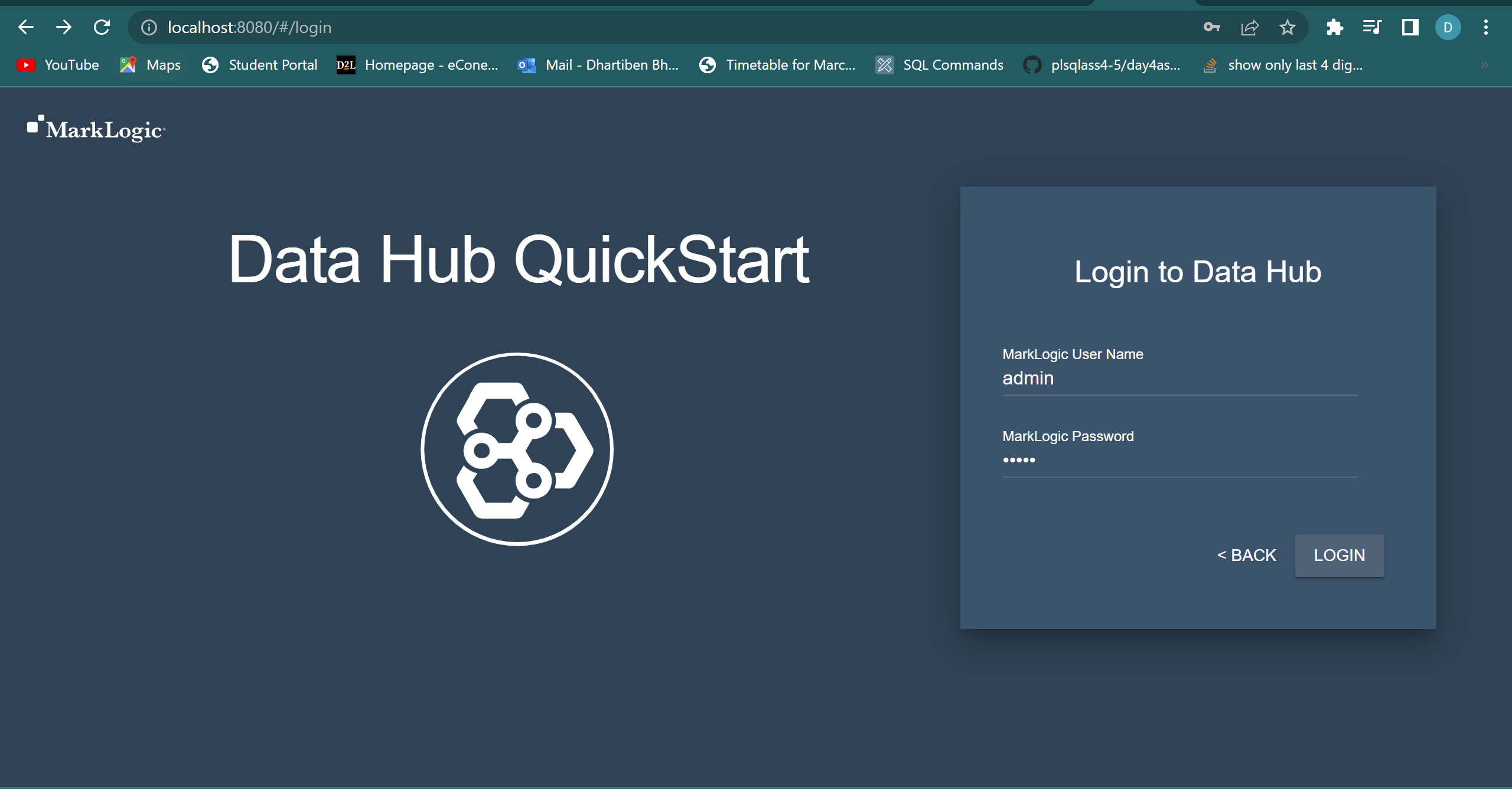
As I have downloaded Datahub version 5.2.2, have to run that version. Once it is downloaded, it is moved to Marklogic folder on Desktop.

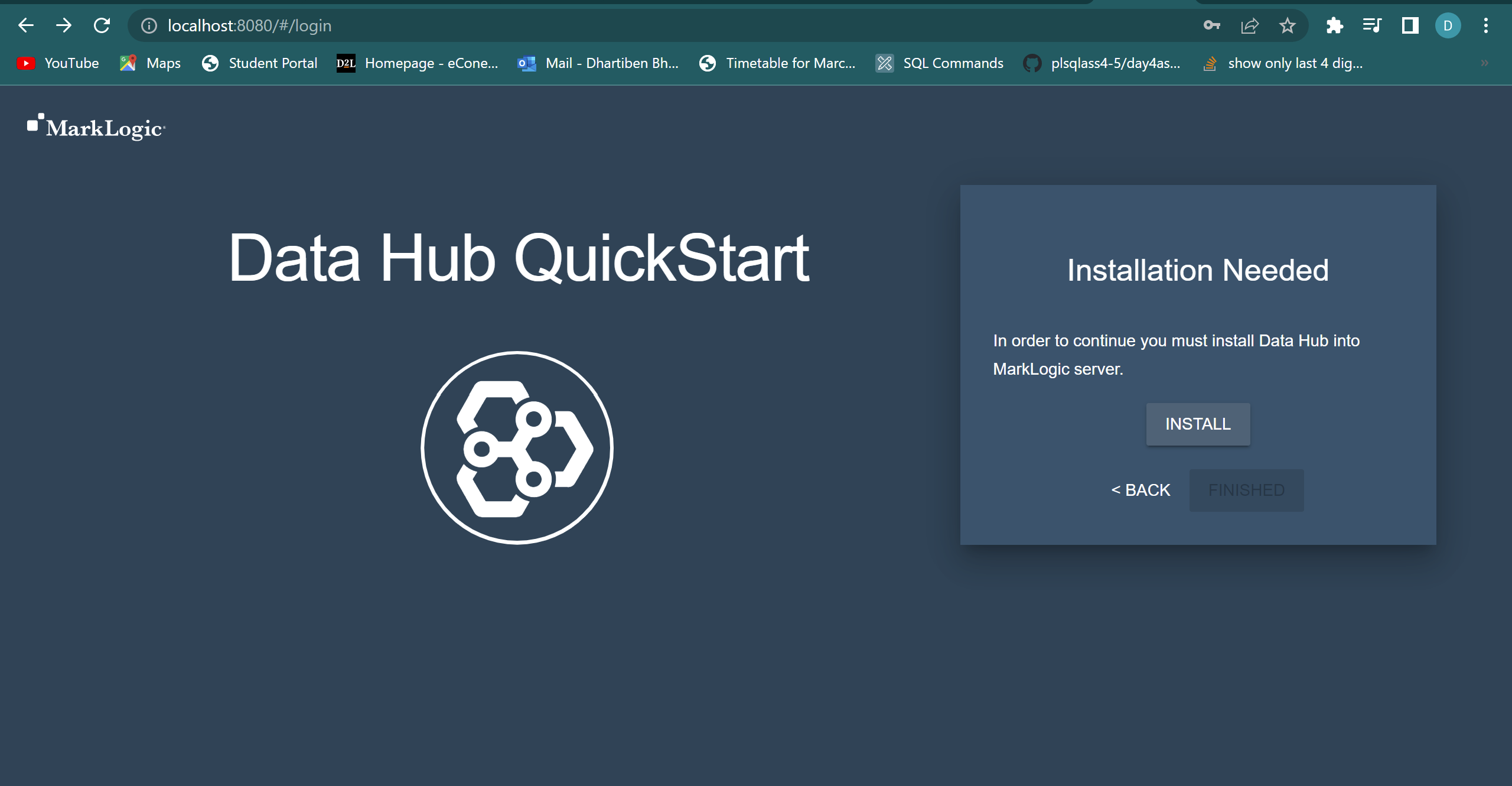
STEP 4: Open localhost:8080 to start up with Datahub and set up Data Hub Qickstart.

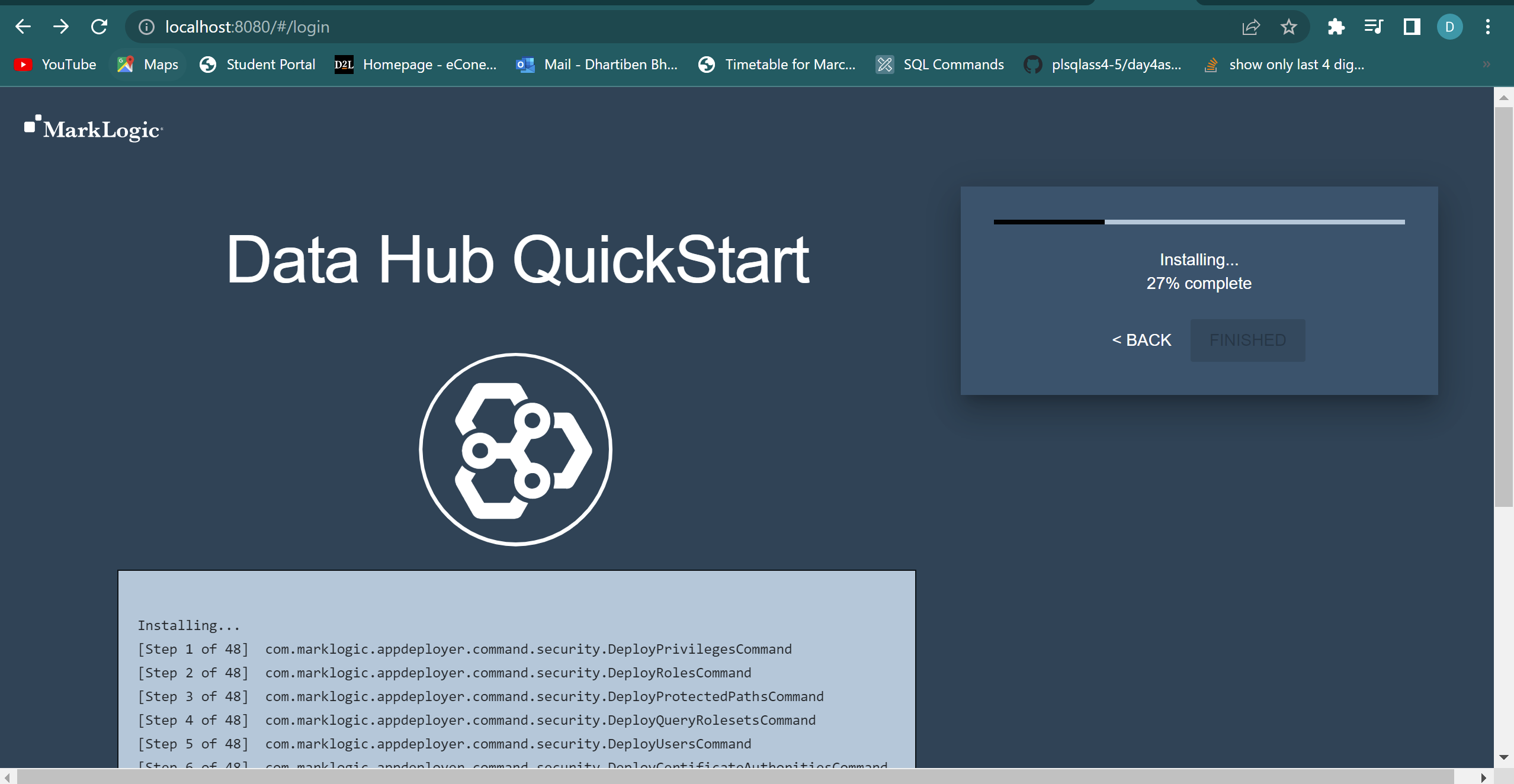


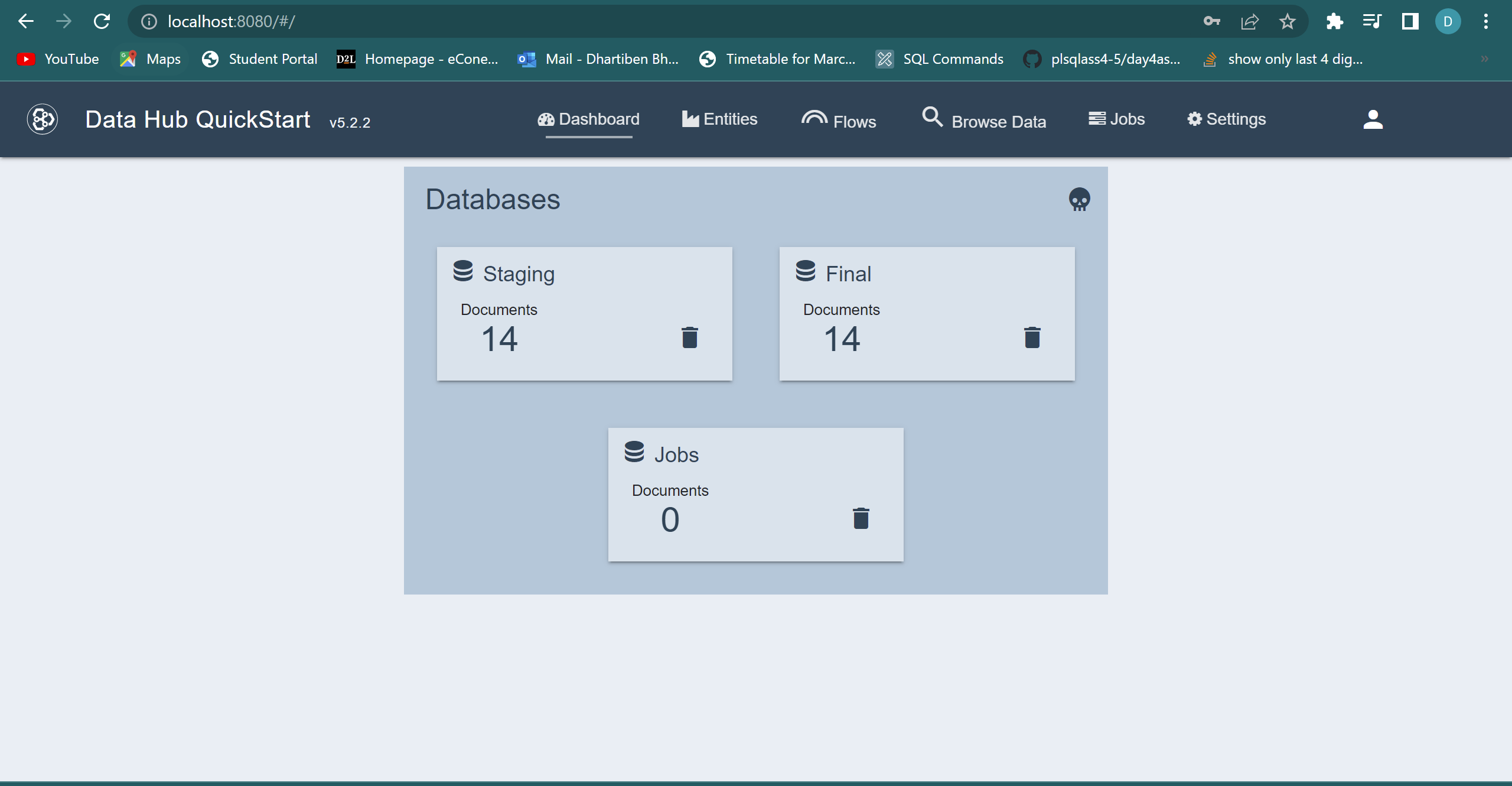


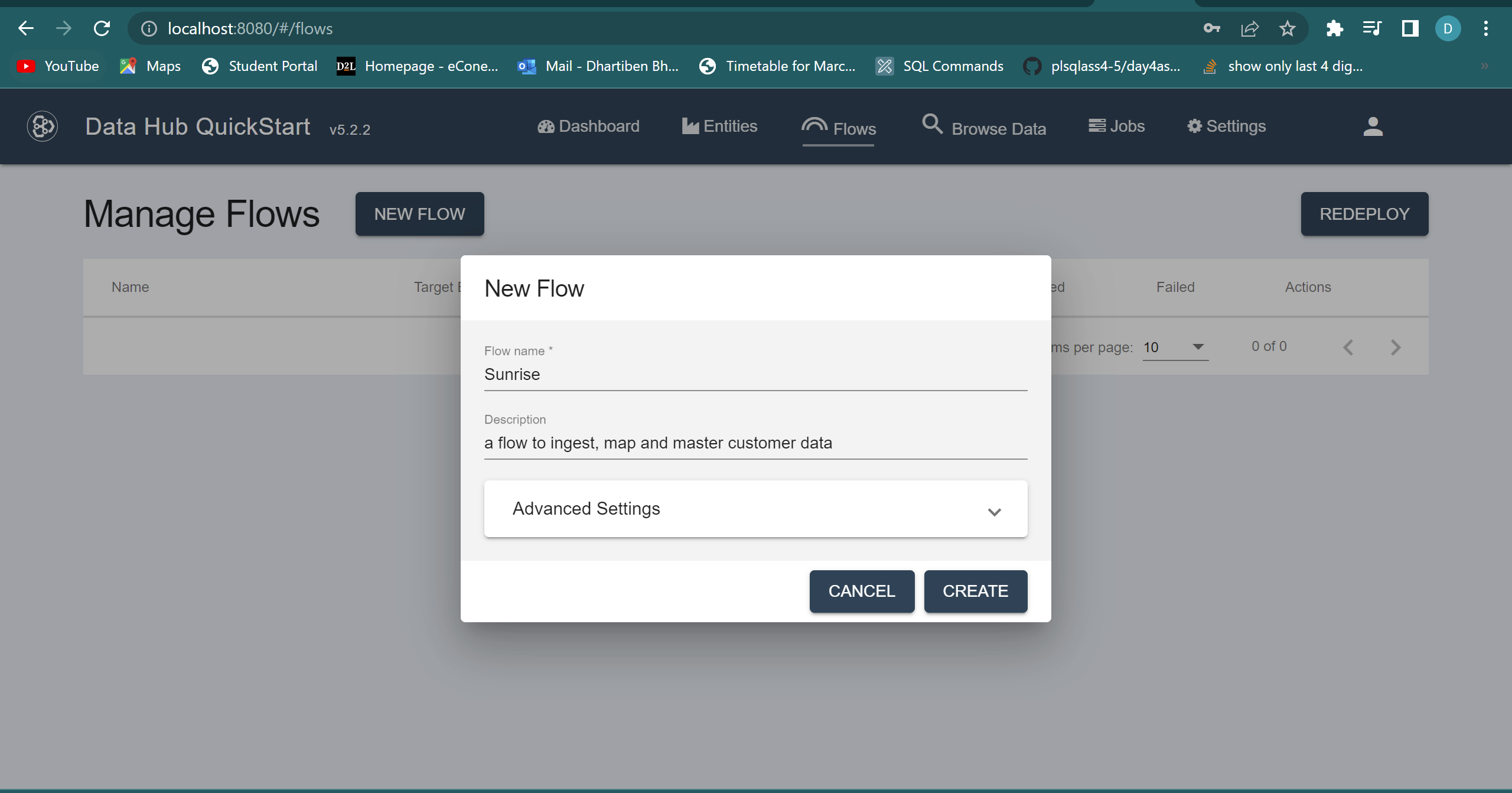




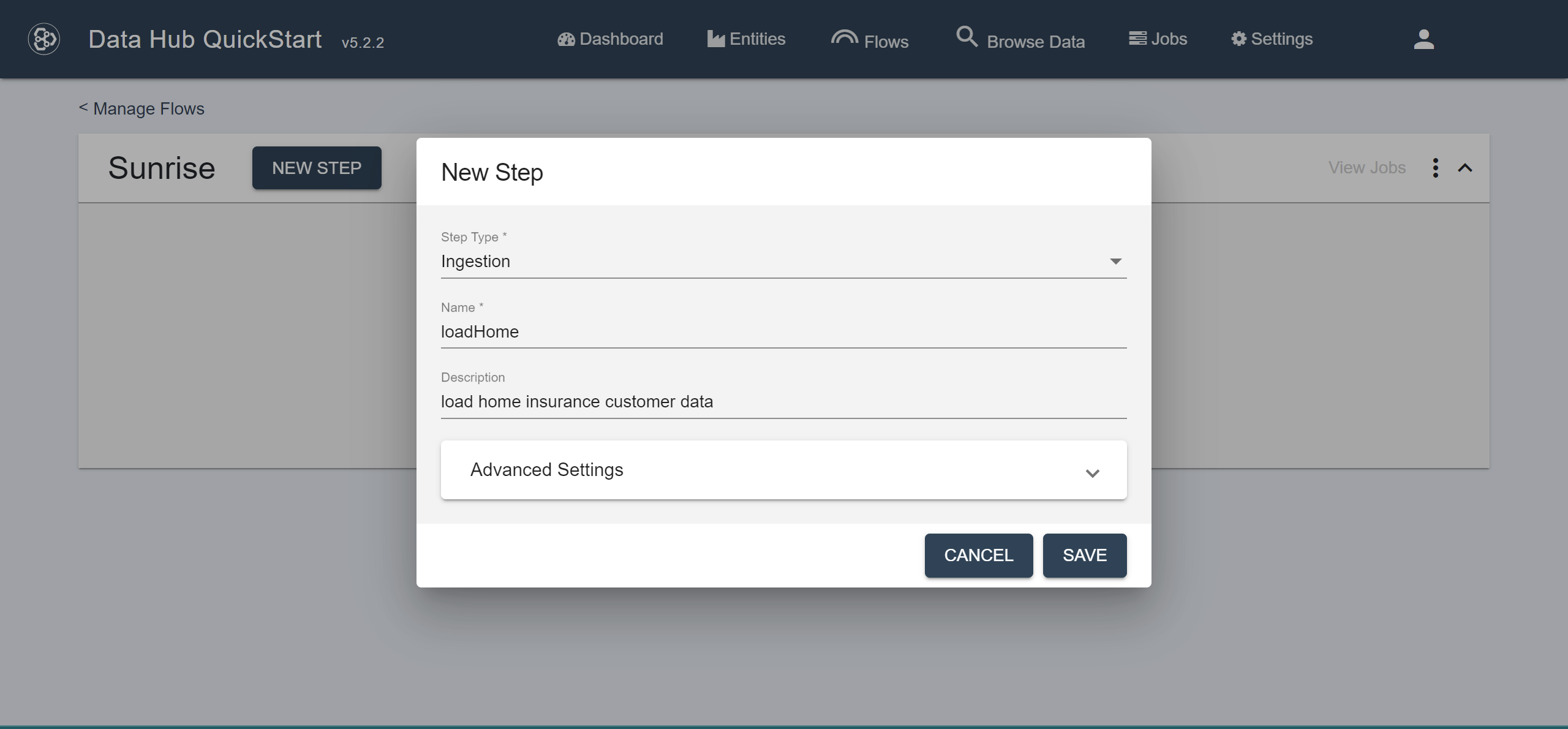


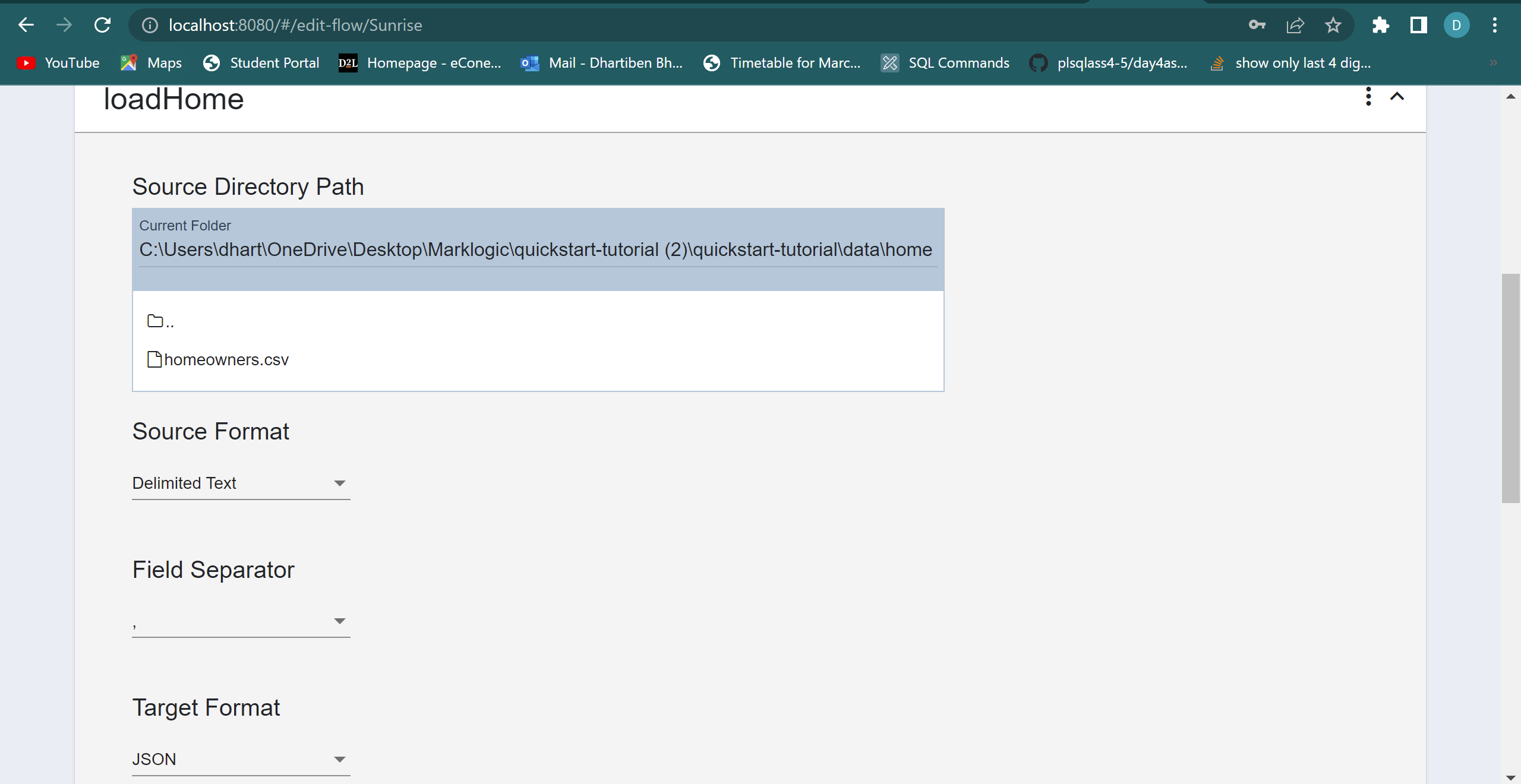


STEP 5: Logged in using the password and username which we created at the time of setting up Marklogic.

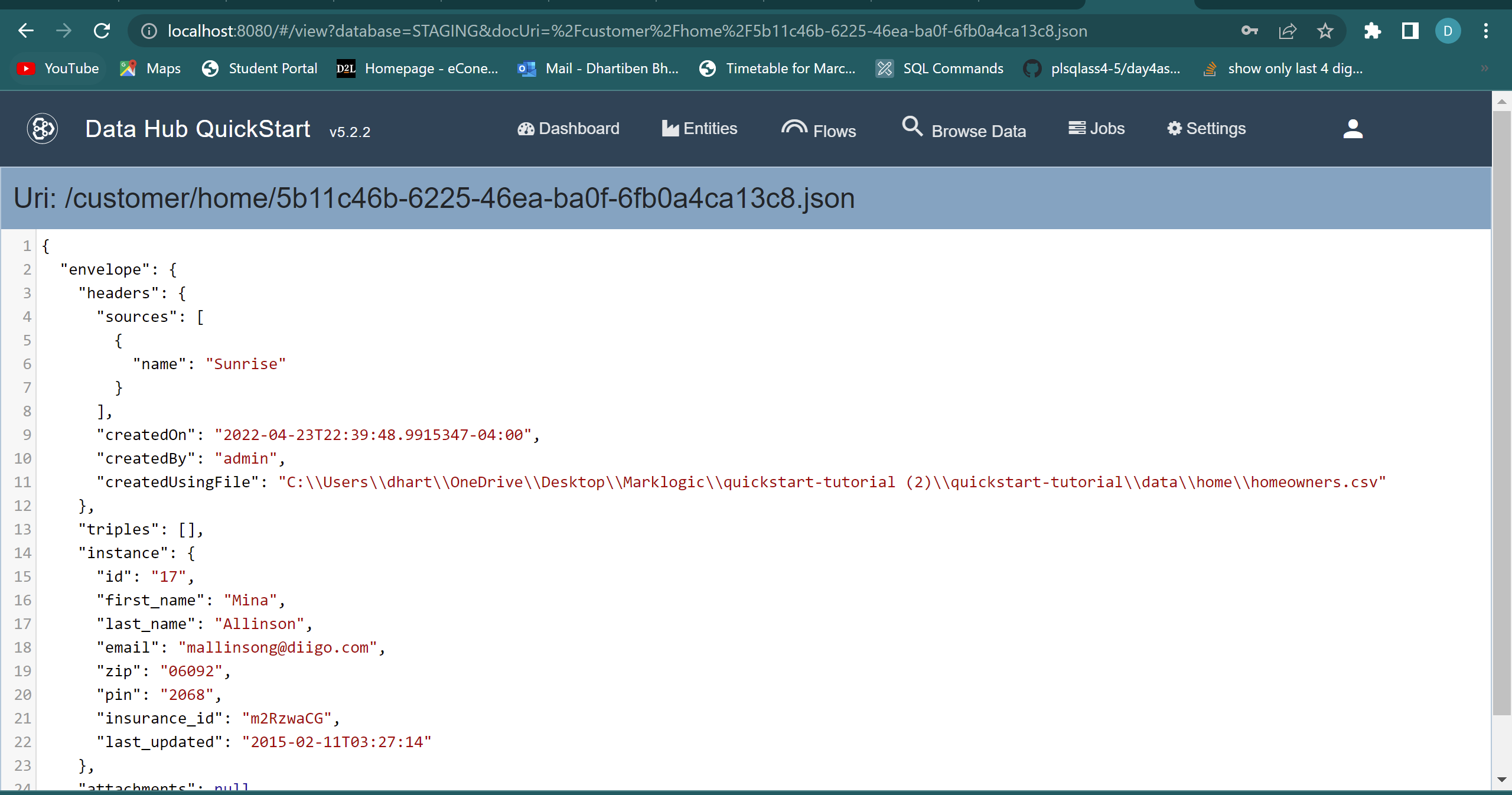
STEP 6: Creating a flow for Sunrise Customer data.

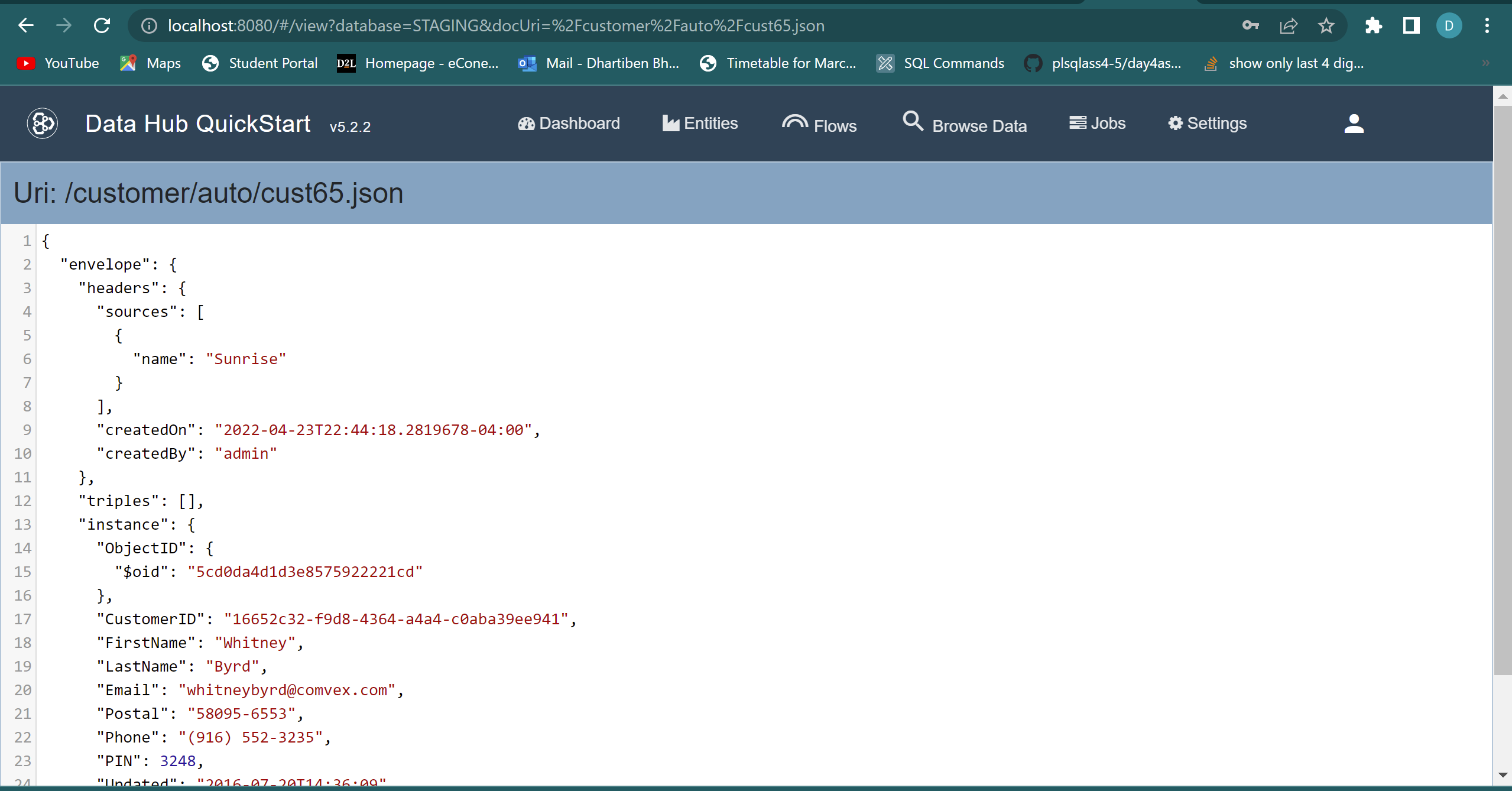
STEP 7: Loading the data which is ready sample data downloaded.





STEP 8: Running the Sunrise Flow



STEP 9: Loading Auto insurance data of customers which is JSON file and loading the data.

We've now put our data into our staging database from the original sources. We loaded data from our file system for simplicity and to make it easy for you to follow along on your own PC. Keep in mind that in a real-world project, data can flow straight into the MarkLogic Data Hub from a variety of sources using data orchestration technologies such as Apache NiFi and Mulesoft.

But for now, compare the house and car customer data that we placed into our data hub from the file system. It's worth noting that both kinds of data are about the same broad business object: a client. They share several characteristics, such as the customer's name. Each source's schema, on the other hand, is unique. For example, the auto data has a property named FirstName that contains the customer's first name, but the home data has a property called first name that contains the customer's first name.

**Difficulties faced during the setup:**

The main hard situation was, that I was not able to log in to the Data Hub Quickstart as Marklogic was not supporting the old version. After that, I downloaded, Data Hub Quickstart version 5.2.2 and copied the .war file to the main folder of the Marklogic project. At the same time, the mistake I did was, that I did not start the Marklogic server while running Data Hub. Once I started it from machine services, it all worked well. I gained knowledge about how to access data from the hub.

Now, **What is** **Curate?**

The process of data curation is done in order to model the data in order to get it into a shape that can power the data services you are going to deliver. Curation makes your data better– better suited to deliver the data service your customer needs.

Curation starts with creating entities and defining the key data properties that your data services will need to consume. From there, you may take the many different shapes of data that you have loaded from various systems and map key properties to that entity configuration. You might also have requirements that require you to enrich the data by iteratively processing, identifying, and tagging references within the data, as well as transforming properties, modeling relationships between entities using triples, or mastering your data to match and merge duplicates.

**Reference**

* <https://developer.marklogic.com/learn/data-hub-central/>
* <https://docs.marklogic.com/guide/installation/procedures#id_28962>
* <https://developer.marklogic.com/>
* <https://www.youtube.com/watch?v=_lwXBb4hhHs>