APRIL 23, 2022

A

GROUP PROJECT

DHARTI PATEL
8807575
ZARNA GOHIL
8800060
BASVARAJ JALIMINCHE
8800149

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.

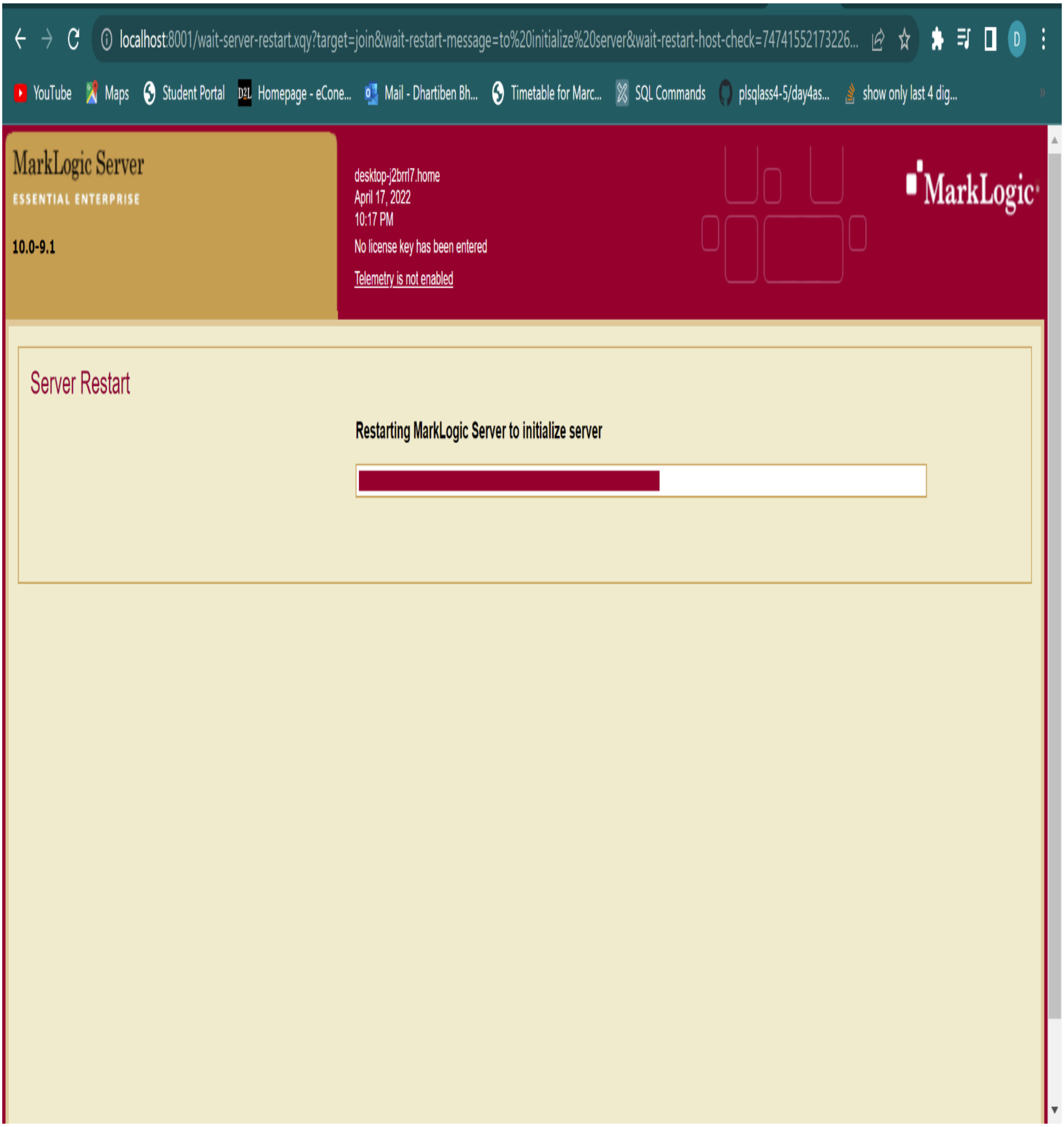
MarkLogic Server currently operates in a variety of industries. Though the data stored in and extracted from MarkLogic is different in each type of industry, many customers have similar data-management challenges.

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



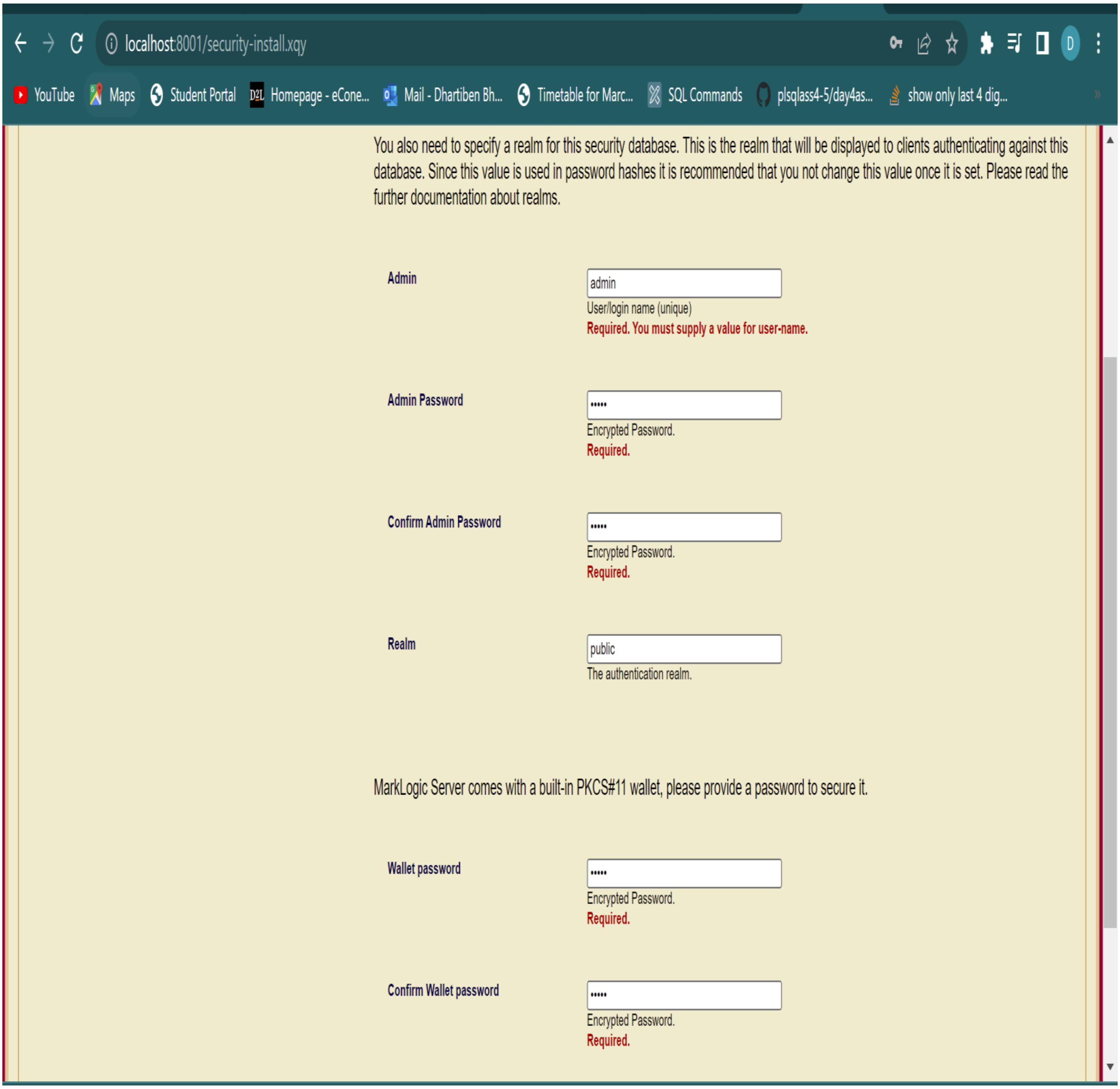
10.0-9.1

desktop-j2brri7.home
April 17, 2022
10:17 PM
No license key has been entered
Telemetry is not enabled

Server Restart

Restarting MarkLogic Server to initialize server

STEP 2: Setting up password and admin



You also need to specify a realm for this security database. This is the realm that will be displayed to clients authenticating against this database. Since this value is used in password hashes it is recommended that you not change this value once it is set. Please read the further documentation about realms.

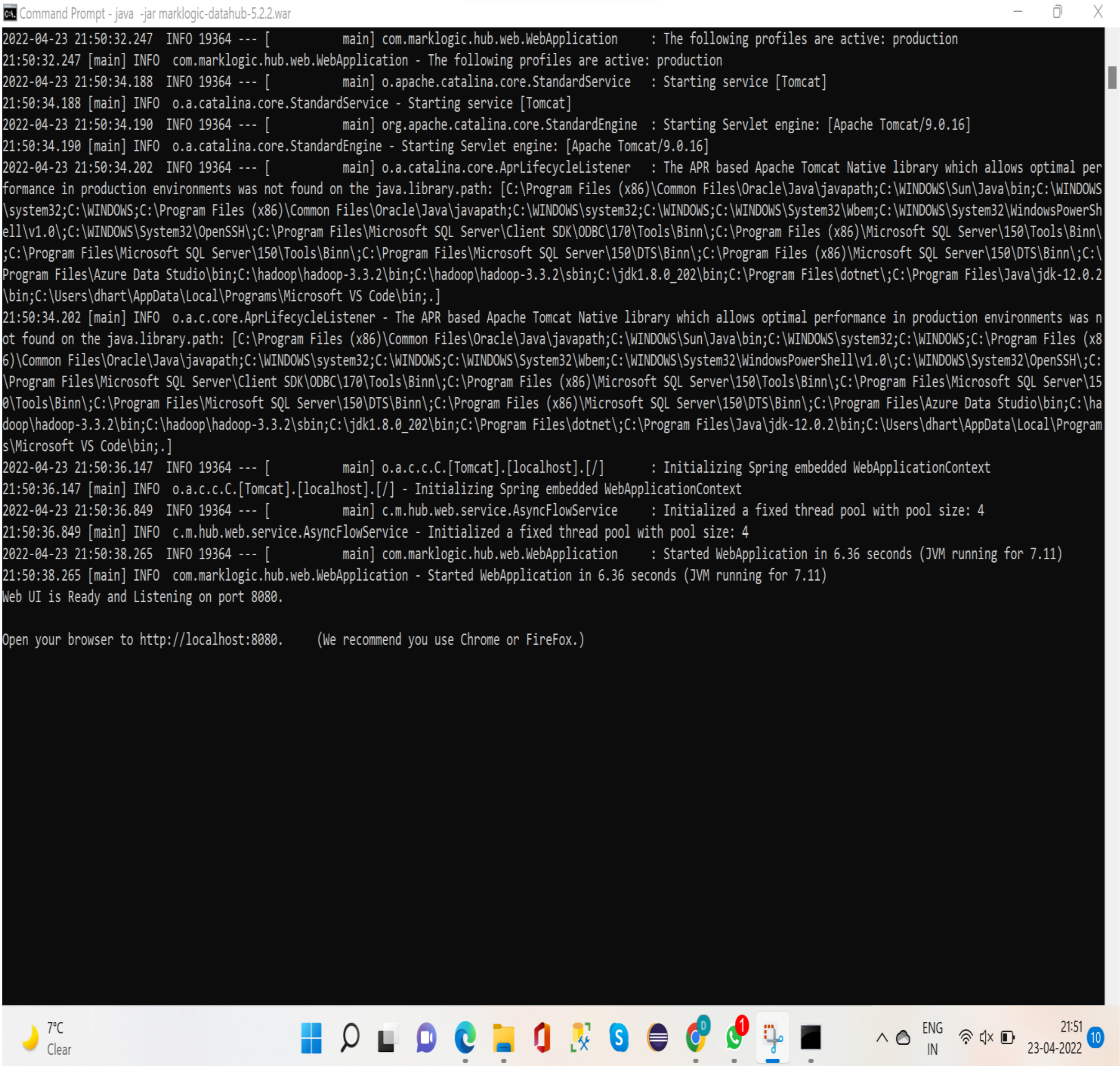
Admin	admin User/login name (unique) Required. You must supply a value for user-name.
Admin Password	••••• Encrypted Password. Required.
Confirm Admin Password	••••• Encrypted Password. Required.
Realm	public The authentication realm.

MarkLogic Server comes with a built-in PKCS#11 wallet, please provide a password to secure it.

Wallet password	••••• Encrypted Password. Required.
Confirm Wallet password	••••• Encrypted Password. Required.

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

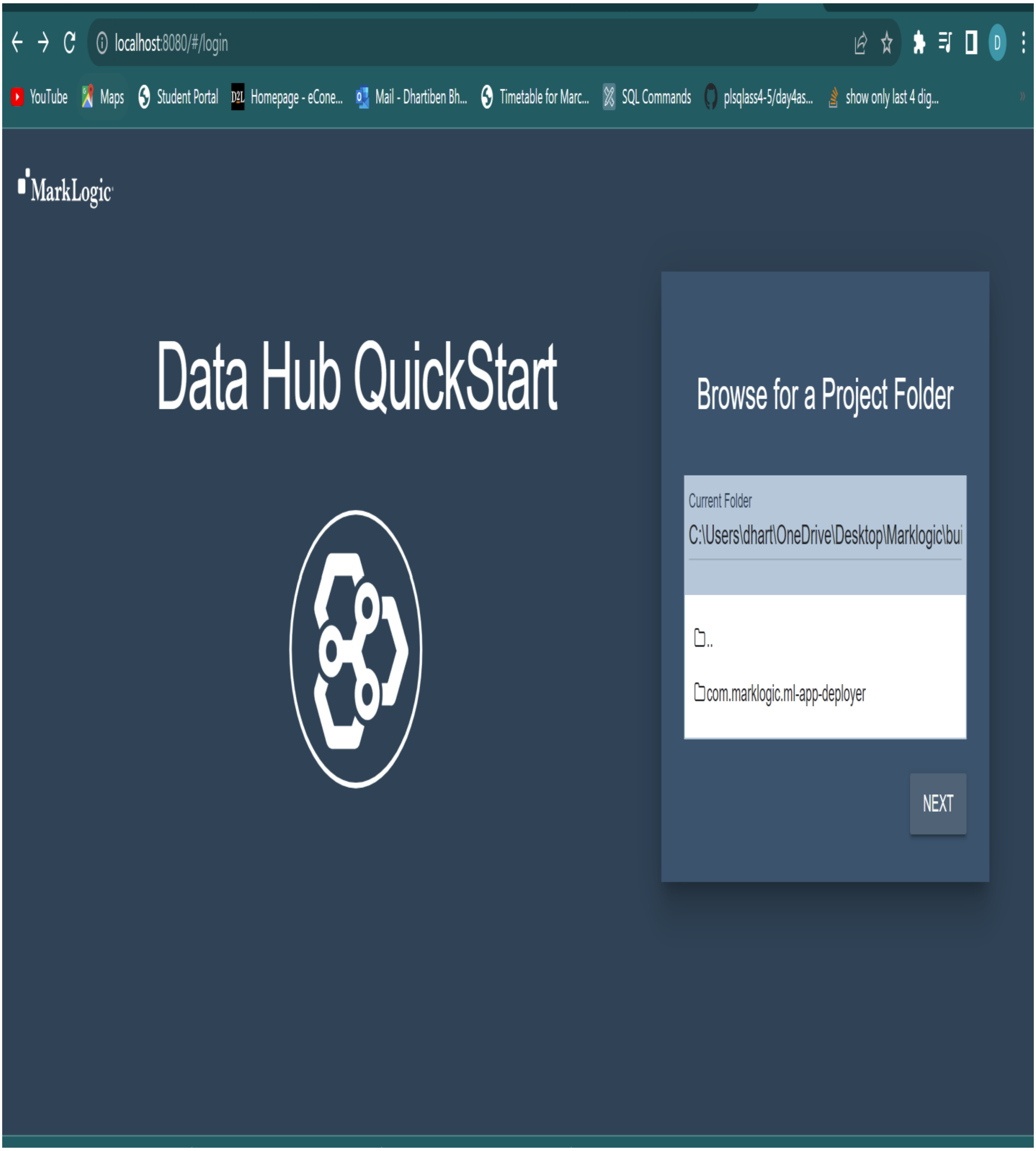
As I have downloaded Datahub version 5.2.2, have to run that version. Once it is downloaded, it is moved to Marklogic folder on Desktop.



```
2022-04-23 21:50:32.247  INFO 19364 --- [           main] com.marklogic.hub.web.WebApplication     : The following profiles are active: production
21:50:32.247 [main] INFO  com.marklogic.hub.web.WebApplication - The following profiles are active: production
2022-04-23 21:50:34.188  INFO 19364 --- [           main] o.apache.catalina.core.StandardService   : Starting service [Tomcat]
21:50:34.188 [main] INFO  o.a.catalina.core.StandardService - Starting service [Tomcat]
2022-04-23 21:50:34.190  INFO 19364 --- [           main] org.apache.catalina.core.StandardEngine  : Starting Servlet engine: [Apache Tomcat/9.0.16]
21:50:34.190 [main] INFO  o.a.catalina.core.StandardEngine - Starting Servlet engine: [Apache Tomcat/9.0.16]
2022-04-23 21:50:34.202  INFO 19364 --- [           main] o.a.catalina.core.AprLifecycleListener   : The APR based Apache Tomcat Native library which allows optimal per
formance in production environments was not found on the java.library.path: [C:\Program Files (x86)\Common Files\Oracle\Java\javapath;C:\WINDOWS\Sun\Java\bin;C:\WINDOWS
\system32;C:\WINDOWS;C:\Program Files (x86)\Common Files\Oracle\Java\javapath;C:\WINDOWS\system32;C:\WINDOWS;C:\WINDOWS\System32\Wbem;C:\WINDOWS\System32\WindowsPowerSh
ell\v1.0\;C:\WINDOWS\System32\OpenSSH\;C:\Program Files\Microsoft SQL Server\Client SDK\ODBC\170\Tools\Binn\;C:\Program Files (x86)\Microsoft SQL Server\150\Tools\Binn\
;C:\Program Files\Microsoft SQL Server\150\Tools\Binn\;C:\Program Files\Microsoft SQL Server\150\DTS\Binn\;C:\Program Files (x86)\Microsoft SQL Server\150\DTS\Binn\;C:\
Program Files\Azure Data Studio\bin;C:\hadoop\hadoop-3.3.2\bin;C:\hadoop\hadoop-3.3.2\sbin;C:\jdk1.8.0_202\bin;C:\Program Files\dotnet\;C:\Program Files\Java\jdk-12.0.2
\bin;C:\Users\dhart\AppData\Local\Programs\Microsoft VS Code\bin;.]
21:50:34.202 [main] INFO  o.a.c.core.AprLifecycleListener - The APR based Apache Tomcat Native library which allows optimal performance in production environments was n
ot found on the java.library.path: [C:\Program Files (x86)\Common Files\Oracle\Java\javapath;C:\WINDOWS\Sun\Java\bin;C:\WINDOWS\system32;C:\WINDOWS;C:\Program Files (x8
6)\Common Files\Oracle\Java\javapath;C:\WINDOWS\system32;C:\WINDOWS;C:\WINDOWS\System32\Wbem;C:\WINDOWS\System32\WindowsPowerShell\v1.0\;C:\WINDOWS\System32\OpenSSH\;C:
\Program Files\Microsoft SQL Server\Client SDK\ODBC\170\Tools\Binn\;C:\Program Files (x86)\Microsoft SQL Server\150\Tools\Binn\;C:\Program Files\Microsoft SQL Server\15
0\Tools\Binn\;C:\Program Files\Microsoft SQL Server\150\DTS\Binn\;C:\Program Files (x86)\Microsoft SQL Server\150\DTS\Binn\;C:\Program Files\Azure Data Studio\bin;C:\ha
doop\hadoop-3.3.2\bin;C:\hadoop\hadoop-3.3.2\sbin;C:\jdk1.8.0_202\bin;C:\Program Files\dotnet\;C:\Program Files\Java\jdk-12.0.2\bin;C:\Users\dhart\AppData\Local\Program
s\Microsoft VS Code\bin;.]
2022-04-23 21:50:36.147  INFO 19364 --- [           main] o.a.c.c.C.[Tomcat].[localhost].[/]       : Initializing Spring embedded WebApplicationContext
21:50:36.147 [main] INFO  o.a.c.c.C.[Tomcat].[localhost].[/] - Initializing Spring embedded WebApplicationContext
2022-04-23 21:50:36.849  INFO 19364 --- [           main] c.m.hub.web.service.AsyncFlowService     : Initialized a fixed thread pool with pool size: 4
21:50:36.849 [main] INFO  c.m.hub.web.service.AsyncFlowService - Initialized a fixed thread pool with pool size: 4
2022-04-23 21:50:38.265  INFO 19364 --- [           main] com.marklogic.hub.web.WebApplication     : Started WebApplication in 6.36 seconds (JVM running for 7.11)
21:50:38.265 [main] INFO  com.marklogic.hub.web.WebApplication - Started WebApplication in 6.36 seconds (JVM running for 7.11)
Web UI is Ready and Listening on port 8080.

Open your browser to http://localhost:8080.     (We recommend you use Chrome or FireFox.)
```

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



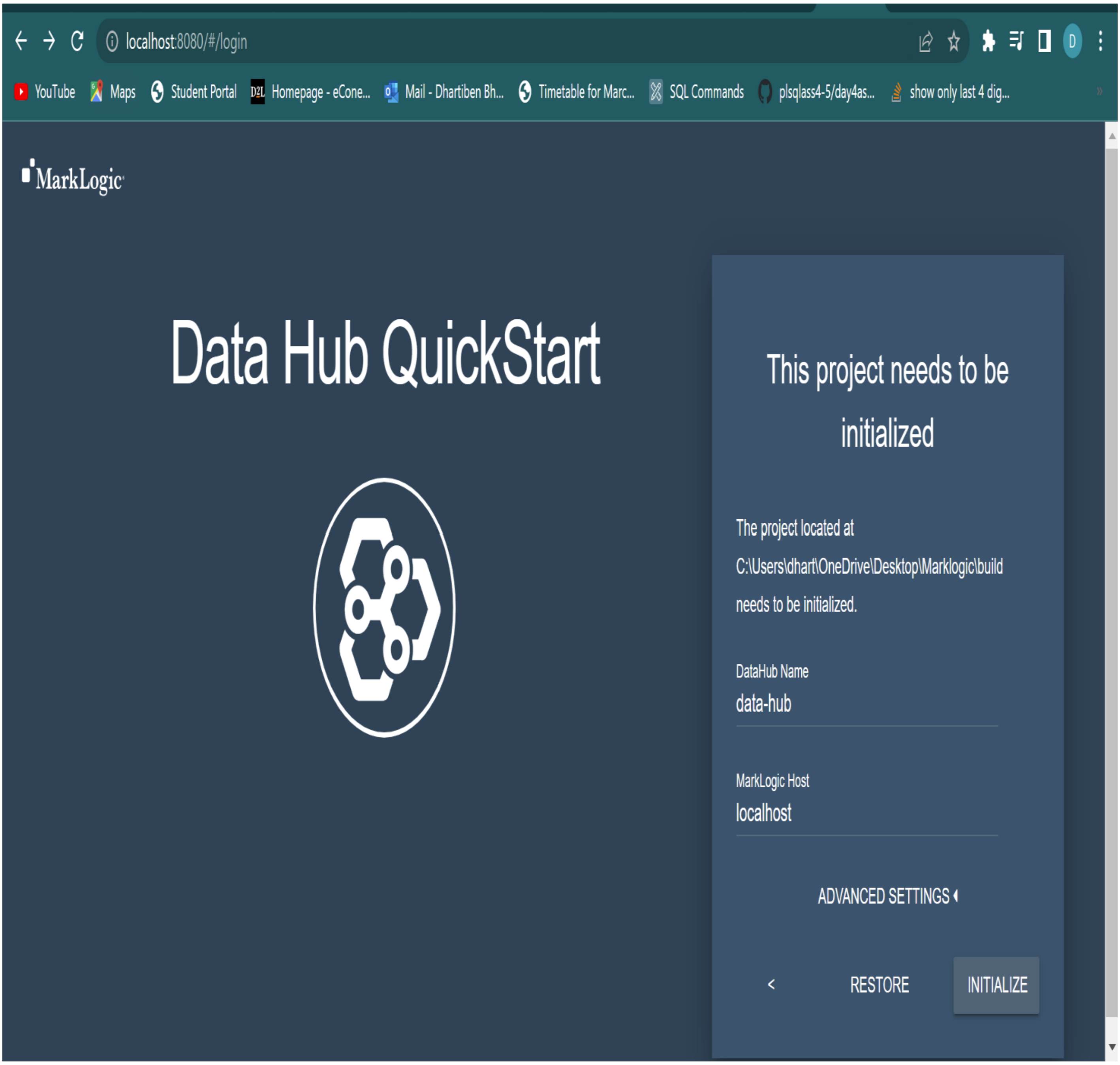
Browse for a Project Folder

Current Folder
C:\Users\dhart\OneDrive\Desktop\Marklogic\bui

..

com.marklogic.ml-app-deployer

NEXT

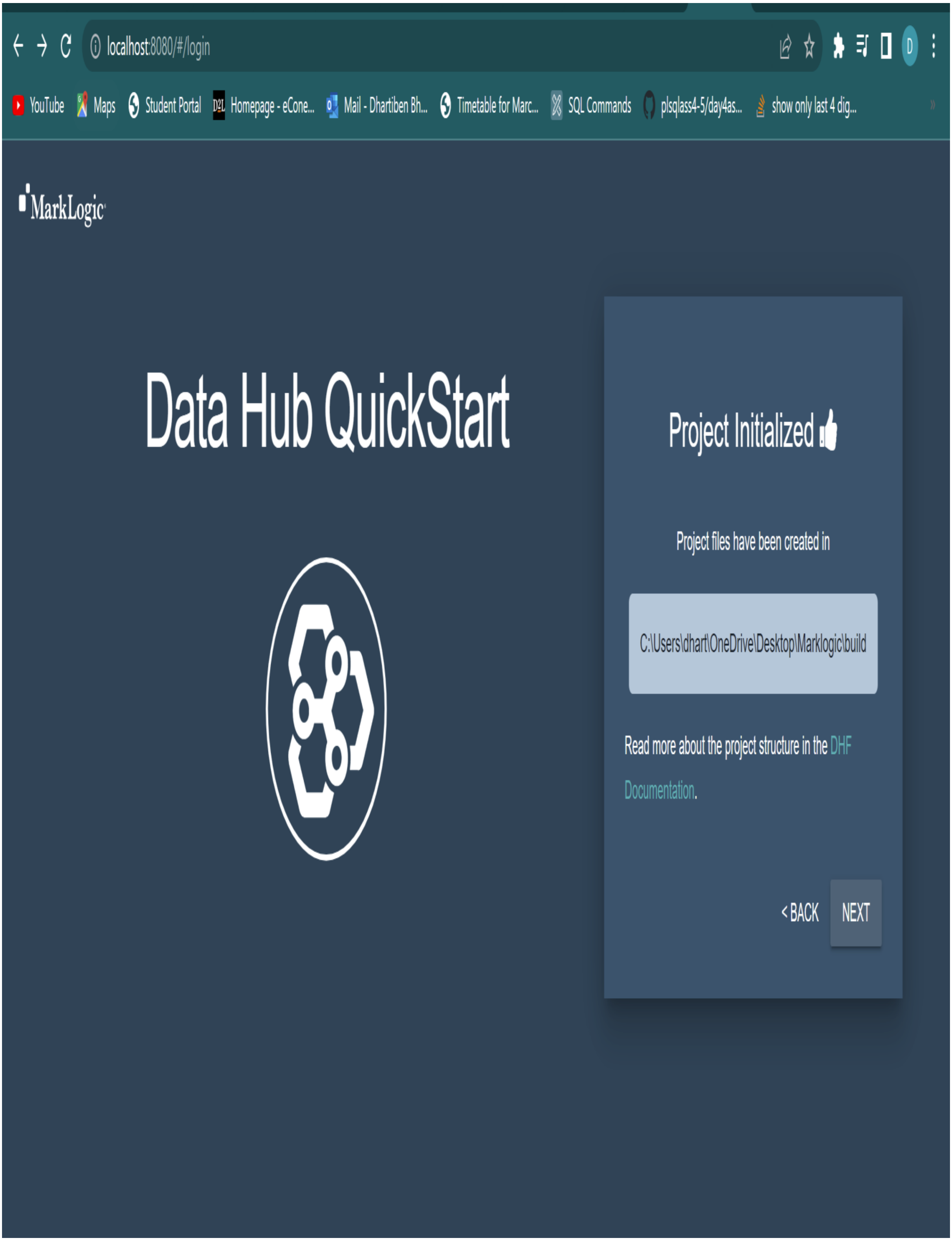


This project needs to be initialized

The project located at C:\Users\dhart\OneDrive\Desktop\Marklogic\build needs to be initialized.

DataHub Name
data-hub

MarkLogic Host
localhost



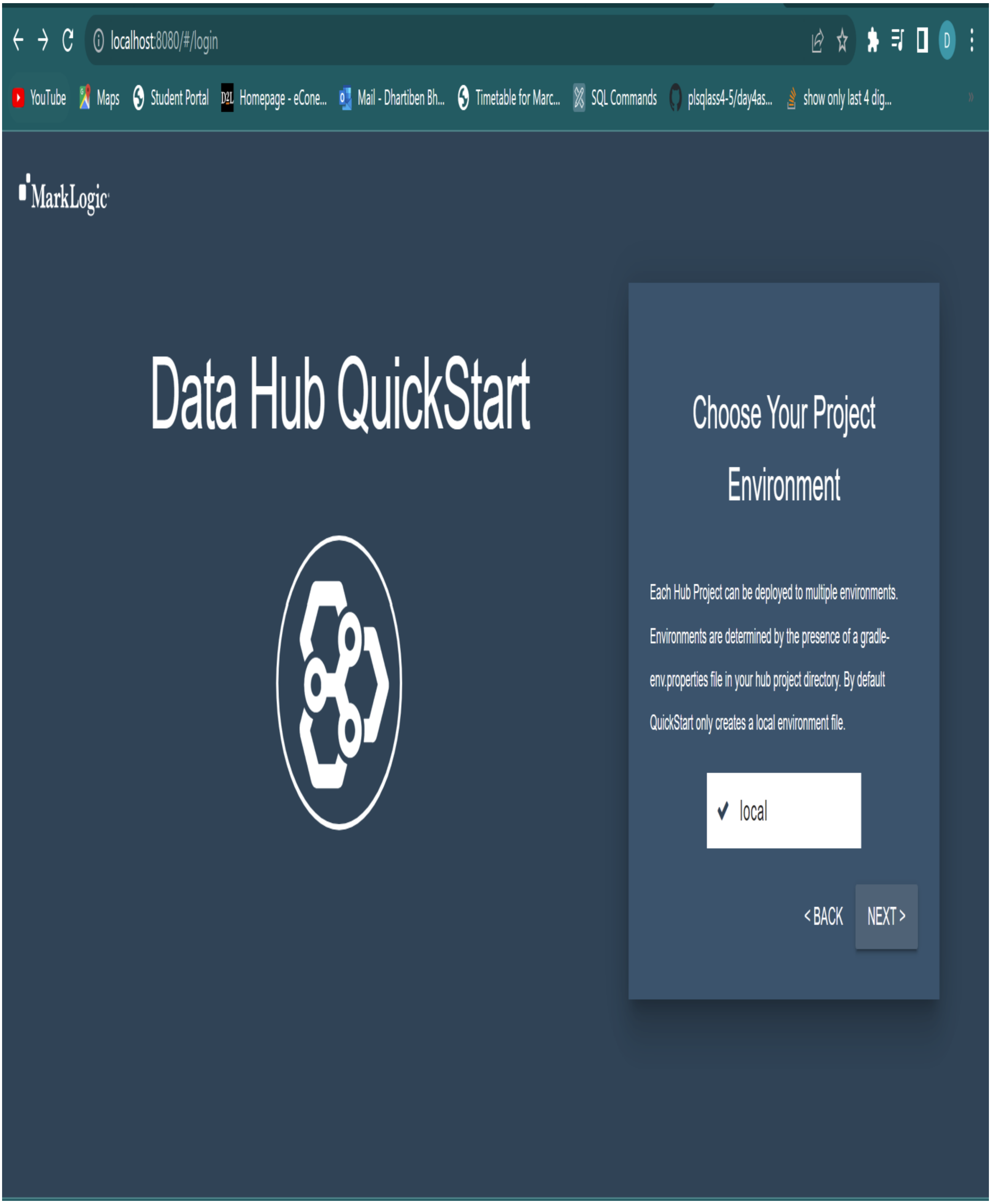
Data Hub QuickStart

Project Initialized

Project files have been created in

C:\Users\dhart\OneDrive\Desktop\Marklogic\build

Read more about the project structure in the DHF Documentation.



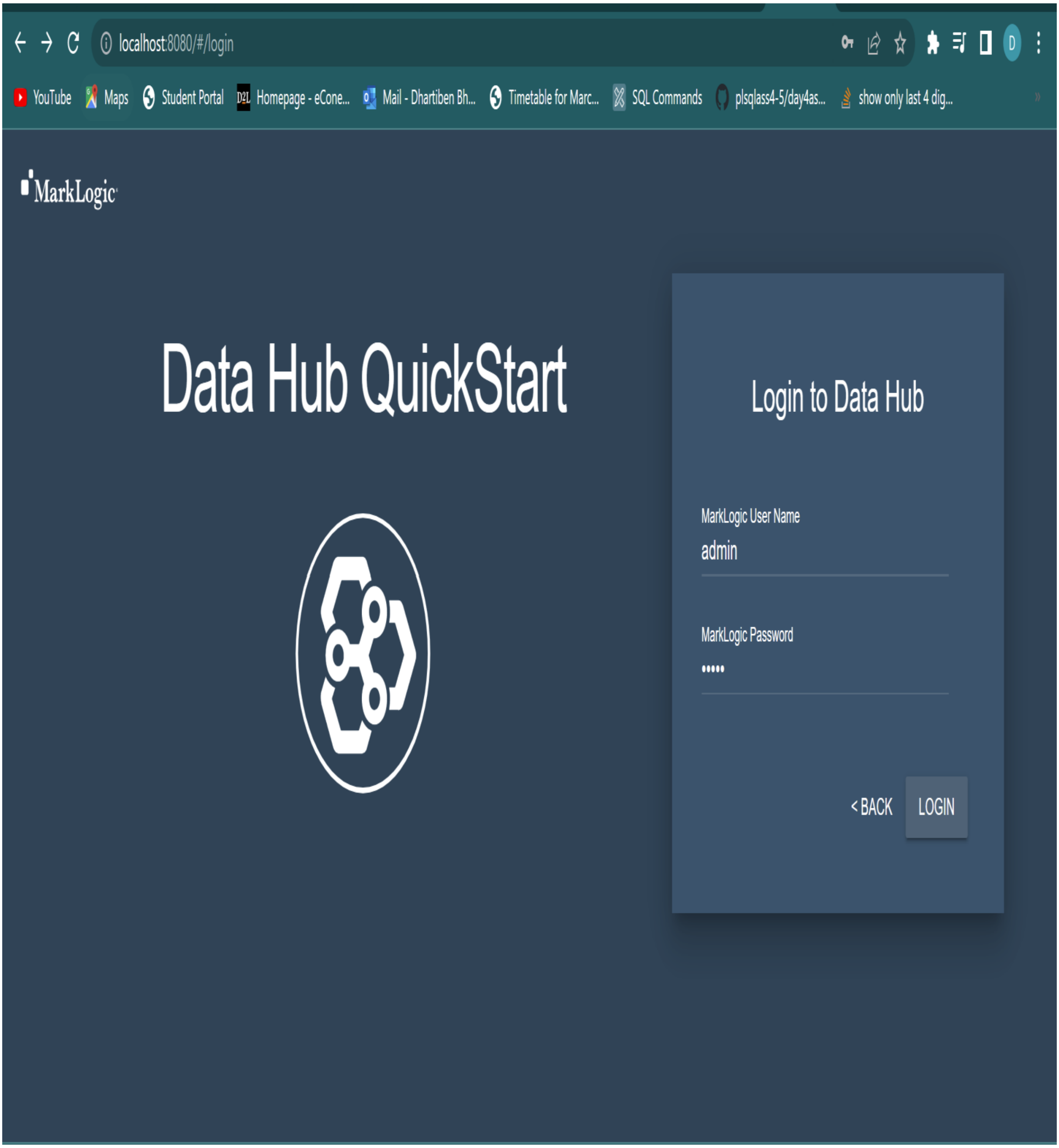
Data Hub QuickStart

Choose Your Project Environment

Each Hub Project can be deployed to multiple environments. Environments are determined by the presence of a gradle-env.properties file in your hub project directory. By default QuickStart only creates a local environment file.

✓ local

< BACK NEXT >



localhost:8080/#/login

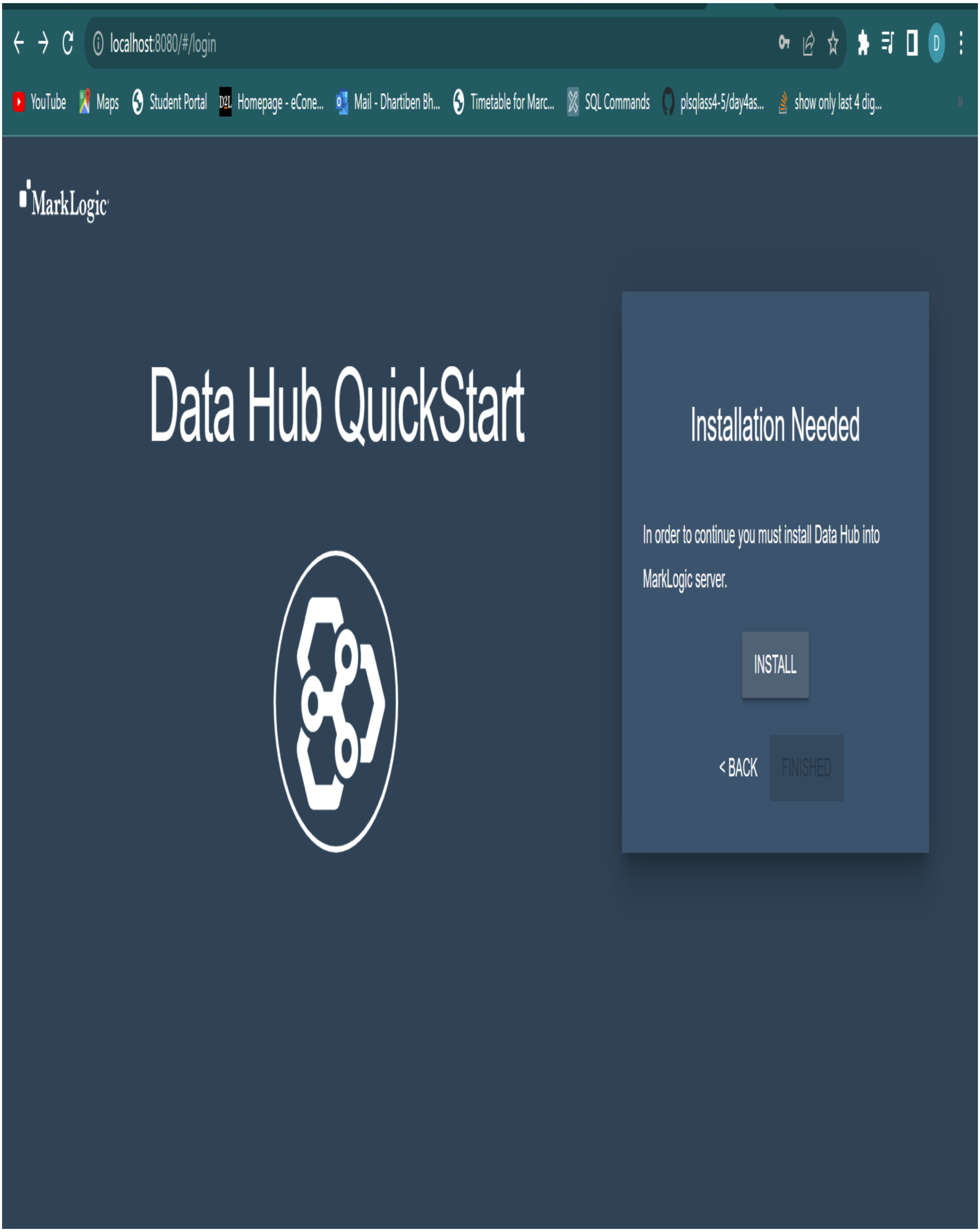
MarkLogic

Data Hub QuickStart

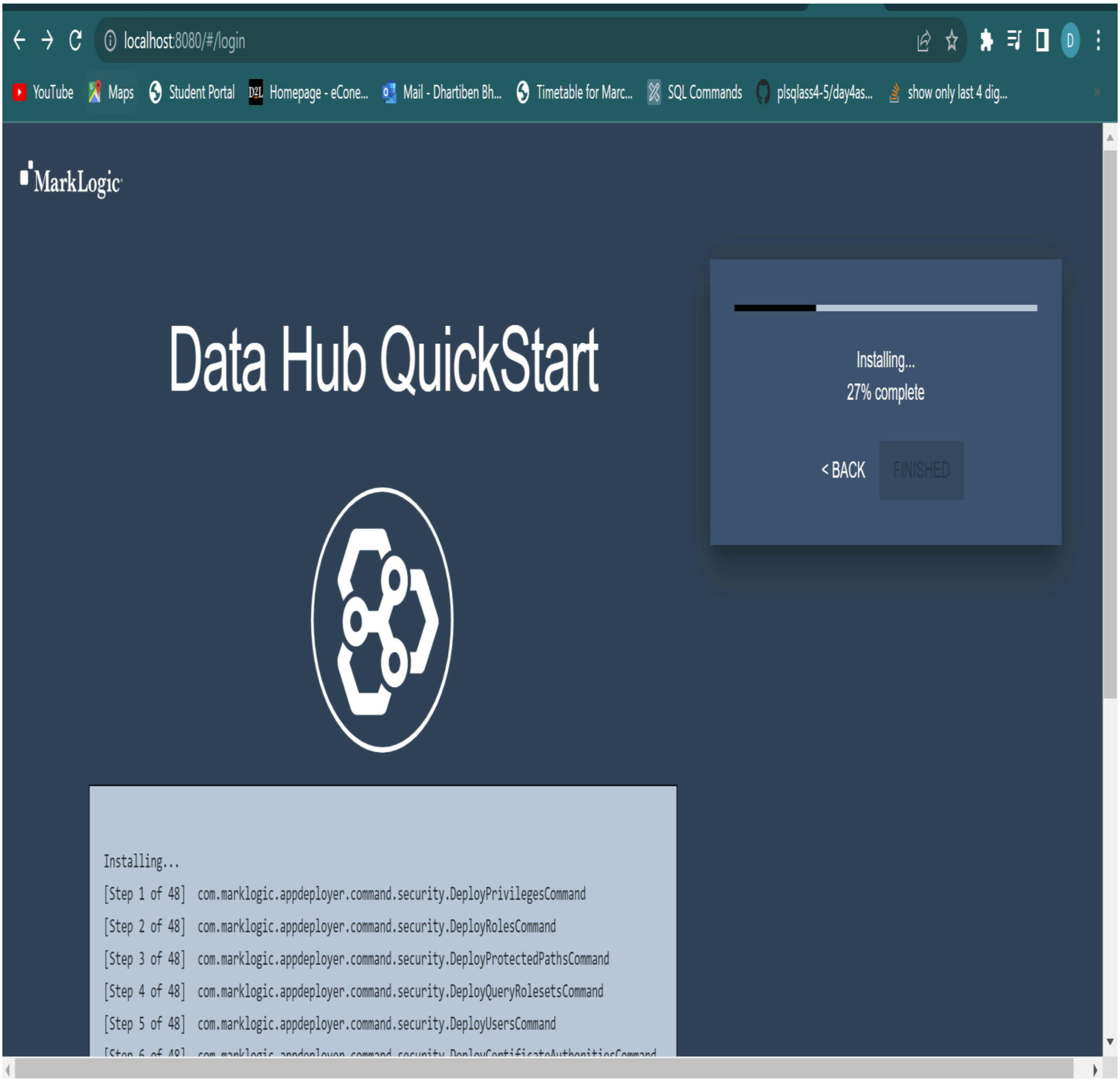
Login to Data Hub

MarkLogic User Name
admin

MarkLogic Password
•••••



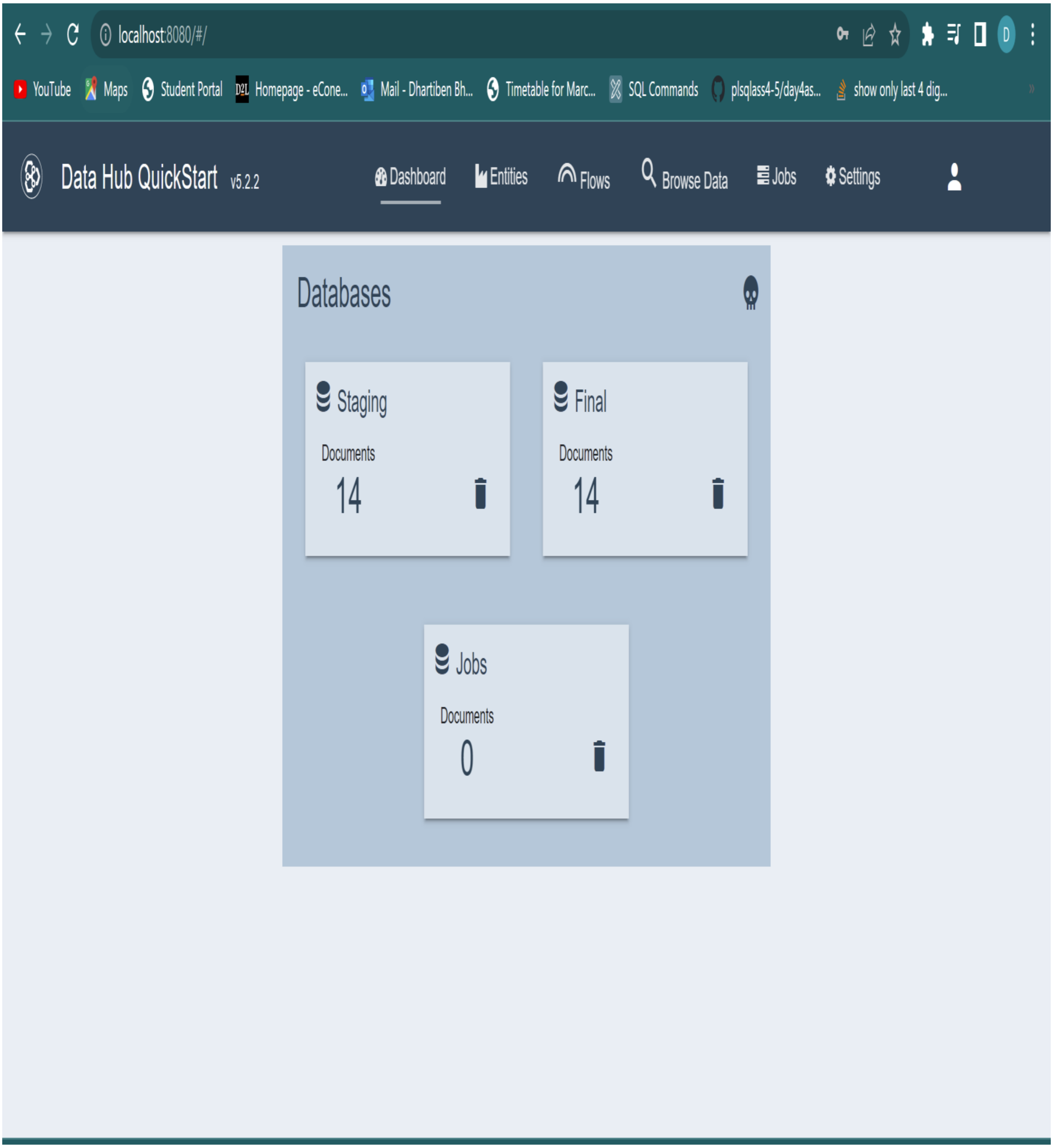
In order to continue you must install Data Hub into MarkLogic server.



Installing...
27% complete

```
Installing...
[Step 1 of 48]  com.marklogic.appdeployer.command.security.DeployPrivilegesCommand
[Step 2 of 48]  com.marklogic.appdeployer.command.security.DeployRolesCommand
[Step 3 of 48]  com.marklogic.appdeployer.command.security.DeployProtectedPathsCommand
[Step 4 of 48]  com.marklogic.appdeployer.command.security.DeployQueryRolesetsCommand
[Step 5 of 48]  com.marklogic.appdeployer.command.security.DeployUsersCommand
[Step 6 of 48]  com.marklogic.appdeployer.command.security.DeployCertificateAuthoritiesCommand
```

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



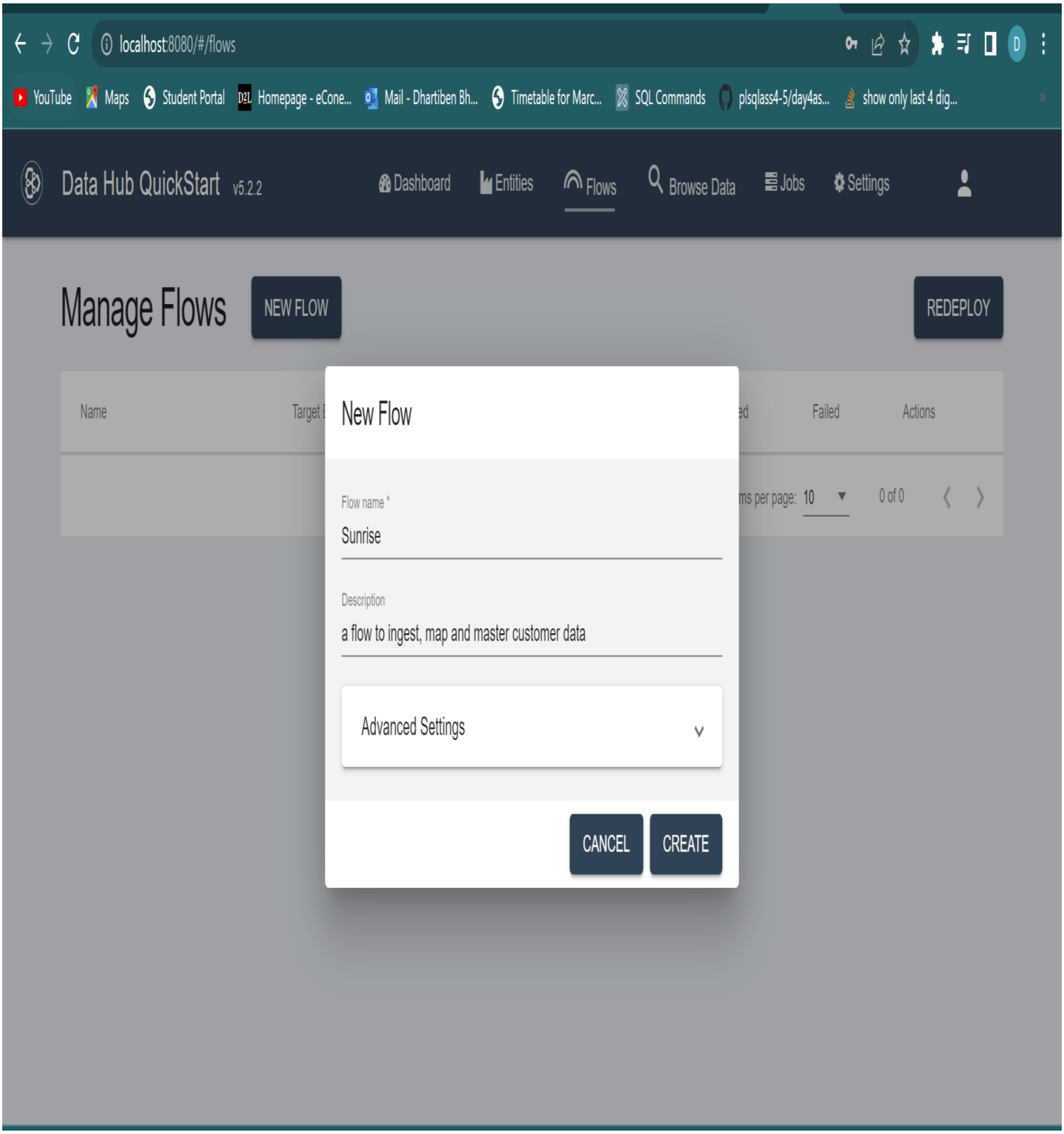
Databases

Staging
Documents
14

Final
Documents
14

Jobs
Documents
0

STEP 6: Creating a flow for Sunrise Customer data.

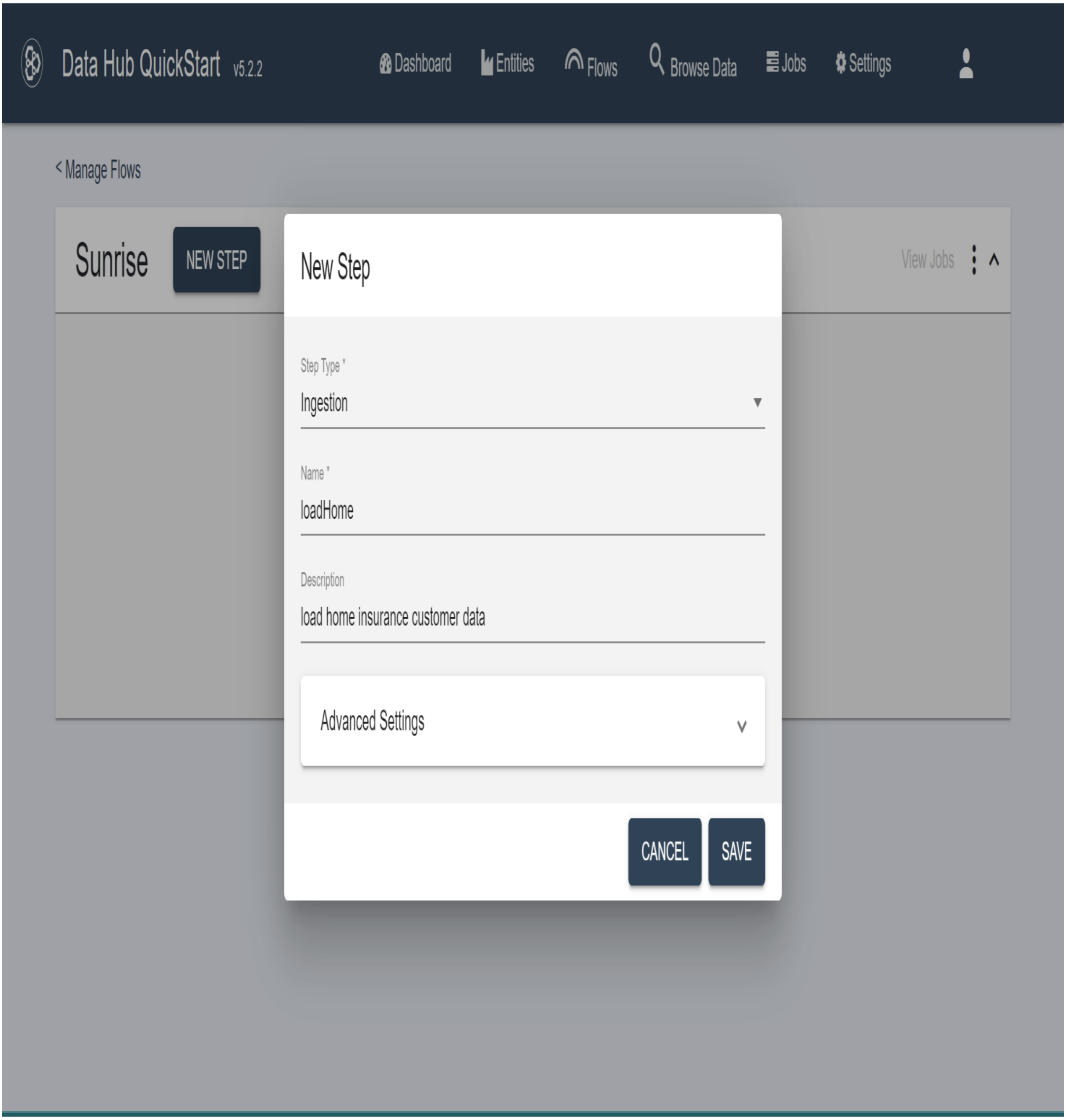


New Flow

Flow name *: Sunrise

Description: a flow to ingest, map and master customer data

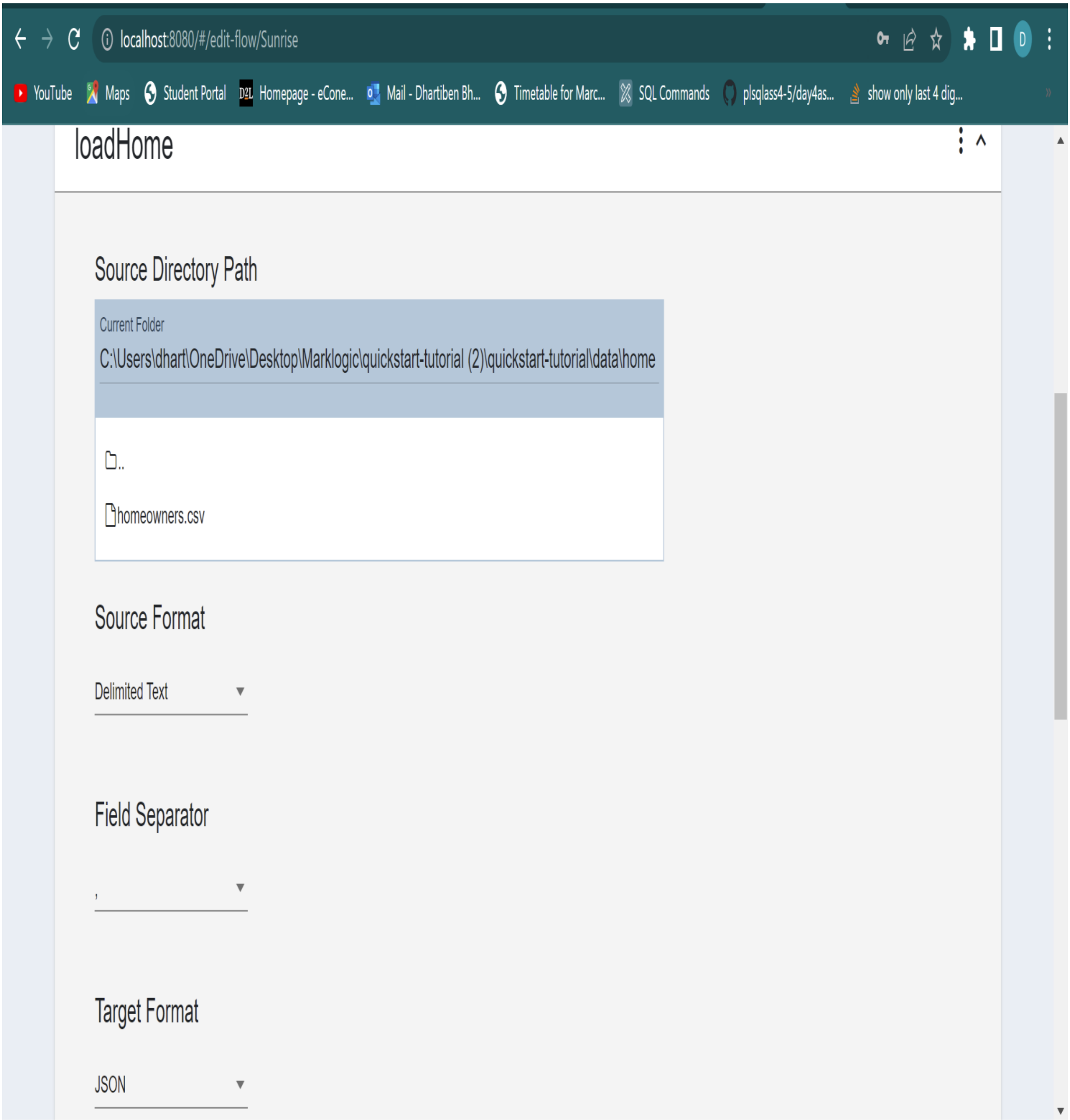
STEP 7: Loading the data which is ready sample data downloaded.



Step Type *
Ingestion

Name *
loadHome

Description
load home insurance customer data



loadHome

Source Directory Path

Current Folder
C:\Users\dhart\OneDrive\Desktop\Marklogic\quickstart-tutorial (2)\quickstart-tutorial\data\home

..

homeowners.csv

Source Format

Delimited Text ▼

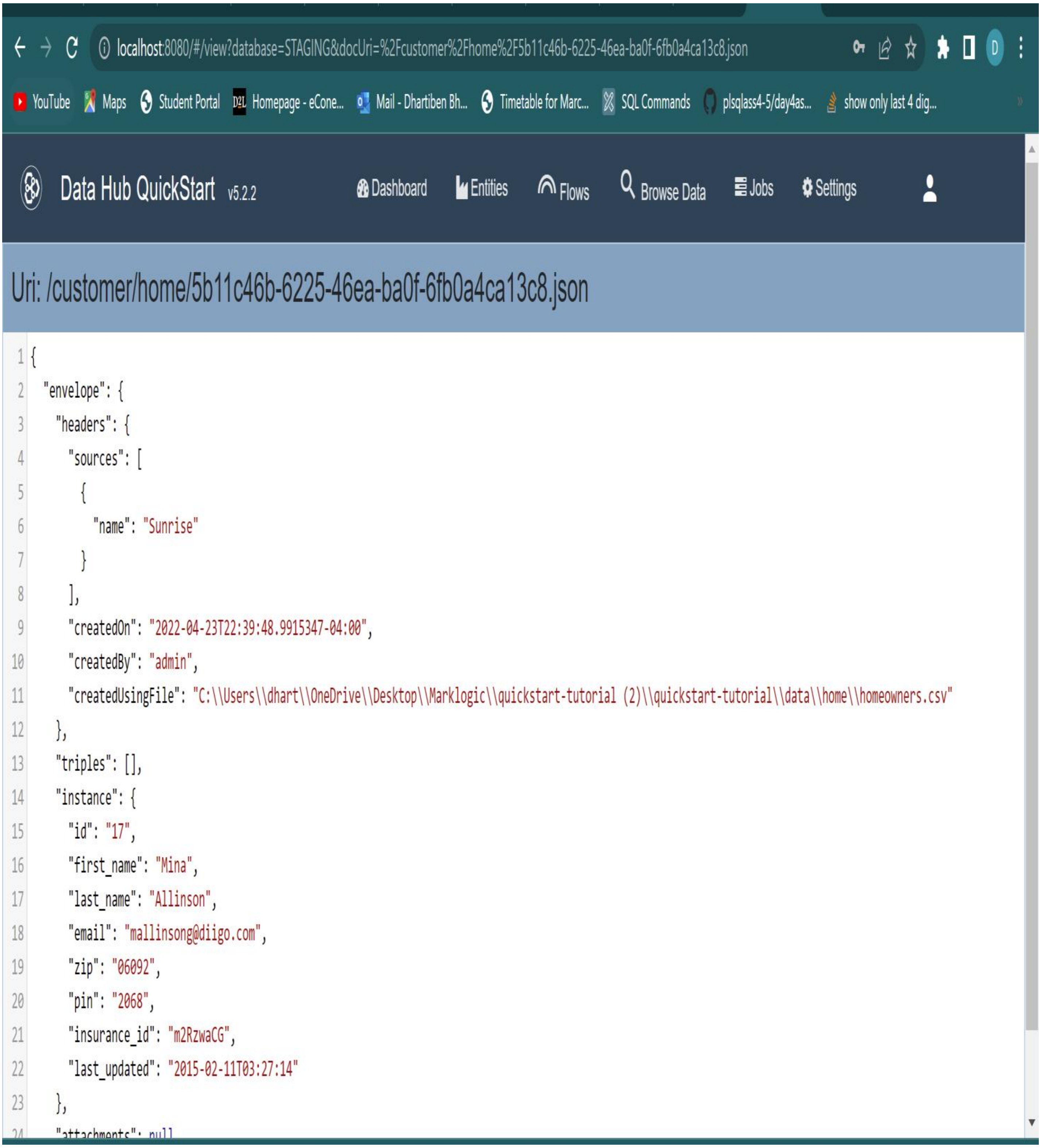
Field Separator

, ▼

Target Format

JSON ▼

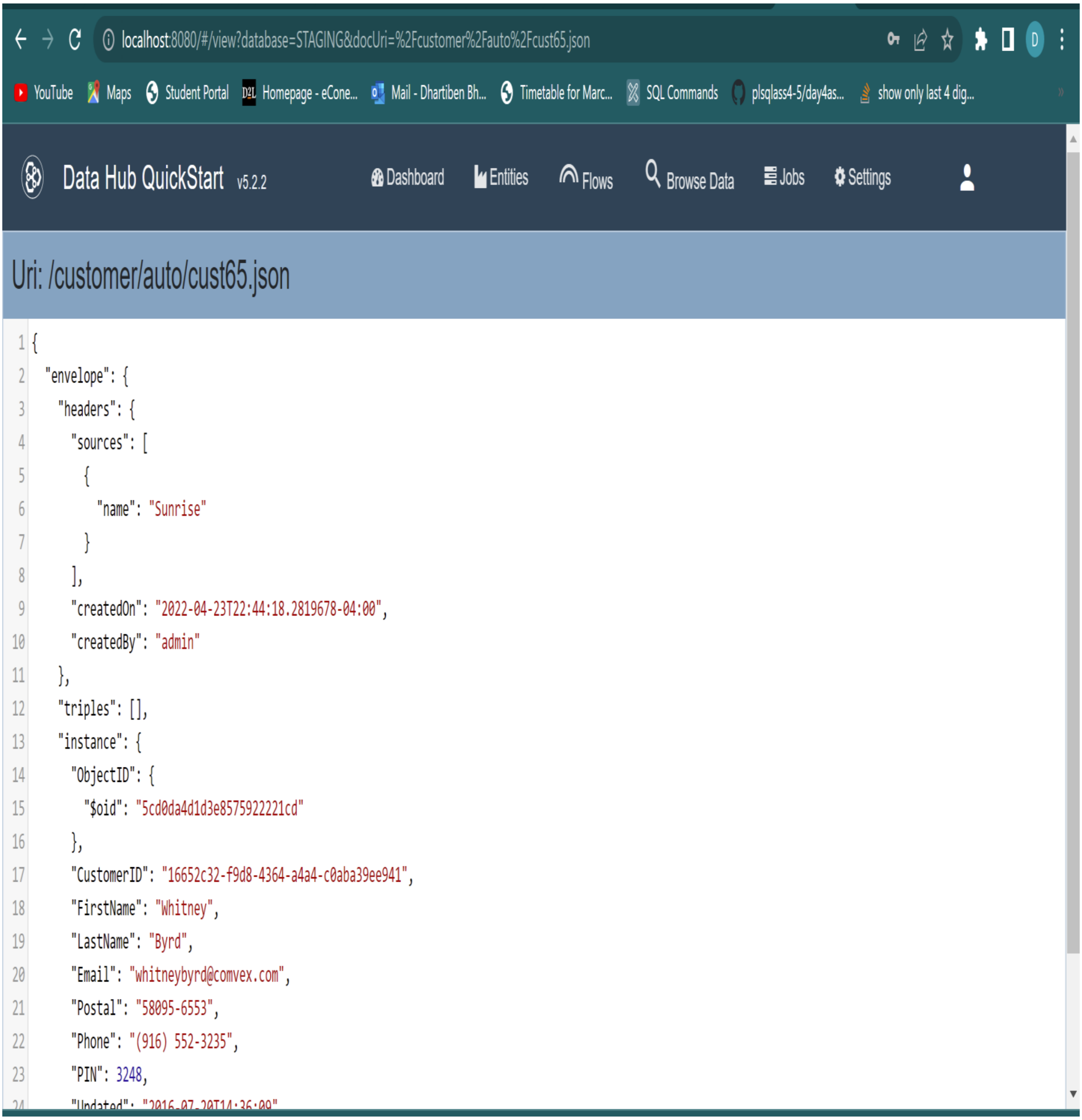
STEP 8: Running the Sunrise Flow



Uri: /customer/home/5b11c46b-6225-46ea-ba0f-6fb0a4ca13c8.json

```
{
  "envelope": {
    "headers": {
      "sources": [
        {
          "name": "Sunrise"
        }
      ],
      "createdOn": "2022-04-23T22:39:48.9915347-04:00",
      "createdBy": "admin",
      "createdUsingFile": "C:\\Users\\dhart\\OneDrive\\Desktop\\Marklogic\\quickstart-tutorial (2)\\quickstart-tutorial\\data\\home\\homeowners.csv"
    },
    "triples": [],
    "instance": {
      "id": "17",
      "first_name": "Mina",
      "last_name": "Allinson",
      "email": "mallinsong@diigo.com",
      "zip": "06092",
      "pin": "2068",
      "insurance_id": "m2RzwaCG",
      "last_updated": "2015-02-11T03:27:14"
    },
    "attachments": null
```

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



localhost:8080/#/view?database=STAGING&docUri=%2Fcustomer%2Fauto%2Fcust65.json

Uri: /customer/auto/cust65.json

```
{
  "envelope": {
    "headers": {
      "sources": [
        {
          "name": "Sunrise"
        }
      ],
      "createdOn": "2022-04-23T22:44:18.2819678-04:00",
      "createdBy": "admin"
    },
    "triples": [],
    "instance": {
      "ObjectID": {
        "$oid": "5cd0da4d1d3e8575922221cd"
      },
      "CustomerID": "16652c32-f9d8-4364-a4a4-c0aba39ee941",
      "FirstName": "Whitney",
      "LastName": "Byrd",
      "Email": "whitneybyrd@comvex.com",
      "Postal": "58095-6553",
      "Phone": "(916) 552-3235",
      "PIN": 3248,
      "Updated": "2016-07-20T14:36:00"
```

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