INFO 607-001

Dr. Il-Yeol Song

Final Project Documentation

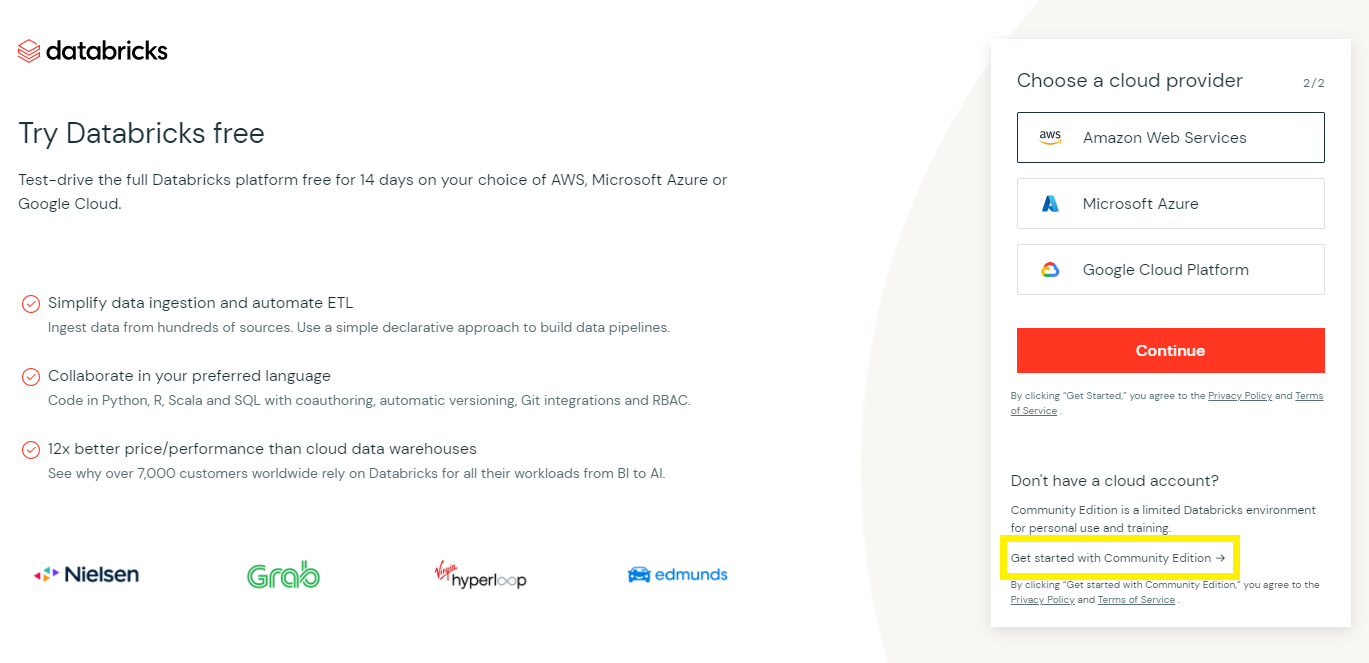
Spring 2023

**Data Lake and Data Lakehouse Technology**

**Using Databricks**

|  |  |  |
| --- | --- | --- |
| Patrick Artac | pa423 | PA |
| Allie Schneider | as5664 | AS |
| Francis Villamater | fv48 | FV |
| Zoe Weiner | zrw29 | ZW |

1. ***Download dataset from Kaggle.com***
   1. Go to <https://www.kaggle.com/datasets/iamprateek/store-transaction-data>
   2. Click the “Download” button in the top right hand corner
   3. Go to your Downloads folder on your computer and open the zip file
   4. Save the file titled “Hackathon\_Ideal\_Data” somewhere on your computer where you can access it later
2. ***Create a Databricks Community Edition Account***
   1. Go to databricks.com
   2. Click “Try Databricks” on the left hand menu
   3. Type your information into the box on the right titled “Create your Databricks account”
   4. Click “Continue”
   5. In the box on the right-hand side titled “Choose a cloud provider”, towards the bottom, click “Get started with Community Edition →” - highlighted in the screenshot below.



* 1. Complete the verification puzzle

1. ***Login to Databricks.com***
   1. Go to <https://community.cloud.databricks.com/login.html>
   2. Type in your username and password
   3. Click “Sign In”
2. ***Upload the Dataset***
   1. Hover your cursor over the icons on the left
   2. Click “Data” on the left hand menu
   3. Click the “Create Table” button in the top right hand corner of the Data pop up
   4. Click in the files box, where it says “Drop files to upload, or click to browse”
   5. A file finder pop up will come up - find where you saved the file titled “Hackathon\_Ideal\_Data”
   6. Click “Open”
   7. Once the file is uploaded, click the “Create Table in Notebook” button under the file upload
3. ***Working in the Notebook***
   1. Databricks will give you a template with instructions to help get you started. In the second cell of the notebook, “Cmd 2”, change line 7 from: first\_row\_is\_header = “false” to: first\_row\_is\_header = “true”
   2. Press CTRL + Enter on your keyboard to run the second cell
   3. **(Optional)** In the third cell of the notebook, “Cmd 3”, change line 3 from: temp\_table\_name = “Hackathon\_Ideal\_Data\_csv” to: temp\_table\_name = “Sales\_Data\_csv”
   4. Press CTRL + Enter on your keyboard to run the third cell
   5. **(Optional)** In the fourth cell of the notebook, “Cmd 4”, change line 5 from: select \* from `Hackathon\_Ideal\_Data\_csv` to: select \* from `Sales\_Data\_csv`
   6. Press CTRL + Enter on your keyboard to run the fourth cell
   7. **(Optional)** In the fifth cell of the notebook, “Cmd 5”, change line 5 from: permanent\_table\_name = “Hackathon\_Ideal\_Data\_csv” to: permanent\_table\_name = “Transactional\_Sales\_Data\_csv”
   8. Press CTRL + Enter on your keyboard to run the fifth cell
4. ***Query the Data Using SQL***
   1. Hover your mouse right underneath the last cell in the notebook until you a button with a “+” appears and click the “+” button to create a new cell
   2. In the first line of your new cell, type “%sql” (without the quotes) to tell the program you are writing code in SQL
   3. In line 2 of your new cell, create a comment by starting the line with “--” (without the quotes) to identify the purpose of the query you are about to write in this cell
   4. In line 3 of your new cell, write out your query in SQL
   5. Repeat steps 6.1 - 6.4 for each query you would like to run