**Student Guide for CS4225/CS5425 Assignment 2**

Table of Contents

[1 Overview 2](#_Toc129007657)

[1.1 Task 1: Spark SQL (15 Marks) 2](#_Toc129007658)

[Dataset 2](#_Toc129007659)

[Questions 3](#_Toc129007660)

[1.2 Task 2: Spark ML (10 Marks) 4](#_Toc129007661)

[Datasets 4](#_Toc129007662)

[Question 4](#_Toc129007663)

[1.3 Getting Started 5](#_Toc129007664)

[2 Using Databricks 6](#_Toc129007665)

[2.1 Databricks Community Edition Registration 6](#_Toc129007666)

[2.2 Upload Dataset 7](#_Toc129007667)

[3 Google Colab (Backup) 9](#_Toc129007668)

[3.1 Getting started in Google Colab 9](#_Toc129007669)

[4 Testing and Submitting 10](#_Toc129007670)

# 1 Overview

## 1.1 Task 1: Spark SQL (15 Marks)

Answer the following questions using Spark, the following sections explains the task and helps you to get started working on it.

### Dataset

It consists of three tables coming from the database of a shop, with products, sales, and sellers.

Table

Description automatically generated with medium confidence

**Sales Table**

Each row in this table is an order and every order can contain only one product. Each row stores the following fields:

* order\_id: The order ID
* product\_id: The single product sold in the order. **All orders have exactly one product**
* seller\_id: The selling employee ID that sold the product
* num\_of\_items\_sold: The number of units sold for the specific product in the order

**Products Table**

Each row represents a distinct product. The fields are:

* product\_id: The product ID
* product\_name: The product name
* price: The product price

**Sellers Table**

This table contains the list of all the sellers:

* seller\_id: The seller ID
* seller\_name: The seller name
* rating: The seller’s rating (scale of 1 – 5)

### Questions

1. Output the top 3 most popular products sold among all sellers
   1. Your table should have 1 column(s): [product\_name]
2. Find out the total sales of the products sold by sellers with seller\_id 1 to 10 and output the top most sold product
   1. Your table should have 1 column(s): [product\_name]
3. Compute the combined revenue earned from sellers where seller\_id ranges from 1 to 500 inclusive.
   1. Your table should have 1 column(s): [total\_revenue]
4. Among sellers with rating >= 4 who have achieved a combined number of products sold >= 3000, find out the top 10 most expensive product(product\_name) sold by any of the sellers. (If there are multiple products at the same price, please sort them in ascending order of product\_id)
   1. Your table should have 1 column(s): [product\_name]
   2. To get the full mark, your query should not run for more than 1 min

## 1.2 Task 2: Spark ML (10 Marks)

### Datasets

We will be using the bank marketing campaign data. The datasets give you information about a marketing campaign done by a financial institution. Detailed information (i.e. column description) is provided below.

Graphical user interface, text, application, letter

Description automatically generated

We have already split data into training set (bank\_train.csv) and testing set (bank\_test.csv)

### Question

Build ML model to predict whether the customer will subscribe bank deposit service or not. Train the model using training set and evaluate the model performance (e.g. accuracy) using testing set.

* You can explore different methods to pre-process the data and select proper features
* You can utilize different machine learning models and tune model hyperparameters
* Present the final testing accuracy.

Focus is on building a proper ml pipeline so that you can adjust the different parts of pipeline to achieve the best performance.

Graphical user interface

Description automatically generated

A baseline model using Random Forest Classifier achieved a test accuracy of around 78.5%. You should at least create a Random Forest model and achieve the similar or higher test accuracy. On top of that, you can also utilize other machine learning models and achieve similar or better test accuracy.

## 1.3 Getting Started

For this assignment, we will be using Databricks (easy setup and allows editing code through notebook interface, similar to Google Colab), a web-based platform for working with Spark. You can also easily track the runtime for each cell and view Spark UI for each job. Google Colab will be used as a backup in case of Databricks service interruption.

The assignment is designed to be solved using Spark. You can use Spark in a language of your choice (Python, Scala, R, etc.) but you have to use Spark to complete all the tasks. It is NOT allowed to use other packages (e.g. pandas, sklearn, pytorch, tensorflow and etc.) in this assignment.

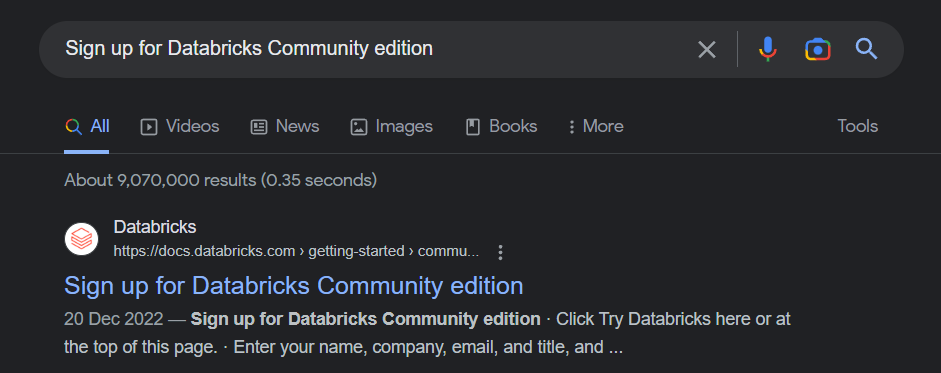
This assignment does not require using the school cluster but if you want to experience to run Spark codes on a cluster environment (with multiple nodes) feel free to try that using school cluster.

# 2 Using Databricks

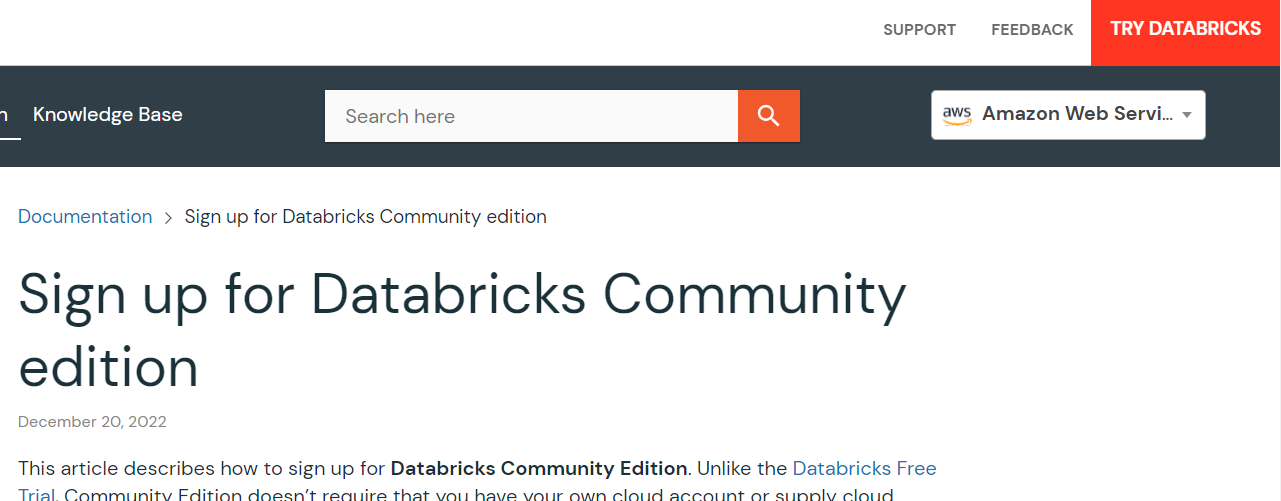
## 2.1 Databricks Community Edition Registration

To get started, we would be using the free version, databricks community edition and you would have to register for it.

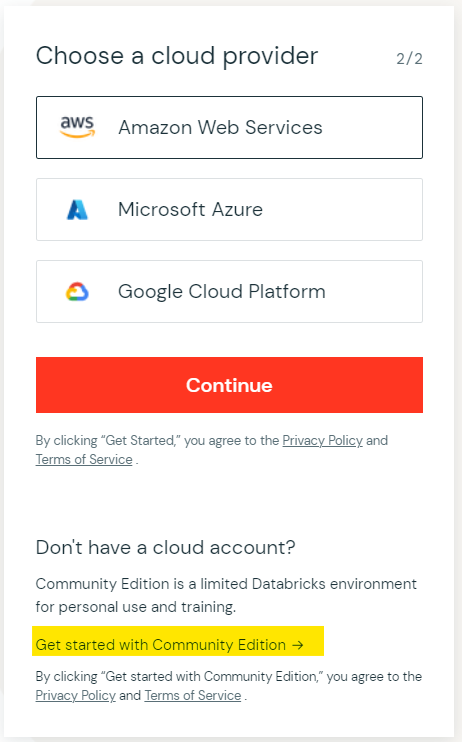
1. Google “Sign up for Databricks Community edition” and click on the first link or click <https://docs.databricks.com/getting-started/community-edition.html>



1. click “TRY DATABRICKS” on the top right corner of the page



1. Fill in the necessary details and click continue to create an account
2. You will reach the page below and since we are using the Community Edition, please click on the portion highlighted below

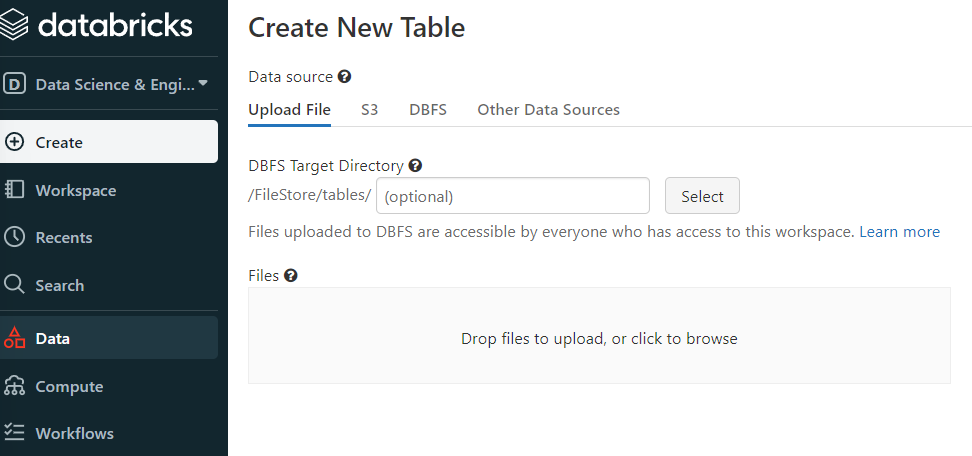


1. Validate your email address and login to databricks

## 2.2 Upload Dataset

Once you are logged in, you can start to upload the dataset we have given you, consisting of the 3 files. We have also given you a notebook file to read the tables so that you can get started with the questions.

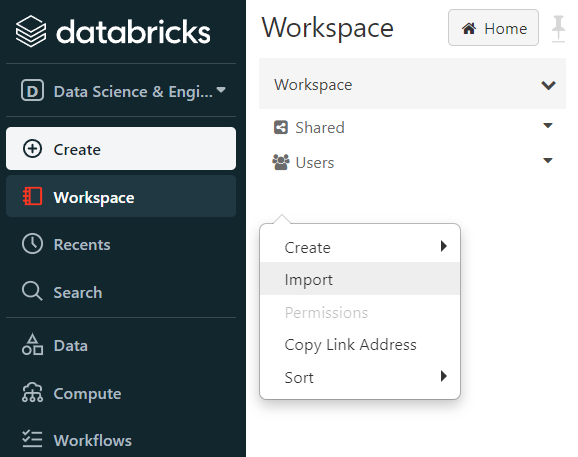
Click on Data > Create Table upload the data files into the default target directory in the picture below.



Once uploaded, the files should be found under DBFS > FileStore > tables

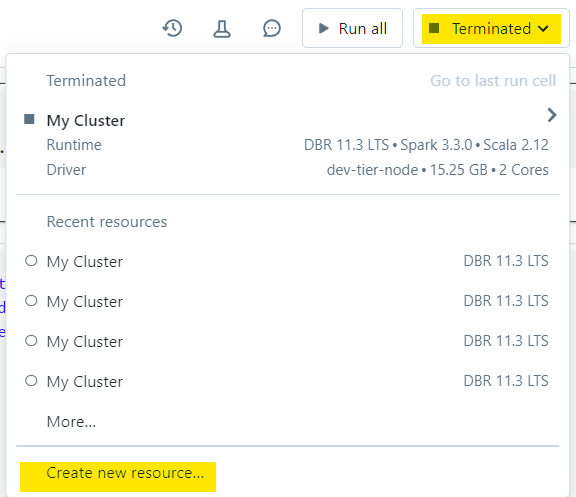
Our given .ipynb file would be using the default file locations and original file names, hence it should be able to read all the tables into the data frames so that you can make use of the tables to attempt the questions.

To upload the given .ipynb file, go to Workspace > right-click > import

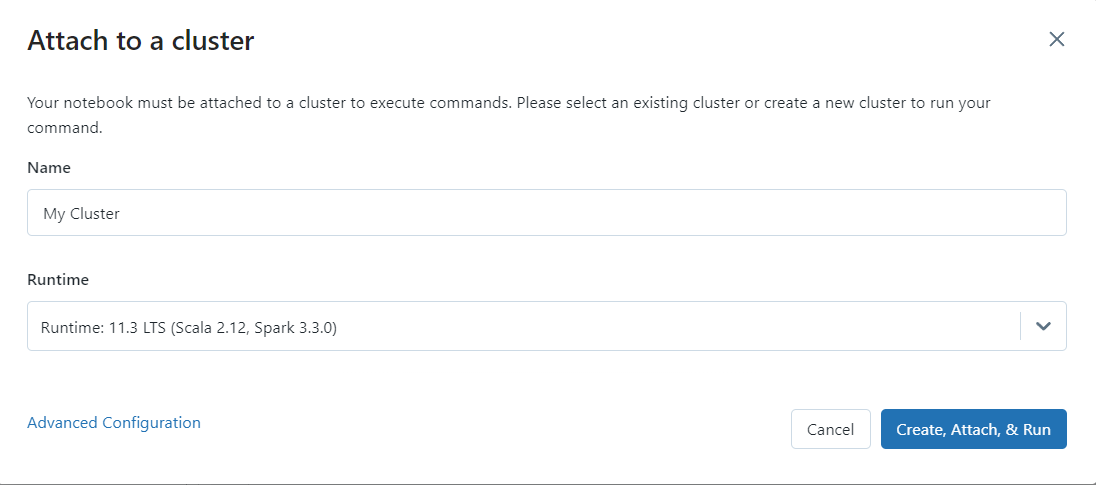


Once the notebook file is opened, you would need to create and attach a cluster to run your code. Note that every cluster you create will automatically terminate after an idle period of two hours for the Community Edition. Hence you would need to re-create a new cluster to run your code if your cluster has terminated.

1. On the top-right of the notebook if it is not connected, click on it to create new resource



1. Click on Create, Attach, & Run and you would be ready to run your code



You are allowed to choose **any** runtime version for your cluster.

A few things to take note when answering the questions:

1. Every question must be answered in one cell
2. We should be able to see the generated answer table after running that cell
3. Please follow the table name conventions given together with the questions, although we will be marking based on your answer and explanation.

# 3 Google Colab (Backup)

To get started with Google Colab, visit the Colab notebooks:

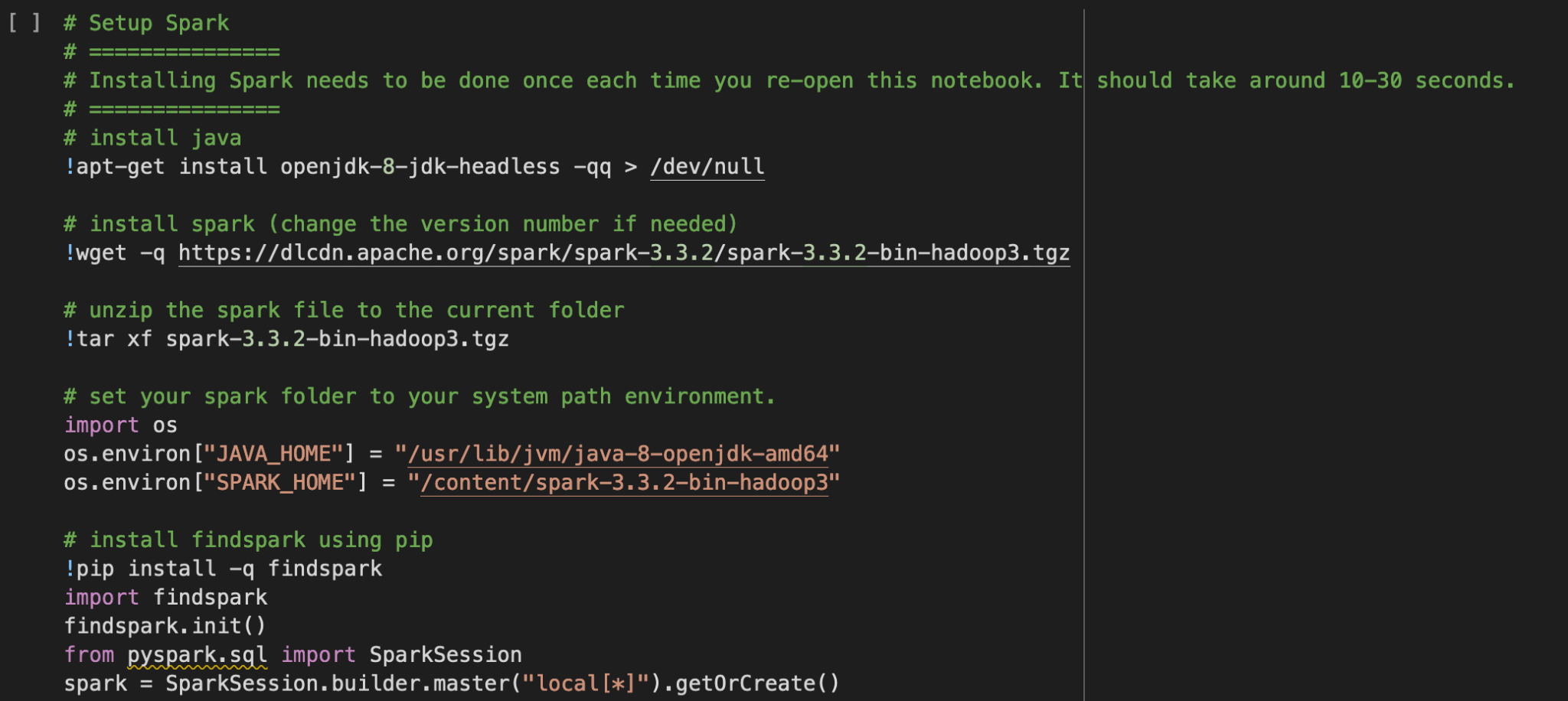
* <https://drive.google.com/file/d/1JCk9gXtAFV5q7PPfX8QH8bIX87hBL9bB/view?usp=sharing>

To copy the notebooks so you can edit them, go to File > Save a copy in Drive. The data files will be downloaded in the given notebook for reading into dataframes, you do not need to upload the dataset.

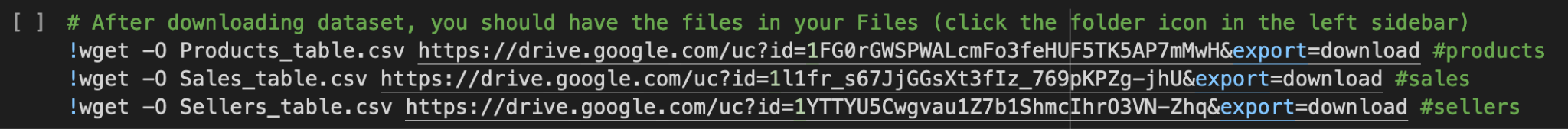
## 3.1 Getting started in Google Colab

For each part, the first 3 cells should not be modified. The cells will do the following:

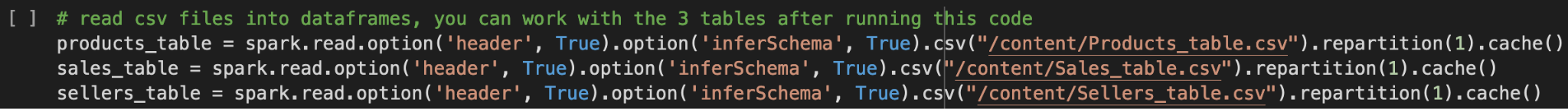
1. Set up Spark in the local environment



1. Retrieve dataset from cloud



1. Read dataset



You are to run the first three cells **before** starting on either parts.

After running your Spark code in Google Colab, ideally you should have figured out the answers for the questions and filled it in the Colab notebook. Similar to Databricks, you would need to submit your notebook file.

Click on File > Download > Download .ipynb

# 4 Testing and Submitting

At this point, you should have a completed Databricks notebook with the answers. Running your notebook should also show how your code can generate the results.

Once you are successful, please submit on Canvas. There are 2 things to submit:

1. Download your completed Databricks notebook by clicking File > Export > IPython Notebook. We will run your .ipynb file to ensure that your code generates your given answers. We should be able to see the generated tables at the respective question cells after running the notebook file.
2. Record a video not longer than 5 minutes to explain your code, onto Canvas. Make sure the video captures both your face and your code throughout the full duration.

Zip the two files together and submit the zipped file to Canvas by Apr 3, 2023, Mon 11:59pm.

**Links and References**

For a basic guide + API reference for Spark, see <https://spark.apache.org/docs/latest/sql-getting-started.html>, or <https://spark.apache.org/docs/latest/api/python/getting_started/index.html>

for PySpark. For a PySpark ‘cheatsheet’, see <https://s3.amazonaws.com/assets.datacamp.com/blog_assets/PySpark_SQL_Cheat_Sheet_Python.pdf>.