**Hướng dẫn thực hành Spark trên Databricks**

1. Tạo tài khoản

Vào link databricks.com.

Chọn Login.

A screenshot of a computer

Description automatically generated with medium confidence

Chọn Sign up để tạo tài khoản.

A screenshot of a login form

Description automatically generated with medium confidence

Nhập các thông tin cá nhân vào đây.

A screenshot of a computer

Description automatically generated with medium confidence

Chọn vào mục nhỏ xíu này để tạo tài khoản trên Community Cloud Databricks.

A screenshot of a computer

Description automatically generated with low confidence

Sau đó sẽ nhận email về việc xác nhận tạo tài khoản.

Ấn vào đường link trong mail để đổi password là xong!

Giao diện của Community Databricks:

A screenshot of a computer

Description automatically generated

* **Create**: Thao tác nhanh để tạo Notebook, Table và Cluster
* **Workspace**: Nơi quản lý thư mục và các file Notebook
* **Recents**: Link đến những Notebook vừa truy cập
* **Search**: Tìm kiếm file, notebook
* **Data**: Nơi lưu trữ các file dữ liệu, các bảng do người dùng tạo ra
* **Compute**: Quản lý các cụm tính toán do người dùng tạo ra
* **Workflows**: Quản lý Data Lineage, nạp tiền để dùng

1. Thao tác với Databricks
   1. Khởi tạo cụm tính toán

Chọn Compute, xuất hiện màn hình quản lý các cụm tính toán của người dùng.

A screenshot of a computer

Description automatically generated with low confidence

Chọn Create Compute, nhập tên cụm. Có thể chọn các phiên bản spark và scala khác nhau cho môi trường cụm.

A screenshot of a computer

Description automatically generated with medium confidence

Nhấn chọn Create Cluster.

Về lại màn hình Compute sẽ thấy cụm đang chạy khởi tạo. Đợi một lúc đến khi cụm khởi tạo thành công.

A picture containing screenshot, text

Description automatically generated

Có thể ấn vào cụm để xem các thông tin của cụm đó: Cấu hình, Notebook nào sử dụng, Thư viện, Spark UI,…

A screenshot of a computer

Description automatically generated with medium confidence

* 1. Tải dữ liệu

Vào mục Data, chọn Create Table.

A screenshot of a computer

Description automatically generated with medium confidence

Xuất hiện màn hình quản lý các file mình đẩy lên Databricks.

A screenshot of a computer

Description automatically generated with medium confidence

Các Data Source:

* **DBFS**: Databricks File System – nơi mình sẽ upload các file dữ liệu phục vụ cho việc chạy notebook của mình. Đường dẫn root của nó là /FileStore/tables/
* **S3**: Dữ liệu trên S3 trỏ vào
* **Other Data Sources**: Các data sources khác mình có thể cấu hình để lấy

Gõ đường dẫn để upload file lên thư mục mình muốn. Ấn Files để tải các file lên.

A screenshot of a computer

Description automatically generated with medium confidence

* 1. Tạo Notebook

Vào Workspace để quản lý các Notebook.

A screenshot of a computer

Description automatically generated

Ấn phải chuột để tương tác:

* Create: Tạo Folder, Notebook, …
* Import: Import Notebook từ file

A screenshot of a computer

Description automatically generated

Chọn tạo Notebook, sẽ xuất hiện Notebook mới trong folder mình tương tác.

A picture containing text, line, software, screenshot

Description automatically generated

* 1: Tên Notebook
* 2: Ngôn ngữ và môi trường để chạy Notebook
* 3: Cụm tính toán dùng để xử lý Notebook
* 4: Khu vực làm việc
  1. Import Notebook

Ấn phải chuột vào thư mục cần Import, chọn Import.

Upload file có đuôi .dbc lên