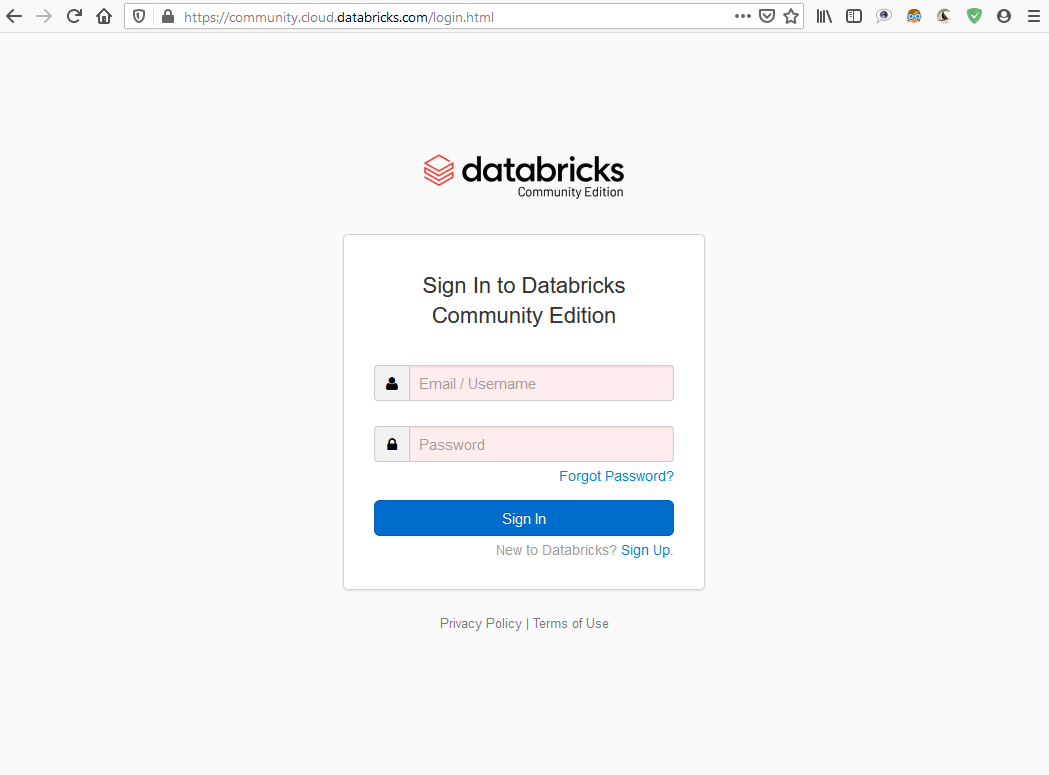
# **Databricks Community**

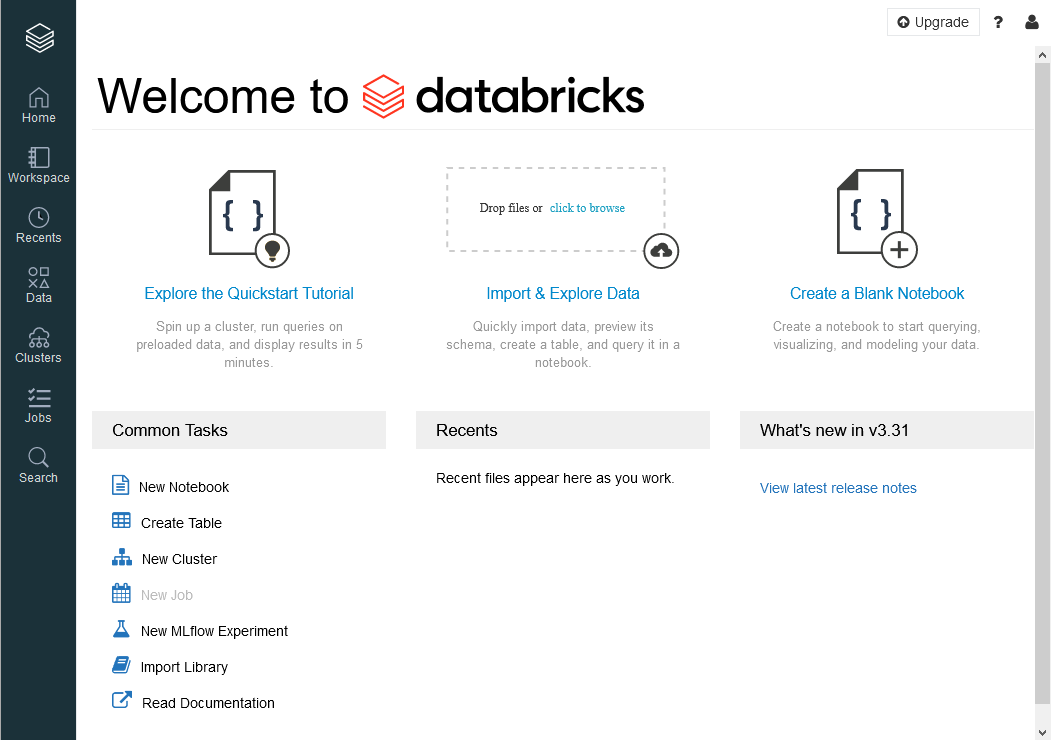
Databricks Community Edition (DCE) cung cấp miễn phí một mini-cluster (cloud-base) dễ dàng và nhanh chóng để tạo, quản lý môi trường phát triển với Apache Spark. Rất phù hợp với developer, data scientists, data engineer… muốn có môi trường khi bắt đầu với Apache Spark.

Link reference: <https://community.cloud.databricks.com>

Link đăng kí (free): <https://databricks.com/try-databricks>

  
*Hình 1. Databricks sign-in page*

Sau khi đăng nhập thành công, màn hình chính của DCE sẽ như dưới đây:

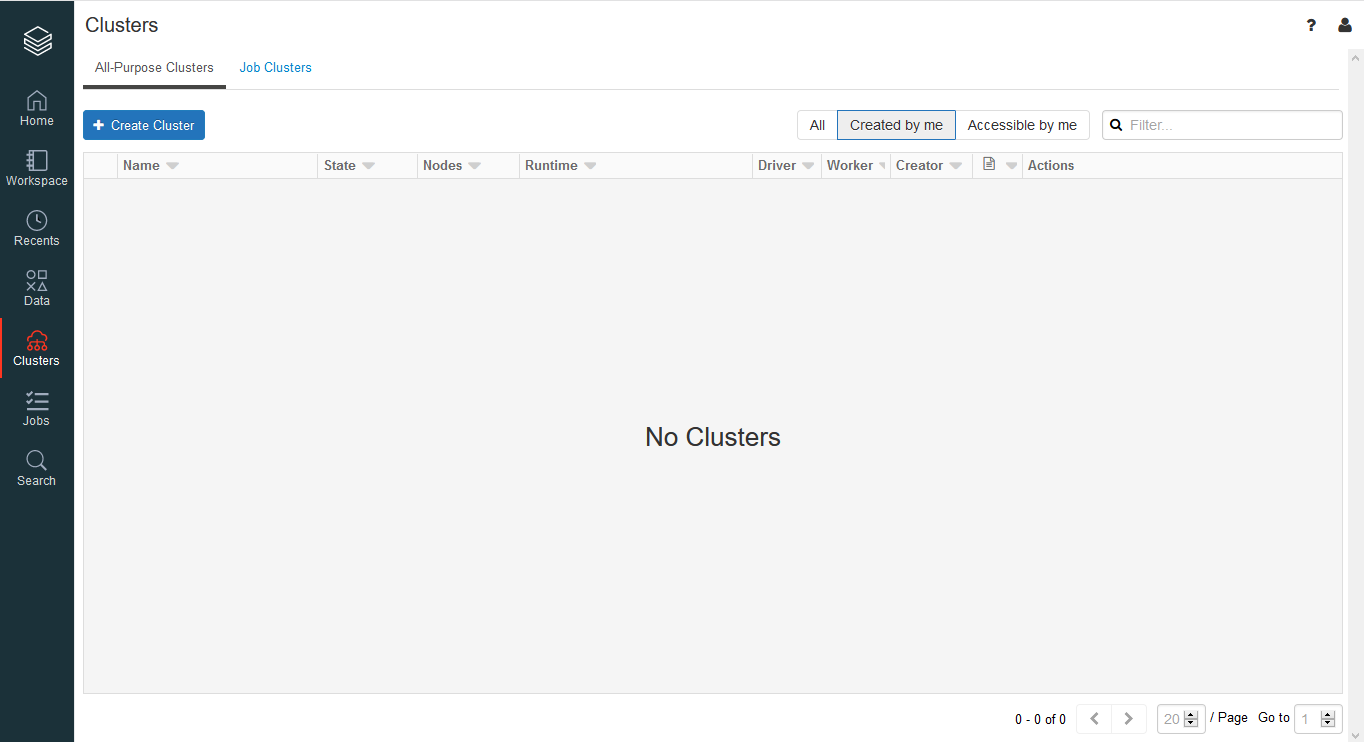
  
*Hình 2. DCE welcome page*

Document này sẽ hướng dẫn 1 số thao tác cơ bản để làm việc với DCE:

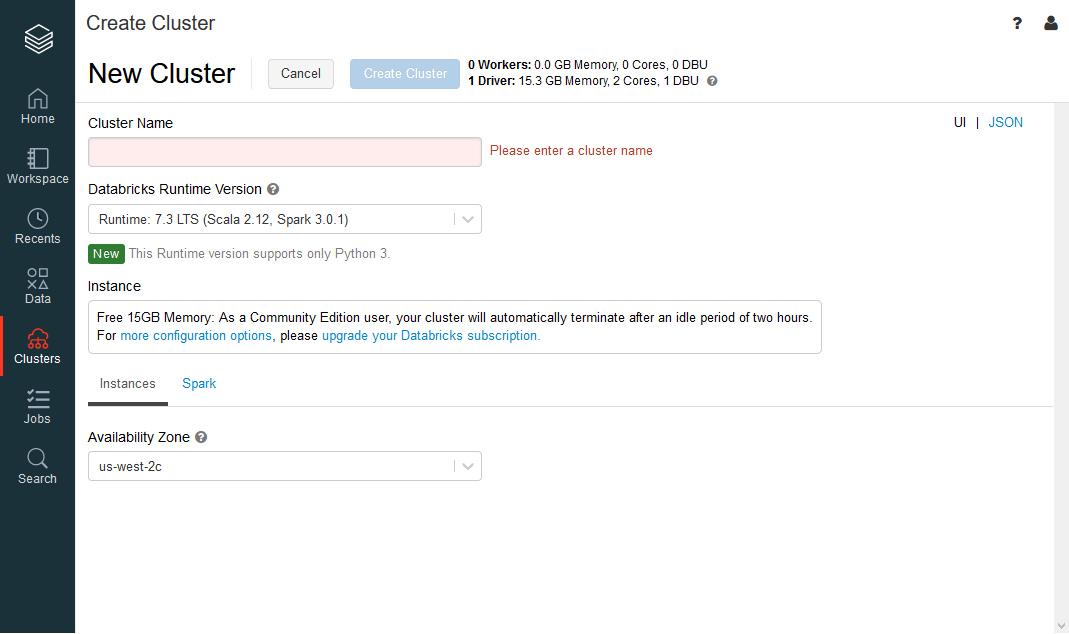
1. Tạo một Cluster mới
2. Tạo mới Folder
3. Tạo mới Notebook

## **Tạo một cluster mới**

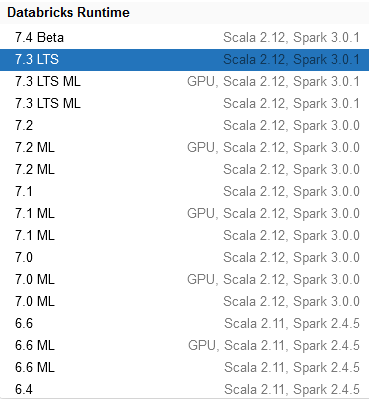
Để tạo một cụm, click vào biểu tượng Cluster trong thanh điều hướng dọc ở bên trái của trang.

  
*Hình 3. DCE Cluster page khi chưa có Cluster nào active*

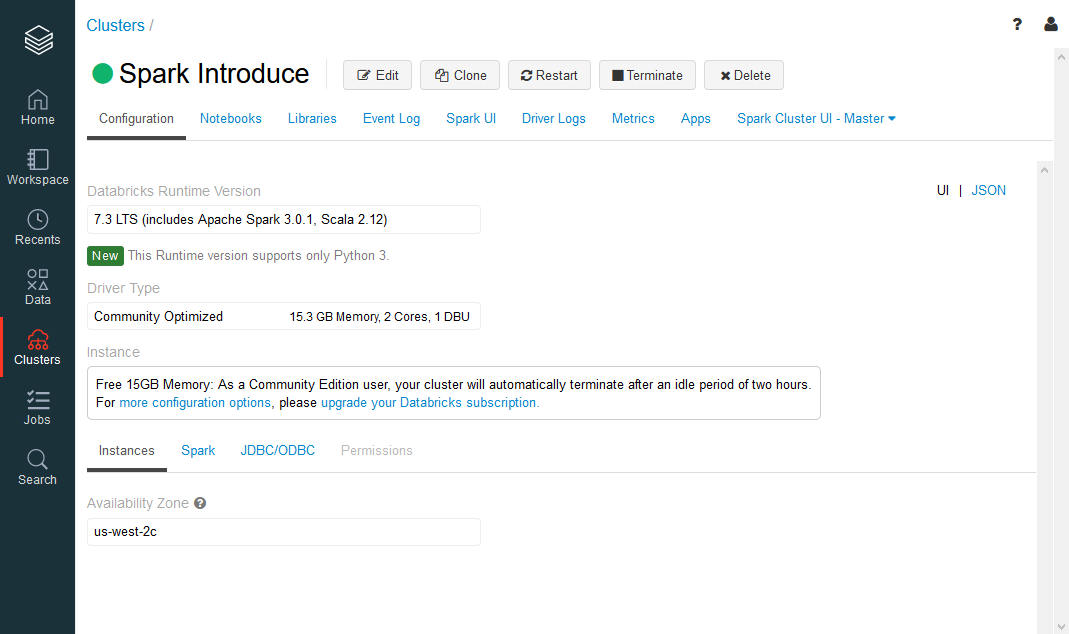
Tiếp theo, click Create Cluster

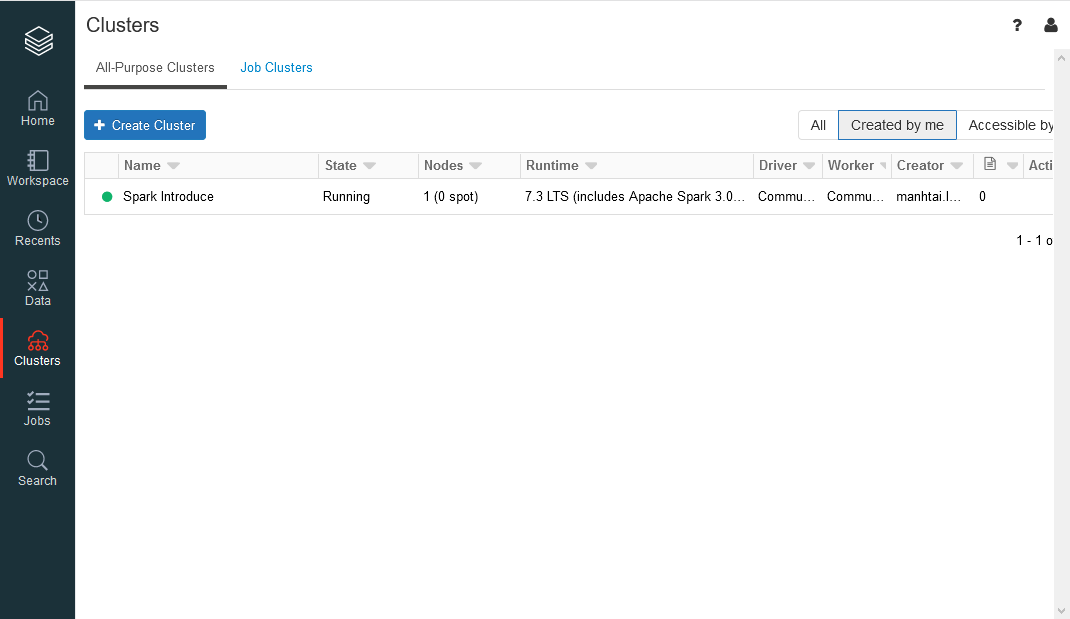
  
*Hình 4. Form tạo mới Cluster*

\*Ở phần Databricks Runtime Version, DCE có support nhiều phiên bản Spark khác nhau, bên cạnh đó có hỗ trợ cả GPU.

  
*Hình 5. DCE environment*

Sau khi hoàn thành Form để tạo Cluster, tùy thuộc vào lựa chọn sẽ mất khoảng vài phút để tạo 1 single-node Spark Cluster.

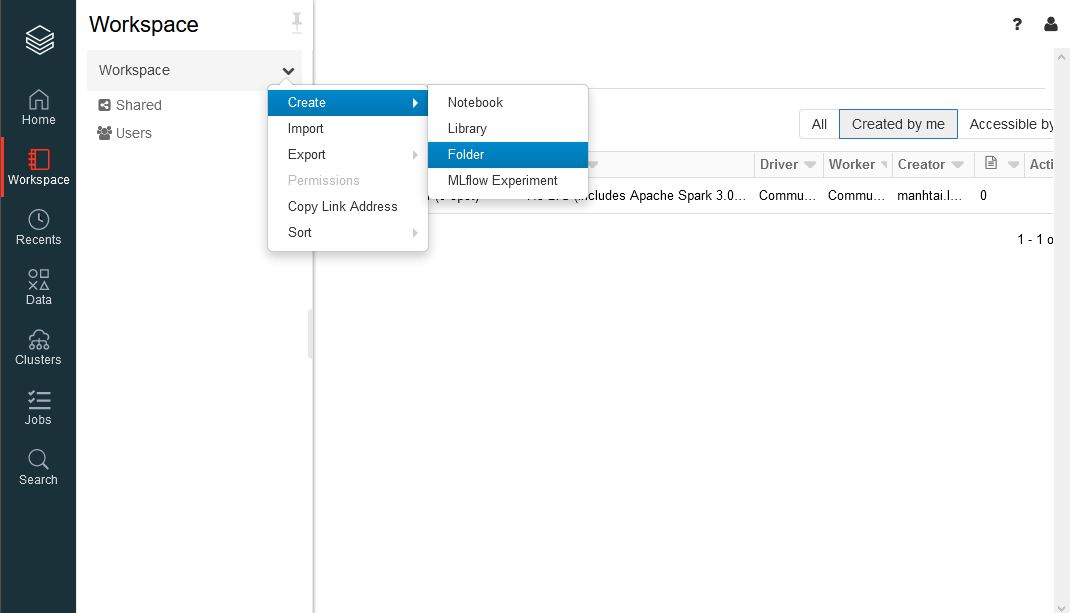
  
*Hình 6. Cluster sau khi active*

  
*Hình 7. Cluster page khi có một cluster active*

## **Tạo mới Folder**

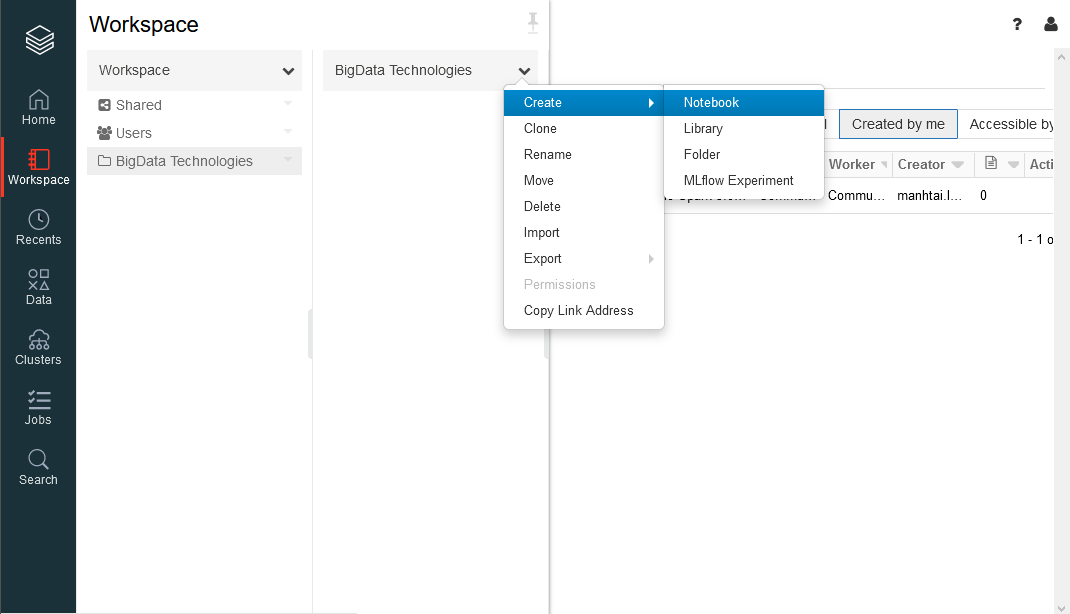
DCE tổ chức workspace theo cách coi nó như là thư mục root trên máy tính, có nghĩa là có thể đặt các file ở đó hoặc tạo các folder để giúp quản lý các tệp theo một cách cụ thể.

Để tạo mới folder, click Workspace icon ở tab điều hướng bên trái, sau đó click mũi tên **” ˅ ”** (hoặc click chuột phải ở vùng Workspace), chọn Create => Folder.

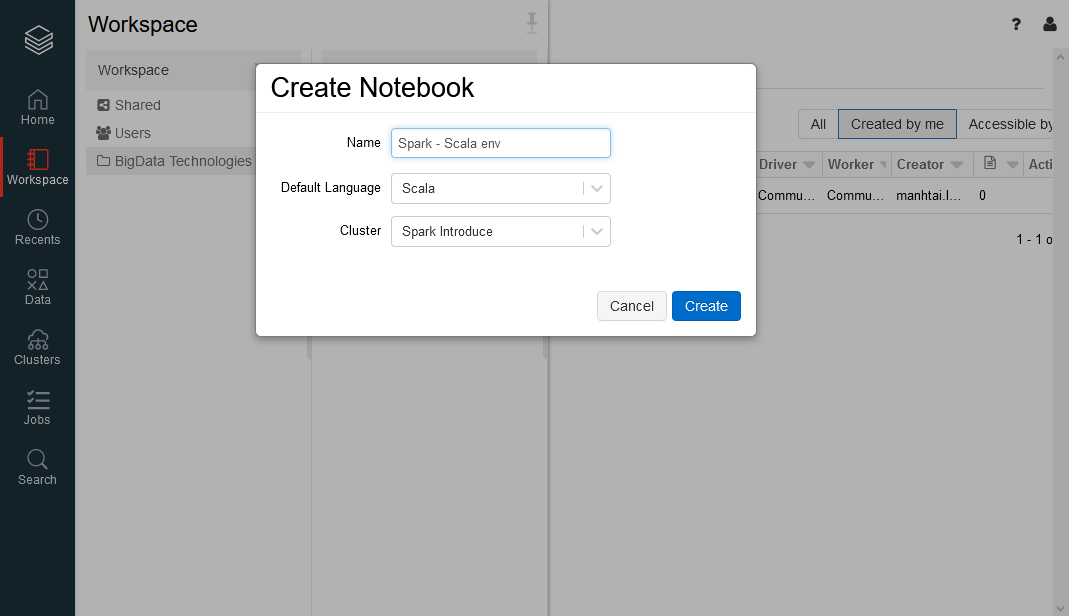
  
*Hình 8. Menu tạo mới folder*

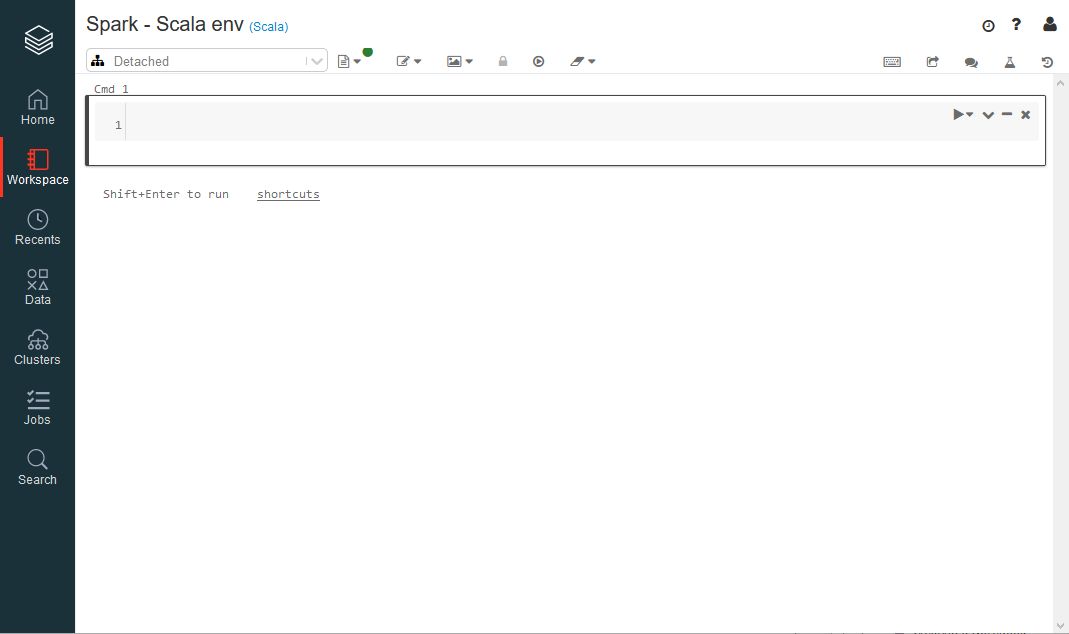
## **Tạo mới Notebook**

Sau khi đã tạo được Folder, có thể tạo Notebook bằng cách click phải vào Folder, chọn Create => Notebook

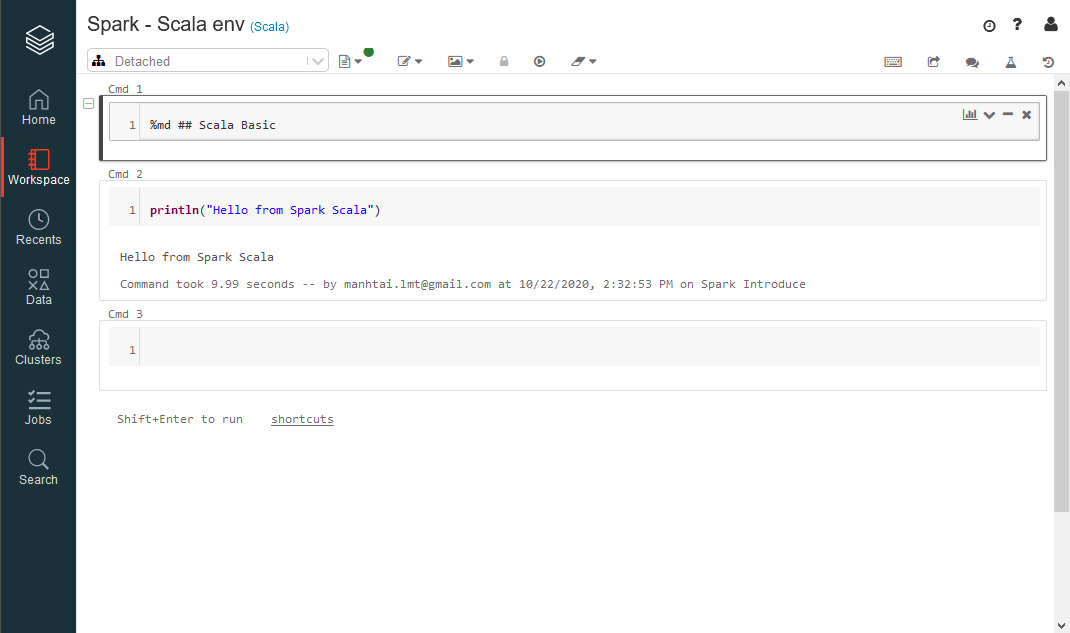
  
*Hình 9. Menu tạo mới Notebook*

Ở form tạo notebook, có thể chọn ngôn ngữ muốn dùng trong notebook, và cluster mà notebook hoạt động

  
*Hình 10. Form tạo notebook*

  
*Hình 11. Scala notebook*

Notebook về cơ bản là một môi trường tính toán tương tác (tương tự như Spark shell nhưng tốt hơn), trong đó có thể thực thi Spark code, viết code bằng nhiều ngôn ngữ như Scala, Markdown, HTML, … và trực quan hóa kết quả phân tích dữ liệu với nhiều loại biểu đồ khác nhau.

  
*Hình 12. Một số ví dụ khi tương tác với notebook shell*

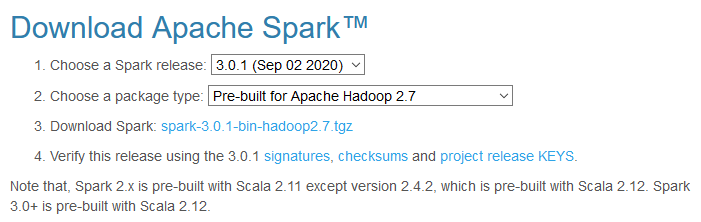
Tất cả các paragraph shell trong 1 notebook thực thi trên cùng 1 environment, mỗi shell có thể dễ dàng update bằng việc chạy lại shell và các shell có thể sử dụng biến/hàm đã định nghĩa ở các shell khác mà không cần viết lại, shortcut để chạy 1 shell là “Shift + Enter”

  
*Hình 13. Một ví dụ đơn giản khác về việc tương tác giữa các shell*

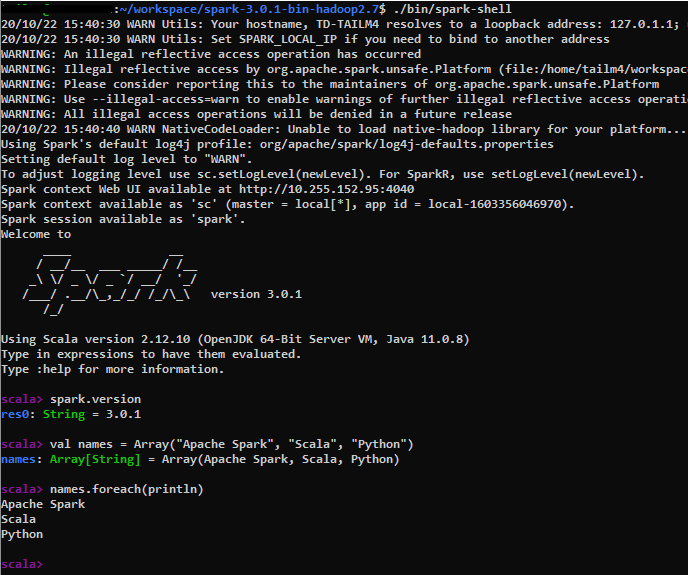
# **Spark-shell**

Một cách khác có thể sử dụng để tương tác khi bắt đầu với Apache Spark đó là Spark-shell, tuy nhiên cách này không được khuyến khích bằng việc sử dụng Notebook bởi tính linh hoạt và trực quan của Notebook vượt trội hơn nhiều

Để sử dụng Spark-shell, đầu tiên sẽ bắt với việc Download phiên bản Spark phù hợp từ trang chủ: <https://spark.apache.org/downloads.html>

  
*Hình 14. Apache Spark download option*

Sau khi tải được package Spark mong muốn, giải nén sẽ có được tập thực thi để khởi chạy Spark-shell ở: bin/spark-shell

  
*Hình 15. Spark-shell*