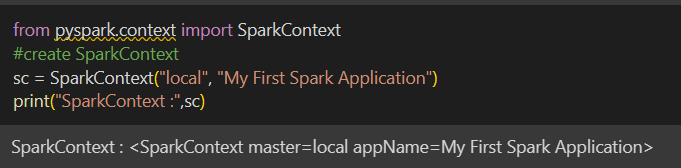
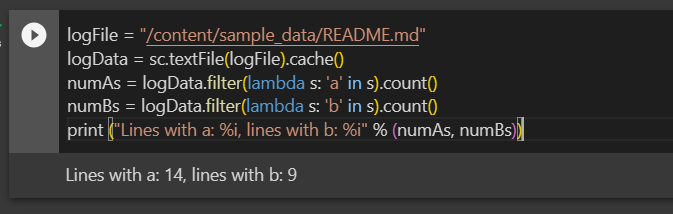
**Bài thực hành số 01**

**Bài 1: Cài đặt Apache Spark trên Colab**

- Bước 3: Chạy chương trình





**Bài 2: Cài đặt Apache Spark trên Windows 10**

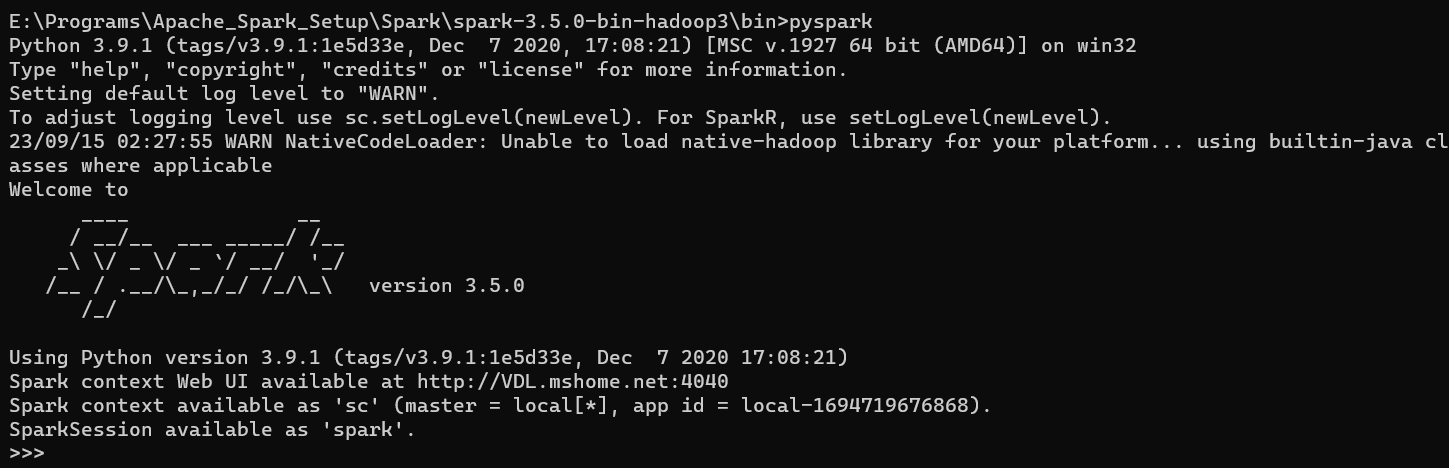
- JDK: 17.0.8

- Hadoop: 3.0.0

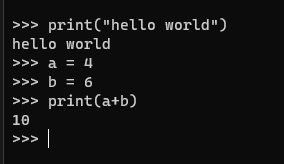
- Python: 3.9.1

- Spark: 3.5.0

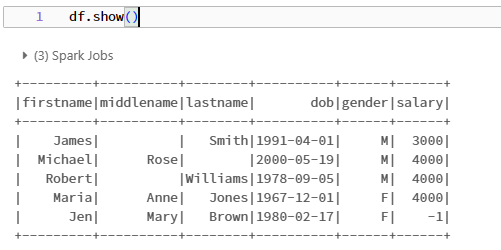
* Bước 7: Chạy Spark dùng Python chế độ tương tác



- Gõ lệnh tương tác (làm theo ví dụ em bị lỗi thư viện của python nên em demo ví dụ khác):



* Bước 9: Lập trình bằng Jupyter Notebook (Databricks Community)



**Bài 10: Lưu ra tệp**

* Đọc dữ liệu từ file readme.md và ghi những dòng chứa chữ ‘Spark’ ra thư mục Bai10.

file\_path = "/FileStore/tables/README.md"

readme\_data = spark.read.text(file\_path)

spark\_lines = readme\_data.filter(readme\_data.value.contains("Spark"))

output = "/dbfs/FileStore/tables/Bai10"

spark\_lines.write.text(output)

* Ghi dữ liệu từ df ra tệp ra tệp csv.

df.write.csv("/dbfs/FileStore/tables/output.csv", header=True)