**Using the official docker image**

Make sure that [docker](https://www.docker.com/community-edition) is installed in your local machine.

Use this command to launch Apache Zeppelin in a container.

docker run -p 8080:8080 --rm --name zeppelin apache/zeppelin:0.9.0

To persist logs and notebook directories, use the [volume](https://docs.docker.com/engine/reference/commandline/run/#mount-volume--v-read-only) option for docker container.

docker run -p 8080:8080 --rm -v $PWD/data:/data -v $PWD/notebook:/notebook -e ZEPPELIN\_DATA\_DIR='/logs' -e ZEPPELIN\_NOTEBOOK\_DIR='/notebook' --name zeppelin apache/zeppelin:0.9.0

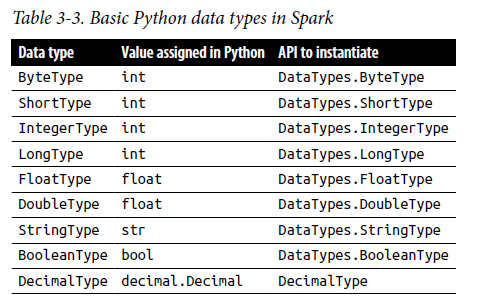
If you have trouble accessing localhost:8080 in the browser, Please clear browser cache.

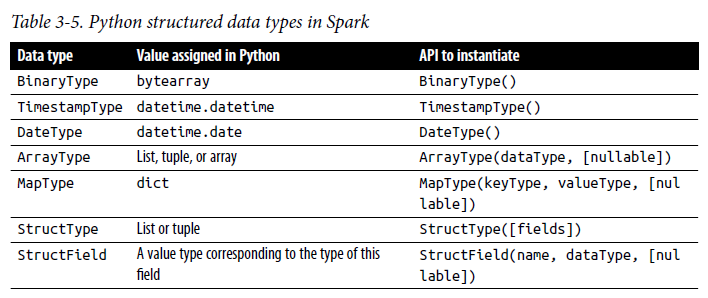
Thực hiện một số truy vấn đơn giản sau:

|  |  |
| --- | --- |
| Liệt kê tất cả các báo từ trang Dân Trí | %spark.pyspark  # show all article from DanTri  DanTri\_df = (data\_df.select("\_id", "headline", "organization")  .where(data\_df.organization=="DanTri"))  print("Count: ", DanTri\_df.count())  DanTri\_df.show() |
| Chuyển đổi tất cả lượt like, share,.. từ -1 sang 0  => Sử dụng cho bước tiền xử lý dữ liệu. |  |
| Thực hiện biểu diễn số bài viết theo ngày |  |
| Thể hiện số bài viết theo nguồn trang | %spark.pyspark  source\_df = (data\_df.groupBy("organization")  .agg(count("organization").alias("Total"))  .orderBy("Total", ascending=False))  z.show(source\_df) |
| - Đếm xem có bao nhiêu bài báo nói về BitCoin (Trong Title có chữ BitCoin)  - Hiển thị Pie Chart xem báo nào đăng nhiều nhất => Báo đó nói về tài chính nhiều.  - Sắp xếp chúng theo biểu đồ thời gian => Tìm xem có liên hệ gì với tình hình thế giới hoặc giá bitcon ko? |  |

Note:

|  |  |
| --- | --- |
| Sử dụng một số hàm trong truy vấn sql (count, sum, avg,…) | from pyspark.sql.functions import \* |
|  | **import org.apache.spark.sql.SparkSession** |
| Using with PySpark types | **import org.apache.spark.sql.types.\_** |
| sds |  |





File path when working in above work space: