databricks

14 DAYS

AI CHALLENGE

DAY 01

Topic:

Platform Setup & First Steps

Challenge:

1. Create Databricks Community Edition account

2. Navigate Workspace, Compute, Data Explorer

3. Create first notebook

4. Run basic PySpark commands

#DatabricksWithIDC

```python
%pip install kaggle
```

Configure Kaggle API

```python
import os

os.environ["KAGGLE_USERNAME"] = "souvanikchaudhury"
os.environ["KAGGLE_KEY"] = "b27df572e24f5ef21668f11a564bf1a4"

print("Kaggle credentials configured!")
```

```
Kaggle credentials configured!
```

Create Schema

```
04:09 PM (3s)                                              5

spark.sql("""
CREATE SCHEMA IF NOT EXISTS workspace.ecommerce
""")
```
> ⏸ See performance (1)

DataFrame[]

Create Volume

```
04:11 PM (1s)                                              7

spark.sql("""
CREATE VOLUME IF NOT EXISTS workspace.ecommerce.ecommerce_data
""")
```
> ⏸ See performance (1)

DataFrame[]

Download Dataset to Volume

```
%sh
cd /Volumes/workspace/ecommerce/ecommerce_data
kaggle datasets download -d mkechinov/ecommerce-behavior-data-from-multi-category-store
```

Dataset URL: https://www.kaggle.com/datasets/mkechinov/ecommerce-behavior-data-from-multi-category-store
License(s): copyright-authors
Downloading ecommerce-behavior-data-from-multi-category-store.zip to /Volumes/workspace/ecommerce/ecommerce_data
100%|████████████| 4.29G/4.29G [00:26<00:00, 176MB/s]

Unzip Dataset

```
%sh
cd /Volumes/workspace/ecommerce/ecommerce_data
unzip -o ecommerce-behavior-data-from-multi-category-store.zip
ls -lh
```

Archive: ecommerce-behavior-data-from-multi-category-store.zip
  inflating: 2019-Nov.csv
  inflating: 2019-Oct.csv
total 18G
-rwxrwxrwx 1 spark-f3ac2c7a-ca17-48b2-b58a-61 nogroup 8.4G Jan 9 10:43 2019-Nov.csv
```

## Remove ZIP (keep CSVs)

```sh
%sh
cd /Volumes/workspace/ecommerce/ecommerce_data
rm -f ecommerce-behavior-data-from-multi-category-store.zip
ls -lh
```

04:22 PM (1s)   13   Python

```
total 14G
-rwxrwxrwx 1 spark-f3ac2c7a-ca17-48b2-b58a-61 nogroup 8.4G Jan  9 10:43 2019-Nov.csv
-rwxrwxrwx 1 spark-f3ac2c7a-ca17-48b2-b58a-61 nogroup 5.3G Jan  9 10:46 2019-Oct.csv
```

Read Code (CSV File)

04:31 PM (58s)   15

```python
oct_events = spark.read.csv(
    "/Volumes/workspace/ecommerce/ecommerce_data/2019-Oct.csv",
    header = True,
    inferSchema = True
)
```

```
nov_events.display()
```

> ⯈ ıllı See performance (1)

Table ⌄　＋　　　　　　　　　　　　　　　　　🔍 ⏷ ⋮ ▢

	🕑 event_time	ᴬᴮ꜀ event_type	¹²₃ product_id	¹²₃ category_id	ᴬᴮ꜀ category_code ⋮ ⇅	ᴬᴮ꜀ brand	1.2 price	¹²₃ user_
5	2019-11-01T00:00:01.000+00:...	view	1004775	20530135556318826...	electronics.smartphone	xiaomi	183.27	55
6	2019-11-01T00:00:01.000+00:...	view	1306894	20530135589202171...	computers.notebook	hp	360.09	52
7	2019-11-01T00:00:01.000+00:...	view	1306421	20530135589202171...	computers.notebook	hp	514.56	51
8	2019-11-01T00:00:02.000+00:...	view	15900065	20530135581904082...	null	rondell	30.86	51
9	2019-11-01T00:00:02.000+00:...	view	12708937	20530135535598963...	null	michelin	72.72	53
10	2019-11-01T00:00:02.000+00:...	view	1004258	20530135556318826...	electronics.smartphone	apple	732.07	53
11	2019-11-01T00:00:03.000+00:...	view	17200570	20530135597926324...	furniture.living_room.sofa	null	437.33	51
12	2019-11-01T00:00:03.000+00:...	view	2701517	20530135639114392...	appliances.kitchen.refrigerators	null	155.11	51
13	2019-11-01T00:00:03.000+00:...	view	16700260	20530135599016843...	furniture.kitchen.chair	null	31.64	56
14	2019-11-01T00:00:04.000+00:...	view	34600011	20609813205819064...	null	null	20.54	51
15	2019-11-01T00:00:05.000+00:...	view	4600658	20530135639449936...	appliances.kitchen.dishwasher	samsung	411.83	52
16	2019-11-01T00:00:06.000+00:...	view	24900193	20530135621833858...	null	null	1.09	51
17	2019-11-01T00:00:07.000+00:...	view	27400066	20530135633913454...	null	null	8.55	55
18	2019-11-01T00:00:07.000+00:...	view	5100503	20530135533753469...	null	xiaomi	22.68	52

10

▶ ✓ 04:33 PM (1s)   17   🗑

oct_events.display()

> 📊 See performance (1)

Table ⌄   + ▯   🔍 🔽 ▯ ▱

	event_time	event_type	product_id	category_id	category_code	brand	price	user_
1	2019-10-01T00:00:00.000+00:...	view	44600062	21038074595953877...	null	shiseido	35.79	54
2	2019-10-01T00:00:00.000+00:...	view	3900821	20530135523267709...	appliances.environment.water_heat...	aqua	33.2	55
3	2019-10-01T00:00:01.000+00:...	view	17200506	20530135597926324...	furniture.living_room.sofa	null	543.1	51
4	2019-10-01T00:00:01.000+00:...	view	1307067	20530135589202171...	computers.notebook	lenovo	251.74	55
5	2019-10-01T00:00:04.000+00:...	view	1004237	20530135556318826...	electronics.smartphone	apple	1081.98	53
6	2019-10-01T00:00:05.000+00:...	view	1480613	20530135610928667...	computers.desktop	pulser	908.62	51
7	2019-10-01T00:00:08.000+00:...	view	17300353	20530135538534976...	null	creed	380.96	55
8	2019-10-01T00:00:08.000+00:...	view	31500053	20530135580310246...	null	luminarc	41.16	55
9	2019-10-01T00:00:10.000+00:...	view	28719074	20530135654801090...	apparel.shoes.keds	baden	102.71	52
10	2019-10-01T00:00:11.000+00:...	view	1004545	20530135556318826...	electronics.smartphone	huawei	566.01	53
11	2019-10-01T00:00:11.000+00:...	view	2900536	20530135547762445...	appliances.kitchen.microwave	elenberg	51.46	55
12	2019-10-01T00:00:11.000+00:...	view	1005011	20530135556318826...	electronics.smartphone	samsung	900.64	53
13	2019-10-01T00:00:13.000+00:...	view	3900746	20530135523267709...	appliances.environment.water_heat...	haier	102.38	55
14	2019-10-01T00:00:15.000+00:...	view	44600062	21038074595953877...	null	shiseido	35.79	54