

Week 6 (24 Sep - 30 Sep) Week 7 (1 Oct - 7 Oct)

Project: Building an ETL Pipeline to Transform Data in Data Lake (From Landing Zone to Cleaned Zone)

สามารถเริ่มต้นจากโค้ดที่ <https://github.com/zkan/swu-ds525> ได้
ในโปรเจคนี้เราจะใช้ชื่อ模 GitHub event data จาก [API](#) หรือสามารถดาวน์โหลดไฟล์ JSON ได้ที่ [URL](#) นี้
สิ่งที่คาดหวังในโปรเจคนี้
ในโปรเจคนี้เราจะตั้งค่านว่าเรามีไฟล์อยู่บน S3 อญญาแล้ว ดังนั้นทุกคนสามารถที่จะอัปโหลดไฟล์ขึ้นไปไว้บน S3 ได้
โดย

1. มีโค้ดที่ใช้ Spark ทำ ETL จากชื่อ模 JSON ใน S3 ที่ landing zone และสร้างไฟล์ที่ clean แล้ว
ใน cleaned zone ใน data lake (S3)
 2. มีการเขียน documentation อธิบายสิ่งที่ตัวเองทำลงไว้
 3. มี instruction ในการรันโค้ดของตัวเอง
-

หนังสือภาษาไทย Spark Internal

Click <https://github.com/aorjoa/SparkInternals/tree/thai/markdown/thai> link to open resource.

Spark By Examples | Learn Spark Tutorial with Examples

Click <https://sparkbyexamples.com/> link to open resource.

Introduction to PySpark

Click <https://www.datacamp.com/courses/introduction-to-pyspark> link to open resource.

Data Wrangling with Spark

Click <https://github.com/zkan/data-wrangling-with-spark> link to open resource.

Machine Learning with Spark and Zeppelin

Click <https://github.com/zkan/machine-learning-with-spark-and-zeppelin> link to open resource.

Read Nested JSON in Spark DataFrame

Click <https://bigdataprogrammers.com/read-nested-json-in-spark-dataframe/> link to open resource.

How to parse nested JSON objects in spark sql?

Click

<https://stackoverflow.com/questions/29948789/how-to-parse-nested-json-objects-in-spark-sql>

Slides: Week 7

<https://docs.google.com/presentation/d/1gZOfJJ9TzDLQ5ZRAAdTLd74ePKzeCSBFVSJsMt7LRx6M/edit?usp=sharing>

Week 6 (Part 1)

<https://youtu.be/mWkGUzsarAQ>

Week 6 (Part 2)

<https://youtu.be/sesxWvbJnzo>

Week 6 Project Walkthru

<https://youtu.be/1sVj7mwAmTE>

```
spark.sql("""  
    select  
        upc  
        , product_size  
        , split(product_size, ' ')[0] as value  
        , split(product_size, ' ')[1] as unit  
        , case  
            when split(product_size, ' ')[1] = 'OZ' then split(product_size, ' ')[0] * 100  
            else  
                split(product_size, ' ')[0] / 100  
            end as is_oz
```

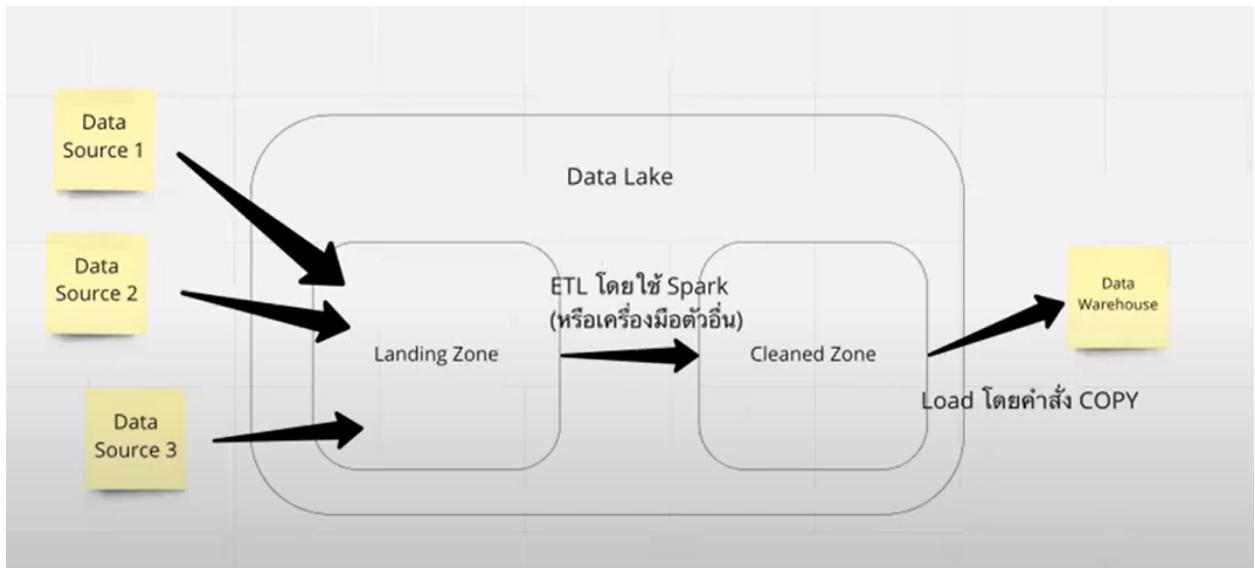
Week 7

<https://youtu.be/c1RtJhyIo9s>

Week 7 Office Hour

https://youtu.be/ZIEPt0OI_w0

Week 6 (24 Sep - 30 Sep)



```
04-building-a-data-lake > {} github_events_01.json > {} org
```

```

1  {
2    "id": "23487929637",
3    "type": "IssueCommentEvent",
4    "actor": { ...
11   },
12   "repo": { ...
16   },
17   "payload": { ...
171  },
172   "public": true,
173   "created_at": "2022-08-17T15:51:05Z",
174   "org": { ...
180  }
181 }
```

Actor, repo, payload เป็นอีก 1 object = nestest

```
04-building-a-data-lake > {} github_events_01.json > {} actor > login
1  {
2    "id": "23487929637",
3    "type": "IssueCommentEvent",
4    "actor": {
5      "id": 1696078,
6      "login": "sukhada",
7      "display_login": "sukhada",
8      "gravatar_id": "",
9      "url": "https://api.github.com/users/sukhada",
10     "avatar_url": "https://avatars.githubusercontent.com/u/1696078?"
11   },
12   "repo": {...},
13   "payload": {...},
14   "public": true,
15   "created_at": "2022-08-17T15:51:05Z",
16   "org": {...}
17 }
```

ต้องแกะ “sukhada” มาใช้

Notebook โดย docker-compose

Run compuse up

```
docker-compose.yml (compose-spec.json)
version: '3.8'

services:
  pyspark-notebook:
    image: jupyter/pyspark-notebook
    volumes:
      - .:/home/jovyan/work
    ports:
      - 8888:8888
      - 4040:4040
      - 4041:4041
```

Copy token

```
gitpod /workspace/swu-ds525 (main) $ cd 04-building-a-data-lake/
gitpod /workspace/swu-ds525/04-building-a-data-lake (main) $ python -m venv ENV
gitpod /workspace/swu-ds525/04-building-a-data-lake (main) $ source ENV/bin/activate
ERROR: Could not open requirements file: [Errno 2] No such file or directory: 'requirements.txt'xt WARNING: You are using pip version 22.0.4; however, version 22.3.1 is available.
You should consider upgrading via the '/workspace/swu-ds525/04-building-a-data-lake/ENV/bin/python -m pip install --upgrade pip' command.
(ENV) gitpod /workspace/swu-ds525/04-building-a-data-lake (main) $ docker-compose up
[+] Running 2/0
  ● Network 04-building-a-data-lake_default          Created          0.0s
  ● Container 04-building-a-data-lake-pyspark-notebook_1  Created          0.0s
Attaching to 04-building-a-data-lake-pyspark-notebook_1
04-building-a-data-lake-pyspark-notebook_1  | Entered start.sh with args: jupyter lab
04-building-a-data-lake-pyspark-notebook_1  | /usr/local/bin/start.sh: running hooks in /usr/local/bin/before-notebook.d as uid / gid: 1000 / 100
04-building-a-data-lake-pyspark-notebook_1  | /usr/local/bin/start.sh: running script /usr/local/bin/before-notebook.d/spark-config.sh
04-building-a-data-lake-pyspark-notebook_1  | /usr/local/bin/start.sh: done running hooks in /usr/local/bin/before-notebook.d
04-building-a-data-lake-pyspark-notebook_1  | Executing the command: jupyter lab
04-building-a-data-lake-pyspark-notebook_1  | [I 2022-11-06 16:05:25.998 ServerApp] jupyterlab | extension was successfully linked.
04-building-a-data-lake-pyspark-notebook_1  | [I 2022-11-06 16:05:26.008 ServerApp] nbclassic | extension was successfully linked.
04-building-a-data-lake-pyspark-notebook_1  | [I 2022-11-06 16:05:26.009 ServerApp] Writing Jupyter server cookie secret to /home/jovyan/.local/share/jupyter/runtime/jupyter_cookie_secret
04-building-a-data-lake-pyspark-notebook_1  | [I 2022-11-06 16:05:26.283 ServerApp] notebook_shim | extension was successfully linked.
04-building-a-data-lake-pyspark-notebook_1  | [I 2022-11-06 16:05:26.303 ServerApp] notebook_shim | extension was successfully loaded.
04-building-a-data-lake-pyspark-notebook_1  | [I 2022-11-06 16:05:26.304 LabApp] JupyterLab extension loaded from /opt/conda/lib/python3.10/site-packages/jupyterlab
04-building-a-data-lake-pyspark-notebook_1  | [I 2022-11-06 16:05:26.308 LabApp] JupyterLab application directory is /opt/conda/share/jupyter/lab
04-building-a-data-lake-pyspark-notebook_1  | [I 2022-11-06 16:05:26.308 ServerApp] jupyterlab | extension was successfully loaded.
04-building-a-data-lake-pyspark-notebook_1  | [I 2022-11-06 16:05:26.312 ServerApp] nbclassic | extension was successfully loaded.
04-building-a-data-lake-pyspark-notebook_1  | [I 2022-11-06 16:05:26.313 ServerApp] Serving notebooks from local directory: /home/jovyan
04-building-a-data-lake-pyspark-notebook_1  | [I 2022-11-06 16:05:26.313 ServerApp] Jupyter Server 1.19.1 is running at:
04-building-a-data-lake-pyspark-notebook_1  | [I 2022-11-06 16:05:26.313 ServerApp] http://5f570511aeab:8888/lab?token=a2de5ed5845647d37c3fc5d5bd1ea663debdbcb7a81d9a29
```

เข้า jupyter

The screenshot shows a Jupyter Notebook interface running in a browser window. The title bar indicates the URL is 8888-penguinaw-swuds525-n5vde124t1b.ws-us74.gitpod.io/lab/tree/work/elt_local.ipynb. The browser toolbar includes tabs for poll_form.php, Canny edge detect..., Your Repositories, Simple guide on ho..., A Neural Network P..., Learner Lab, Image Feature Extra..., Canny edge detect..., How to tune hyper..., and a refresh icon.

The main area has a header with File, Edit, View, Run, Kernel, Tabs, Settings, Help, and a Launcher tab. The Launcher tab is active, showing a list of recent notebooks: ETL with Spark (Local), elt_local.ipynb, and another unnamed notebook. The elt_local.ipynb tab is selected.

The notebook content displays Python code for ETL with Spark:

```
from pyspark.sql import SparkSession
# from pyspark.sql.types import StructType, StructField, DoubleType, StringType, IntegerType, DateType, TimestampType
# import pyspark.sql.functions as F

import pandas as pd
import glob

p = glob.glob("data/*.json")
p
pd.read_json(p)

data = "github_events_01.json"
data_2 = "github_events_02.json"

spark = SparkSession.builder \
    .appName("ETL") \
    .getOrCreate()

data_folder = "data"

data = spark.read.option("multiline", "true").json(data_folder)

# data = spark.read.option("multiline", "true").json(data_2)

data.show()
data.printSchema()
data.select("id", "type").show()
data.createOrReplaceTempView("staging_events")

table = spark.sql"""
    select
```

The status bar at the bottom shows "Simple 0 1 Python 3 (ipykernel) | Idle" on the left and "Mode: Command Ln 1, Col 1 elt_local.ipynb" on the right.

ถ้าหาคำสั่งใช้



spark select columns pyspark

X



```
df.select("firstname", "lastname").show()
df.select(df.firstname, df.lastname).show()
df.select(df["firstname"], df["lastname"]).show()

#By using col() function
from pyspark.sql.functions import col
df.select(col("firstname"), col("lastname")).show()

#Select columns by regular expression
df.select(df.colRegex("^.*name*$")).show()
```

[24]: `data.select("id", "type").show()`

```
+-----+-----+
|      id|      type|
+-----+-----+
|23487929637|IssueCommentEvent|
|23487929676|      PushEvent|
|23487929674|      PushEvent|
|23487929661|      PushEvent|
|23487929682|      PushEvent|
|23487929673|      PushEvent|
|23487929588|      PushEvent|
|23487929636|CreateEvent|
|23487929580|IssuesEvent|
|23487929591|      PushEvent|
|23487929533|      PushEvent|
|23487929573|      PushEvent|
|23487929349|      PushEvent|
|23487929578|      PushEvent|
|23487929597|IssueCommentEvent|
|23487929522|ReleaseEvent|
|23487929560|      PushEvent|
|23487929536|      PushEvent|
|23487929501|DeleteEvent|
|23487929523|      PushEvent|
+-----+
only showing top 20 rows
```

ແກະຂໍ້ມູນລືທີ່ອ່າງຸດ້ານໃນ object ທີ່ nestest ອີງ

```
import org.apache.spark.sql.functions._  
var parseOrdersDF = ordersDF.withColumn("orders", explode($"datasets"))  
  
// Step 3: Fetch Each Order using getItem on explode column  
parseOrdersDF = parseOrdersDF.withColumn("customerId", $"orders".getItem("customerId"))  
                                .withColumn("orderId", $"orders".getItem("orderId"))  
                                .withColumn("orderDate", $"orders".getItem("orderDate"))  
                                .withColumn("orderDetails", $"orders".getItem("orderDetails"))  
                                .withColumn("shipmentDetails", $"orders".getItem("shipmentDetails"))
```

Password or token:  Log in

Token authentication is enabled

If no password has been configured, you need to open the server with its login token in the URL, or paste it above. This requirement will be lifted if you enable a password.

The command:

```
jupyter server list
```

will show you the URLs of running servers with their tokens, which you can copy and paste into your browser. For example:

```
Currently running servers:  
http://localhost:8888/?token=c8de56fa... :: /Users/you/notebooks
```

or you can paste just the token value into the password field on this page.

See [the documentation on how to enable a password](#) in place of token authentication, if you would like to avoid dealing with random tokens.

Cookies are required for authenticated access to the Jupyter server.

Setup a Password

You can also setup a password by entering your token and a new password on the fields below:

Token

New Password

Token

New Password

ด้วยร่างคำสั่งpartition

File Edit View Run Kernel Tabs Settings Help

Launcher etl_local.ipynb

	[id]	type	created_at	date[year]	login	actor_url	name	repo_url
1	23487963576	WatchEvent	2022-08-17T15:52:40Z	2022-08-17	evilgaoshu	https://api.github.com/	spring-project/spring	https://api.github.com/
2	23487963624	CreateEvent	2022-08-17T15:52:40Z	2022-08-17	gurram47	https://api.github.com/	gurram47/AP201190...	https://api.github.com/
3	23487963529	PushEvent	2022-08-17T15:52:40Z	2022-08-17	afbeltran	https://api.github.com/	afbeltran/Agrilab2	https://api.github.com/
4	23487963558	IssueCommentEvent	2022-08-17T15:52:40Z	2022-08-17	karla-vm	https://api.github.com/	OSgov/cms-carts...	https://api.github.com/
5	23487963581	PullRequestEvent	2022-08-17T15:52:40Z	2022-08-17	hsluoyz	https://api.github.com/	casdoor/casdoor-c...	https://api.github.com/
6	23487963532	PushEvent	2022-08-17T15:52:40Z	2022-08-17	mnn1020	https://api.github.com/	mnn1020/obisidian	https://api.github.com/
7	23487963524	PushEvent	2022-08-17T15:52:40Z	2022-08-17	ikj093	https://api.github.com/	ijkj093/Data-Struc...	https://api.github.com/
8	23487963526	PushEvent	2022-08-17T15:52:40Z	2022-08-17	Gabe616	https://api.github.com/	Gabe616/ObbyCreat...	https://api.github.com/
9	23487963492	PushEvent	2022-08-17T15:52:40Z	2022-08-17	BadProfessor	https://api.github.com/	BadProfessor/BadPr...	https://api.github.com/
10	23487963504	DeleteEvent	2022-08-17T15:52:40Z	2022-08-17	allang4	https://api.github.com/	ALMA-FUNDEQUA/vac...	https://api.github.com/
11	23487963536	PullRequestReviewEvent	2022-08-17T15:52:40Z	2022-08-17	QGarchery	https://api.github.com/	morpho-dao/morpho...	https://api.github.com/
12	23487963495	CreateEvent	2022-08-17T15:52:40Z	2022-08-17	Diyouf	https://api.github.com/	Diyouf/nepage.gi...	https://api.github.com/
13	23487963522	PushEvent	2022-08-17T15:52:40Z	2022-08-17	tiltingpenguin	https://api.github.com/	tiltingpenguin/yunni	https://api.github.com/
14	23487963444	PushEvent	2022-08-17T15:52:40Z	2022-08-17	igrek-ovs	https://api.github.com/	igrek-ovs/Irek-o...	https://api.github.com/
15	23487963462	PullRequestEvent	2022-08-17T15:52:40Z	2022-08-17	channan	https://api.github.com/	TeamDearToday/Dea...	https://api.github.com/
16	23487963480	IssuesEvent	2022-08-17T15:52:40Z	2022-08-17	mvashishtha	https://api.github.com/	modin-project/modin	https://api.github.com/
17	23487963457	PushEvent	2022-08-17T15:52:40Z	2022-08-17	na4zagin3	https://api.github.com/	na4zagin3/satvrog...	https://api.github.com/
18	23487963413	PushEvent	2022-08-17T15:52:40Z	2022-08-17	xsidc	https://api.github.com/	xsidc/monniao	https://api.github.com/
19	23487963429	PushEvent	2022-08-17T15:52:40Z	2022-08-17	kkukelka	https://api.github.com/	kkukelka/nft-gallery	https://api.github.com/
20	23487963448	PushEvent	2022-08-17T15:52:40Z	2022-08-17	Sidalivk	https://api.github.com/	Sidalivk/005-Sida...	https://api.github.com/

only showing top 20 rows

```

[15]: output_csv = ".../output_csv"
output_parquet = ".../output_parquet"

[16]: table.write.partitionBy("year").mode("overwrite").csv(output_csv)

[17]: table.write.partitionBy("year").mode("overwrite").parquet(output_parquet)

[ ]: table = spark.sql("""
    select
        id
        , type
        , created_at
        , day(created_at) as day
        , month(created_at) as month
        , year(created_at) as year
        , date(created_at) as date
    from
        staging_events
""")

```

📁 / output_csv / year=2022 /

Name	Last Modified
part-00000-13392ed7-93d2-4693-bb8a-1544ecf3a2b0.c000.csv	a minute ago
part-00001-13392ed7-93d2-4693-bb8a-1544ecf3a2b0.c000.csv	a minute ago

File Edit View Run Kernel Tabs Settings Help

Launcher etl_local.ipynb

	[id]	type	created_at	date[year]	login	actor_url	name	repo_url
1	2347963578	WatchEvent	2022-08-17T15:52:40Z	2022-08-17	evilgaoshu	https://api.github.com/	spring-project/spring	https://api.github.com/
2	2347963524	CreateEvent	2022-08-17T15:52:40Z	2022-08-17	gurram47	https://api.github.com/	gurram47/AP201190...	https://api.github.com/
3	2347963528	PushEvent	2022-08-17T15:52:40Z	2022-08-17	afbeltran	https://api.github.com/	afbeltran/Agrilab2	https://api.github.com/
4	2347963558	IssueCommentEvent	2022-08-17T15:52:40Z	2022-08-17	karla-vm	https://api.github.com/	OSgov/cms-carts...	https://api.github.com/
5	2347963581	PullRequestEvent	2022-08-17T15:52:40Z	2022-08-17	hsluoyz	https://api.github.com/	casdoor/casdoor-c...	https://api.github.com/
6	2347963532	PushEvent	2022-08-17T15:52:40Z	2022-08-17	mnn1020	https://api.github.com/	mnn1020/obisidian	https://api.github.com/
7	2347963526	PushEvent	2022-08-17T15:52:40Z	2022-08-17	ikj093	https://api.github.com/	ijkj093/Data-Struc...	https://api.github.com/
8	2347963492	PushEvent	2022-08-17T15:52:40Z	2022-08-17	Gabe616	https://api.github.com/	Gabe616/ObbyCreat...	https://api.github.com/
9	2347963504	DeleteEvent	2022-08-17T15:52:40Z	2022-08-17	allang4	https://api.github.com/	ALMA-FUNDEQUA/vac...	https://api.github.com/
10	2347963536	PullRequestReviewEvent	2022-08-17T15:52:40Z	2022-08-17	QGarchery	https://api.github.com/	morpho-dao/morpho...	https://api.github.com/
11	2347963545	CreateEvent	2022-08-17T15:52:40Z	2022-08-17	Diyouf	https://api.github.com/	Diyouf/nepage.gi...	https://api.github.com/
12	2347963522	PushEvent	2022-08-17T15:52:40Z	2022-08-17	tiltingpenguin	https://api.github.com/	tiltingpenguin/yunni	https://api.github.com/
13	2347963444	PushEvent	2022-08-17T15:52:40Z	2022-08-17	igrek-ovs	https://api.github.com/	igrek-ovs/Irek-o...	https://api.github.com/
14	2347963462	PullRequestEvent	2022-08-17T15:52:40Z	2022-08-17	channan	https://api.github.com/	TeamDearToday/Dea...	https://api.github.com/
15	2347963480	IssuesEvent	2022-08-17T15:52:40Z	2022-08-17	mvashishtha	https://api.github.com/	modin-project/modin	https://api.github.com/
16	2347963457	PushEvent	2022-08-17T15:52:40Z	2022-08-17	na4zagin3	https://api.github.com/	na4zagin3/satvrog...	https://api.github.com/
17	2347963413	PushEvent	2022-08-17T15:52:40Z	2022-08-17	xsidc	https://api.github.com/	xsidc/monniao	https://api.github.com/
18	2347963429	PushEvent	2022-08-17T15:52:40Z	2022-08-17	kkukelka	https://api.github.com/	kkukelka/nft-gallery	https://api.github.com/
19	2347963448	PushEvent	2022-08-17T15:52:40Z	2022-08-17	Sidalivk	https://api.github.com/	Sidalivk/005-Sida...	https://api.github.com/
20	2347963528	PushEvent	2022-08-17T15:52:40Z	2022-08-17	qikee	https://api.github.com/	qikee/cluster-qikee	https://api.github.com/
21	2347963491	IssueCommentEvent	2022-08-17T15:52:40Z	2022-08-17	mvashishtha	https://api.github.com/	modin-project/modin	https://api.github.com/
22	2347963458	IssueCommentEvent	2022-08-17T15:52:40Z	2022-08-17	meiy400	https://api.github.com/	meiy400/meiy...	https://api.github.com/
23	2347963428	WebEvent	2022-08-17T15:52:40Z	2022-08-17	exphel	https://api.github.com/	bwmarrive/decide...	https://api.github.com/
24	2347963471	PullRequestEvent	2022-08-17T15:52:40Z	2022-08-17	Indriyawett	https://api.github.com/	RedHiroto/Indriy...	https://api.github.com/
25	2347963425	CreateEvent	2022-08-17T15:52:40Z	2022-08-17	wenesis2cm202	https://api.github.com/	wenesis2cm202...	https://api.github.com/
26	2347963398	PushEvent	2022-08-17T15:52:40Z	2022-08-17	sigocia-test-unsigned	https://api.github.com/	google-test/sigocia...	https://api.github.com/
27	2347963384	PushEvent	2022-08-17T15:52:40Z	2022-08-17	LombiqOrchard	https://api.github.com/	LombiqOrchard	https://api.github.com/
28	2347963376	PushEvent	2022-08-17T15:52:40Z	2022-08-17	KHURWILLIAMS	https://api.github.com/	HARPHUNA/gpc-9ml	https://api.github.com/
29	2347963368	PushEvent	2022-08-17T15:52:40Z	2022-08-17	khurwilliams	https://api.github.com/	khurwilliams/microblog	https://api.github.com/

part-00000-13392ed7-93d2-4693-bb8a-1544ecf3a2b0.c000.csv

```
[31]: table = spark.sql("""
    select
        actor.login
        , id as event_id
        , actor.url as actor_url
    from
        staging_events
""")
destination = "../actors"
table.write.mode("overwrite").csv(destination)
```

```
[32]: table = spark.sql("""
    select
        repo.name
        , id as event_id
        , repo.url as repo_url
    from
        staging_events
""")
destination = "../repos"
table.write.mode("overwrite").csv(destination)
```

■ /

Name	Last Modified
■ actors	2 minutes ago
■ events	4 minutes ago
■ output_csv	13 minutes ago
■ output_parquet	13 minutes ago
■ repos	a minute ago
■ work	an hour ago
▣ etl_local.ipynb	36 minutes ago
· ▣ etl_local2.ipynb	23 minutes ago

```
[24]: table = spark.sql("""
    select
        id
        , type
        , created_at
        , year(created_at) as year
    from
        staging_events
""")
```

```
[25]: destination = ".../events"
```

```
[26]: table.write.partitionBy("year").mode("overwrite").csv(destination)
```

The screenshot shows a Jupyter Notebook interface with several tabs at the top: 'Launcher', 'etl_local2.ipynb', 'etl_local.ipynb', and 'part-00000-927ec590-d026-X'. Below the tabs is a file browser with a search bar and a 'Filter files by name' dropdown. The browser lists two main directory entries: '/events / year=2022 /' and two CSV files: 'part-00000-927ec590-d026-42c9-99b6-4ddb624e7654.c000.csv' and 'part-00001-927ec590-d026-42c9-99b6-4ddb624e7654.c000.csv', both modified 'a minute ago'. To the right of the browser is a preview of the 'part-00000...' CSV file, which contains four rows of event data:

	23487963578	WatchEvent	2022-08-17T15:52:40Z
1	23487963624	CreateEvent	2022-08-17T15:52:40Z
2	23487963529	PushEvent	2022-08-17T15:52:40Z
3	23487963558	IssueCommentEvent	2022-08-17T15:52:40Z
4	23487963581	PullRequestEvent	2022-08-17T15:52:40Z

```
[29]: table = spark.sql("""
    select
        id
        , type
        , created_at
        , day(created_at) as day
        , month(created_at) as month
        , year(created_at) as year
    from
        staging_events
""")
[]:
```

```
[30]: table.show()
```

	id	type	created_at	day	month	year
1	23487963576	WatchEvent	2022-08-17T15:52:40Z	17	8	2022
2	23487963624	CreateEvent	2022-08-17T15:52:40Z	17	8	2022
3	23487963529	PushEvent	2022-08-17T15:52:40Z	17	8	2022
4	23487963558	IssueCommentEvent	2022-08-17T15:52:40Z	17	8	2022
5	23487963581	PullRequestEvent	2022-08-17T15:52:40Z	17	8	2022
6	23487963532	PushEvent	2022-08-17T15:52:40Z	17	8	2022
7	23487963524	PushEvent	2022-08-17T15:52:40Z	17	8	2022
8	23487963526	PushEvent	2022-08-17T15:52:40Z	17	8	2022
9	23487963492	PushEvent	2022-08-17T15:52:40Z	17	8	2022
10	23487963504	DeleteEvent	2022-08-17T15:52:40Z	17	8	2022
11	23487963536	PullRequestReviewEvent	2022-08-17T15:52:40Z	17	8	2022
12	23487963495	CreateEvent	2022-08-17T15:52:40Z	17	8	2022
13	23487963522	PushEvent	2022-08-17T15:52:40Z	17	8	2022
14	23487963444	PushEvent	2022-08-17T15:52:40Z	17	8	2022
15	23487963462	PullRequestEvent	2022-08-17T15:52:40Z	17	8	2022
16	23487963488	IssuesEvent	2022-08-17T15:52:40Z	17	8	2022
17	23487963457	PushEvent	2022-08-17T15:52:40Z	17	8	2022
18	23487963413	PushEvent	2022-08-17T15:52:40Z	17	8	2022
19	23487963429	PushEvent	2022-08-17T15:52:40Z	17	8	2022
20	23487963448	PushEvent	2022-08-17T15:52:40Z	17	8	2022

only showing top 20 rows

```
[31]: destination = "../events"
[32]: table.write.partitionBy("year", "month", "day").mode("overwrite").csv(destination)
```

■ / events / year=2022 /

Name

■ month=8

■ / ... / month=8 / day=17 /

Name	Last Modified
part-00000-e93f3216-d9e7-4eba-aa3f-517abc4375b8.c000.csv	2 minutes ago
part-00001-e93f3216-d9e7-4eba-aa3f-517abc4375b8.c000.csv	2 minutes ago

Launcher						
	23487963576	WatchEvent	2022-08-17T15:52:40Z	17	8	2022
1	23487963624	CreateEvent	2022-08-17T15:52:40Z	17	8	2022
2	23487963529	PushEvent	2022-08-17T15:52:40Z	17	8	2022
3	23487963558	IssueCommentEvent	2022-08-17T15:52:40Z	17	8	2022
4	23487963581	PullRequestEvent	2022-08-17T15:52:40Z	17	8	2022

Name	Last Modified
actors	5 minutes ago
events	14 minutes ago
output_csv	26 minutes ago
output_parquet	26 minutes ago
repos	8 minutes ago
work	2 hours ago
etl_local.ipynb	an hour ago
etl_local2.ipynb	an hour ago

Name	Last Modified
actors	17 minutes ago
events	26 minutes ago
output_csv	37 minutes ago
output_parquet	37 minutes ago
repos	19 minutes ago
work	2 hours ago
etl_local.ipynb	2 hours ago
etl_local2.ipynb	4 minutes ago
new-workspace.jupyterlab-workspace	2 minutes ago

■ / work /

Name	Last Modified
■ data	a day ago
■ ENV	a day ago
■ workshop	22 days ago
🔗 04-building-a-data-lake_doc - Google Docs.pdf	a day ago
Y: docker-compose.yml	a day ago
• etl_local.ipynb	2 hours ago
etl.ipynb	2 hours ago
⌚ github_events_01.json	2 hours ago
notes.txt	2 hours ago
M: Readme.md	a day ago
M: README.md	a day ago

← → C 8888-penguina-swuds525-n5vde124t1b.ws-us74.gtpod.io/lab/tree/etl_local2.ipynb

poll_form.php Canny edge detect... Your Repositories Simple guide on ho... A Neural Network P... Learner Lab Image Feature Extra... Canny edge detect... How to tune hyper... Machine Learn

File Edit View Run Kernel Tabs Settings Help

etl_local2.ipynb etl_local.ipynb part-00000-ee867841-6c29..X part-00000-acaff0d-e3cf-4f..X

Filter files by name

■ / work /

Name	Last Modified
■ data	a day ago
■ ENV	a day ago
■ workshop	22 days ago
🔗 04-building-a-data-lake_doc - Google Docs.pdf	a day ago
Y: docker-compose.yml	a day ago
• etl_local.ipynb	2 hours ago
etl.ipynb	2 hours ago
⌚ github_events_01.json	2 hours ago
notes.txt	2 hours ago
M: Readme.md	a day ago
M: README.md	a day ago

ETL with Spark (Local)

```
[1]: from pyspark.sql import SparkSession
# from pyspark.sql.types import StructType, StructField, DoubleType, StringType, IntegerType, DateType, TimeType
# import pyspark.sql.functions as F

[2]: import pandas as pd
import glob

[3]: p = glob.glob("data/*.json")
[4]: p
[5]: #pd.read_json(p)

[6]: data = "github_events_01.json"

[7]: data_2 = "github_events_02.json"
```

ໃຫ້ Notebook ໃນ AWS

```
ddd_v1_w_foQ_1383611@runweb62413:~$ cat ~/.aws/credentials
[default]
aws_access_key_id = ASIARC3CKZR2WNZXQLZA
aws_secret_access_key = puIUVfojgB6Ws4jyvv/AeCjBP10kqB0W8feUsn7T
aws_session_token = FwoGZXIvYXdzECMaDLfAeXU6AITLcZujViLPAYzC1NjTEMc2LtjXgOmEn7nXiv6AUMyTfx5N64EU
uIojF15QZ61MwDMeJJY4vYQpL0ATv5NaZdOo0B0C+vT2qdtL5eVsLxPjWJvw2moTv3He4SKfZagR2T/p43M+3igpVH3L6/1l
TledcIgTqh8A5brllWrg1GJ+eLvLjPpeUBQ4/vdJo+dea169aTM8gNkbgb5YCirrzM56kUkCXVUlcw//xklpxZEe3lc89X1kr
ulY1LCp05WC1K0a10rqPbq5Gcp0fBfxjfORmoyjbmlWBIHyi5p7uZBjItbx/VwVL9mCXoR3+lW3nxGUaoYBM3gNeUNHKFE0X
g/a6Wkvld6Ym4Lm1MN9W
ddd_v1_w_foQ_1383611@runweb62413:~$
```

← → C us-east-1.console.aws.amazon.com/elasticmapreduce/home?region=us-east-1#

poll_form.php Canny edge detect... Simple guide on ho... A Neural Network P... Learner Lab Canny edge detect... Machine Learning u... TensorFlow Lite Co...

aws Services Search for services, features, blogs, docs, and more [Alt+S] N. Virginia v vclabs/user1591029=peeyapak.somvitoon@g.swu.ac.th @ 0748-312... ▾

Amazon EMR

EMR Studio

EMR Serverless New

EMR on EC2

Clusters

Notebooks

Git repositories

Security configurations

Block public access

VPC subnets

Events

EMR on EKS

Virtual clusters

Feedback Looking for language selection? Find it in the new Unified Settings ▾

Welcome to Amazon Elastic MapReduce

Amazon Elastic MapReduce (Amazon EMR) is a web service that enables businesses, researchers, data analysts, and developers to easily and cost-effectively process vast amounts of data.

You do not appear to have any clusters. Create one now:

Create cluster

How Elastic MapReduce Works

Upload Create Monitor

© 2022, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences

Additional Information

More about Elastic MapReduce

EMR overview FAQs Pricing

More Help Using Elastic MapReduce

Forum Documentation Developer Guide API Reference EMR on GitHub Help portal

Welcome to AI

Amazon Elastic MapReduce
analysts, and developers to

You do not appear to have a

Create cluster

Create Cluster - Quick Options [Go to advanced options](#)

General Configuration

Cluster name

Logging [i](#)

S3 folder

Launch mode Cluster [i](#) Step execution [i](#)

Software configuration

Release [i](#)

Applications Core Hadoop: Hadoop 3.2.1 with Hive 3.1.3, Hue 4.10.0, Pig 0.17.0 and Tez 0.9.2

HBase: HBase 2.4.4 with Hadoop 3.2.1, Hive 3.1.3, Hue 4.10.0, Phoenix 5.1.2, and ZooKeeper 3.5.7

Create Cluster - Advanced Options [Go to quick options](#)

Step 1: Software and Steps

Step 2: Hardware

Step 3: General Cluster Settings

Step 4: Security

Software Configuration

Release [i](#)

- | | | |
|---|--|--|
| <input checked="" type="checkbox"/> Hadoop 3.2.1 | <input type="checkbox"/> Zeppelin 0.10.0 | <input type="checkbox"/> Livy 0.7.1 |
| <input type="checkbox"/> JupyterHub 1.4.1 | <input type="checkbox"/> Tez 0.9.2 | <input type="checkbox"/> Flink 1.14.2 |
| <input type="checkbox"/> Ganglia 3.7.2 | <input type="checkbox"/> HBase 2.4.4 | <input checked="" type="checkbox"/> Pig 0.17.0 |
| <input checked="" type="checkbox"/> Hive 3.1.3 | <input type="checkbox"/> Presto 0.272 | <input type="checkbox"/> ZooKeeper 3.5.7 |
| <input type="checkbox"/> JupyterEnterpriseGateway 2.1.0 | <input type="checkbox"/> MXNet 1.8.0 | <input type="checkbox"/> Sqoop 1.4.7 |
| <input checked="" type="checkbox"/> Hue 4.10.0 | <input type="checkbox"/> Phoenix 5.1.2 | <input type="checkbox"/> Trino 378 |
| <input type="checkbox"/> Oozie 5.2.1 | <input type="checkbox"/> Spark 3.2.1 | <input type="checkbox"/> HCatalog 3.1.3 |
| <input type="checkbox"/> TensorFlow 2.4.1 | | |

Multiple master nodes (optional)

Software Configuration

Release [i](#)

- | | | |
|--|---|--|
| <input checked="" type="checkbox"/> Hadoop 3.2.1 | <input type="checkbox"/> Zeppelin 0.10.0 | <input type="checkbox"/> Livy 0.7.1 |
| <input type="checkbox"/> JupyterHub 1.4.1 | <input type="checkbox"/> Tez 0.9.2 | <input type="checkbox"/> Flink 1.14.2 |
| <input type="checkbox"/> Ganglia 3.7.2 | <input type="checkbox"/> HBase 2.4.4 | <input checked="" type="checkbox"/> Pig 0.17.0 |
| <input checked="" type="checkbox"/> Hive 3.1.3 | <input type="checkbox"/> Presto 0.272 | <input type="checkbox"/> ZooKeeper 3.5.7 |
| <input checked="" type="checkbox"/> JupyterEnterpriseGateway 2.1.0 | <input type="checkbox"/> MXNet 1.8.0 | <input type="checkbox"/> Sqoop 1.4.7 |
| <input checked="" type="checkbox"/> Hue 4.10.0 | <input type="checkbox"/> Phoenix 5.1.2 | <input type="checkbox"/> Trino 378 |
| <input type="checkbox"/> Oozie 5.2.1 | <input checked="" type="checkbox"/> Spark 3.2.1 | <input type="checkbox"/> HCatalog 3.1.3 |
| <input type="checkbox"/> TensorFlow 2.4.1 | | |

Amazon EMR

EMR Serverless is now GA.
With EMR Serverless, get the benefits of Amazon EMR such as open source compatibility, latest versions and performance optimized runtime for popular frameworks along with easy provisioning, quick job startup, automatic capacity management, and simple cost controls. [Get Started with EMR Serverless.](#)

Configuration details

- Release label: emr-6.7.0
- Hadoop distribution: Amazon 3.2.1
- Applications: Hive 3.1.3, Pig 0.17.0, Hue 4.10.0, JupyterEnterpriseGateway 2.1.0, Spark 3.2.1
- Log URI: s3://aws-logs-074831285365-us-east-1/elasticmapreduce/
- EMRFS consistent view: Disabled
- Custom AMI ID: --
- Amazon Linux Release: 2.0.20220606.1 [Learn more](#)

Network and hardware

- Availability zone: us-east-1b
- Subnet ID: subnet-04ed0369f48942ed
- Master: Provisioning 1 m5.xlarge
- Core: Provisioning 2 m5.xlarge
- Task: --
- Cluster scaling: Not enabled
- Auto-termination: Terminate if idle for 1 hour

Application user interfaces

- Persistent user interfaces: --
- On-cluster user Not Enabled [Enable an SSH Connection](#)
- interfaces: --

Security and access

- Key name: --
- EC2 instance profile: EMR_EC2_DefaultRole
- EMR role: EMR_DefaultRole
- Auto Scaling role: EMR_AutoScaling_DefaultRole
- Visible to all users: All [Change](#)
- Security groups for Master: sg-0f6468cd497c78455 (ElasticMapReduce-master)
- Security groups for Core & Task: sg-0631eeab7a22ded0 (ElasticMapReduce-Task: slave)

[us-east-1.console.aws.amazon.com/ec2/home?region=us-east-1#instances:instanceState=running](#)

[poll_form.php](#) [Canny edge detect...](#) [Simple guide on ho...](#) [A Neural Network P...](#) [Learner Lab](#) [Canny edge detect...](#) [Machine Learning u...](#) [TensorFlow Lite Co...](#)

New EC2 Experience Tell us what you think

Instances (3) Info

Find instance by attribute or tag (case-sensitive)

Instance state = running	X	Clear filters
Instance state	Running	Running
Name	i-07e31bce190575ae0	Running
Instance ID	i-0999c0a6ae2c66806	Running
Instance type	m5.xlarge	Running
Status check	2/2 checks passed	2/2 checks passed
Alarm status	No alarms	No alarms
Availability Zone	us-east-1b	us-east-1b

[us-east-1.console.aws.amazon.com/elasticmapreduce/home?region=us-east-1#cluster-details:j-30GA8FX0O1ZL1](#)

[poll_form.php](#) [Canny edge detect...](#) [Simple guide on ho...](#) [A Neural Network P...](#) [Learner Lab](#) [Canny edge detect...](#) [Machine Learning u...](#) [TensorFlow Lite Co...](#)

Amazon EMR

EMR Serverless is now GA.
With EMR Serverless, get the benefits of Amazon EMR such as open source compatibility, latest versions and performance optimized runtime for popular frameworks along with easy provisioning, quick job startup, automatic capacity management, and simple cost controls. [Get Started with EMR Serverless.](#)

Cluster: My cluster in class **Running** Running step

Summary **Application user interfaces** **Monitoring** **Hardware** **Configurations** **Events** **Steps** **Bootstrap actions**

Summary

ID: j-30GA8FX0O1ZL1
Creation date: 2022-09-24 17:09 (UTC+7)
Elapsed time: 10 minutes
After last step completes: Cluster waits
Termination protection: On [Change](#)
Tags: -- [View All / Edit](#)
Master public DNS: ec2-3-83-128-235.compute-1.amazonaws.com [Connect to the Master Node Using SSH](#)

Configuration details

- Release label: emr-6.7.0
- Hadoop distribution: Amazon 3.2.1
- Applications: Hive 3.1.3, Pig 0.17.0, Hue 4.10.0,

Application user interfaces

- Persistent user interfaces: Spark history server, YARN timeline server, Tez UI
- On-cluster user Not Enabled [Enable an SSH Connection](#)
- interfaces: --

Amazon EMR

- EMR Studio
- EMR Serverless
- EMR on EC2**
- Clusters
- Notebooks**
- Git repositories
- Security configurations
- Block public access
- VPC subnets
- Events
- EMR on EKS**
- Virtual clusters

Help

EMR Serverless is now GA.
With EMR Serverless, get the benefits of Amazon EMR such as open source compatibility, latest versions and performance optimized runtime for popular frameworks along with easy provisioning, quick job startup, automatic capacity management, and simple cost controls. [Get Started with EMR Serverless.](#)

Notebooks

Use EMR notebooks based on Jupyter to analyze data interactively with live code, narrative text, visualizations, and more. Create and attach notebooks to Amazon EMR clusters Hadoop, Spark, and Livy. Notebooks run free of charge and are saved in Amazon S3 independently of clusters. Standard billing for clusters and Amazon S3 apply. [Learn more](#)

Create notebook **View details** **Open in JupyterLab** **Open in Jupyter** **Start** **Stop** **Delete**

Filter: All notebooks 0 notebooks (all loaded)

Name	Status	Cluster	Creation time (UTC+7)	Last modified

Amazon EMR

- EMR Studio
- EMR Serverless
- EMR on EC2**
- Clusters
- Notebooks**
- Git repositories
- Security configurations
- Block public access
- VPC subnets
- Events
- EMR on EKS**
- Virtual clusters

Help

What's new

EMR Serverless is now GA.
With EMR Serverless, get the benefits of Amazon EMR such as open source compatibility, latest versions and performance optimized runtime for popular frameworks along with easy provisioning, quick job startup, automatic capacity management, and simple cost controls. [Get Started with EMR Serverless.](#)

Name your notebook, choose a cluster or create one, and customize configuration options if desired. [Learn more](#)

Notebook name* mynotebook
Names may only contain alphanumeric characters, hyphens (-), or underscores (_).

Description

256 characters max.

Cluster* Choose an existing cluster **Choose** My cluster in class [j-30GA8FX0O1ZL1](#)
 Create a cluster

Security groups Use default security groups
 Choose security groups (vpc-0644f8397501301da)

AWS service role* LabRole
Make sure this role has the required permissions. [Learn more](#)

aws **Services** [Alt+S] N. Virginia vclabs/user1591029=peeyapak.somvitoon@g.swu.ac.th @ 0748-312...

Amazon EMR

- EMR Studio
- EMR Serverless
- EMR on EC2**
- Clusters
- Notebooks**
- Git repositories
- Security configurations
- Block public access
- VPC subnets
- Events
- EMR on EKS**
- Virtual clusters

Help

What's new

EMR Serverless is now GA.
With EMR Serverless, get the benefits of Amazon EMR such as open source compatibility, latest versions and performance optimized runtime for popular frameworks along with easy provisioning, quick job startup, automatic capacity management, and simple cost controls. [Get Started with EMR Serverless.](#)

Notebook: mynotebook Starting Starting workspace(notebook). Cluster j-30GA8FX0O1ZL1.

Open in JupyterLab **Open in Jupyter** **Stop** **Delete**

Notebook

Notebook ID: e-EH9DJKCMFFK7XD7IO9FU1KF5V
Description: --
Last modified: 7 seconds ago
Last modified by: ...assumed-role/vclabs/user1591029=peeyapak.somvitoon@g.swu.ac.th
Created on: 2022-09-24 17:24 (UTC+7)
Created by: ...assumed-role/vclabs/user1591029=peeyapak.somvitoon@g.swu.ac.th
Service IAM role: LabRole
Notebook tags: creatorUserId = AROARC3CKZR23PB73NMSZ:user1591029=peeyapak.somvitoon@g.swu.ac.th [View All / Edit](#)

Notebook location: s3://aws-emr-resources-074831285365-us-east-1/notebooks/

Cluster

us-east-1.console.aws.amazon.com/elasticmapreduce/home?region=us-east-1#notebook-details:e-EH9DJKCMFFK7XD7iO9FU1KF5V

poll_form.php Canny edge detect... Simple guide on ho... A Neural Network P... Learner Lab Canny edge detect... Machine Learning u... TensorFlow Lite Co...

Search for services, features, blogs, docs, and more [Alt+S] N. Virginia v voclabs/user1591029=peeyapak.somvitoon@g.swu.ac.th @ 0748-512...

Amazon EMR

EMR Studio

EMR Serverless **New**

EMR on EC2

Clusters

Notebooks

- Git repositories
- Security configurations
- Block public access
- VPC subnets
- Events

EMR on EKS

Virtual clusters

Help

What's new

EMR Serverless is now GA.
With EMR Serverless, get the benefits of Amazon EMR such as open source compatibility, latest versions and performance optimized runtime for popular frameworks along with easy provisioning, quick job startup, automatic capacity management, and simple cost controls. [Get Started with EMR Serverless.](#)

Notebook: mynotebook Ready Workspace(notebook) is ready to run jobs on cluster j-30GA8FX0O1ZL1.

[Open in JupyterLab](#) [Open in Jupyter](#) [Stop](#) [Delete](#)

Notebook

Notebook ID: e-EH9DJKCMFFK7XD7iO9FU1KF5V
 Description: --
 Last modified: 8 seconds ago ⓘ
 Last modified by: ...assumed-role/voclabs/user1591029=peeyapak.somvitoon@g.swu.ac.th ⓘ
 Created on: 2022-09-24 17:24 (UTC+7)
 Created by: ...assumed-role/voclabs/user1591029=peeyapak.somvitoon@g.swu.ac.th ⓘ
 Service IAM role: LabRole ⓘ
 Security groups for sg-055a0ddd20ad1afae ⓘ master instance:
 Security groups for sg-0f7d11a332b983be3 ⓘ notebook instance:
 Notebook tags: creatorUserId = AROARC3CKZR23PB73NMSZ:user1591029=peeyapak.somvitoon@g.swu.ac.th [View All / Edit](#)

เรียก spark ขึ้นมา

e-eh9djkcffffk7xd7io9fu1kf5v.emrnotebooks-prod.us-east-1.amazonaws.com/e-EH9DJKCMFFK7XD7iO9FU1KF5V/lab

poll_form.php Canny edge detect... Simple guide on ho... A Neural Network P... Learner Lab Canny edge detect... Machine Learning u... TensorFlow Lite Co...

File Edit View Run Kernel Git Tabs Settings Help

Launcher

Filter files by name

Name / Last Modified

mynotebook.ipynb a minute ago

Notebook

- Python 3
- PySpark
- Spark
- Spark

Console

- Python 3
- PySpark
- Spark
- SparkR

Other

jupyter

Files Running Clusters

Select items to perform actions on them.

Upload New

Name	Last Modified	File size
0	2 minutes ago	72 B
/		
mynotebook.ipynb		

The screenshot shows a Jupyter Notebook interface with the following details:

- Header:** Shows the URL e-h9djkcmffk7xd7io9fu1kf5v.emrnotebooks-prod.us-east-1.amazonaws.com/e-EH9DJKCMFFK7XD7IO9FU1KF5V/tree?username=user1591029=peey... and various browser tabs.
- Title Bar:** Displays the Jupyter logo and a "quit" button.
- Navigation:** Buttons for "Files", "Running", and "Clusters".
- Section:** "Currently running Jupyter processes".
- Terminals:** A section stating "There are no terminals running."
- Notebooks:** A section showing a single notebook named "mynotebook.ipynb".
- Details:** Python 3, Shutdown, and a timestamp indicating it was shutdown "seconds ago".

The screenshot shows a Jupyter Notebook interface with the following details:

- Header:** Shows the URL e-h9djkcmffk7xd7io9fu1kf5v.emrnotebooks-prod.us-east-1.amazonaws.com/e-EH9DJKCMFFK7XD7IO9FU1KF5V/tree?username=user1591029=peey... and various browser tabs.
- Title Bar:** Displays the Jupyter logo and a "quit" button.
- Navigation:** Buttons for "Files", "Running", and "Clusters".
- Section:** "Currently running Jupyter processes".
- Terminals:** A section stating "There are no terminals running."
- Notebooks:** A section showing a single notebook named "mynotebook.ipynb".
- Details:** Python 3, Shutdown, and a timestamp indicating it was shutdown "seconds ago".

The screenshot shows the AWS IAM LabRole configuration page with the following details:

- Header:** Shows the AWS logo, Services, and a search bar for "Search for services, features, blogs, docs, and more".
- Left Sidebar:** Includes sections for Identity and Access Management (IAM), Access management (User groups, Users, Roles, Policies, Identity providers, Account settings), Access reports (Access analyzer, Archive rules, Analyzers, Settings, Credential report), and a global status bar.
- Notification:** A blue banner at the top right says "Introducing the new IAM roles experience" and "We've redesigned the IAM roles experience to make it easier to use. Let us know what you think." with a "Delete" and "Edit" button.
- Role Details:** The "LabRole" is selected under "Roles > Roles".
- Summary:** Shows creation date (September 03, 2022, 17:54 (UTC+07:00)), ARN (arn:aws:iam::074831285365:role/LabRole), instance profile ARN (arn:aws:iam::074831285365:instance-profile/LabRole), last activity (24 minutes ago), maximum session duration (1 hour), and a "Delete" and "Edit" button.
- Permissions:** Tabbed section showing "Permissions", "Trust relationships", "Tags (1)", "Access Advisor", and "Revoke sessions".
- Permissions policies:** Shows 7 managed policies attached to the role.
- Actions:** Buttons for "Simulate", "Remove", and "Add permissions".

← → C e-eh9djkcmffk7xd7io9fu1kf5v.emrnotebooks-prod.us-east-1.amazonaws.com/e-EH9DJKCMFFK7XD7IO9FU1KF5V/lab

poll.form.php Canny edge detect... Simple guide on ho... A Neural Network P... Learner Lab Canny edge detect... Machine Learning u... TensorFlow Lite Co...

File Edit View Run Kernel Git Tabs Settings Help

Launcher

Notebook

Python 3 PySpark Spark SparkR

Console

Python 3 PySpark Spark SparkR

Other

jupyter Untitled Last Checkpoint: 3 minutes ago (unsaved changes)

File Edit View Insert Cell Kernel Widgets Help Trusted | PySpark

In [1]: `from pyspark.sql import SparkSession`

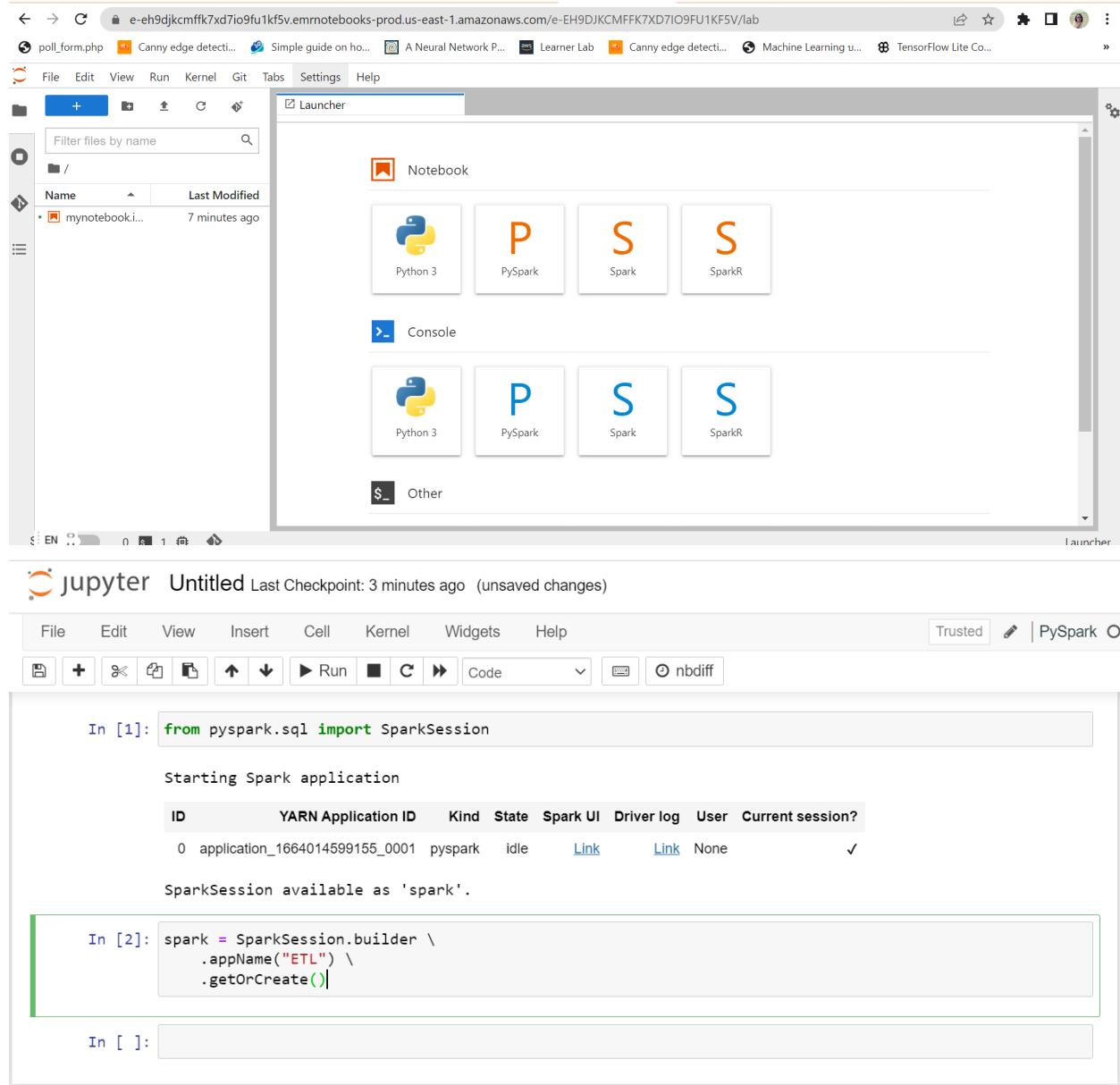
Starting Spark application

ID	YARN Application ID	Kind	State	Spark UI	Driver log	User	Current session?
0	application_1664014599155_0001	pyspark	idle	Link	Link	None	✓

SparkSession available as 'spark'.

In [2]: `spark = SparkSession.builder \
 .appName("ETL") \
 .getOrCreate()`

In []:

The image shows a Jupyter Notebook interface running on an Amazon EMR cluster. At the top, there's a browser-like header with a URL and several tabs. Below it is a 'Launcher' window containing icons for different kernels: Python 3, PySpark, Spark, and SparkR. The main area is a Jupyter notebook titled 'Untitled'. It has a toolbar with various icons. The first cell (In [1]) contains the command 'from pyspark.sql import SparkSession' and outputs the message 'Starting Spark application' followed by a table of application details. The second cell (In [2]) contains the code to build a SparkSession. The bottom cell (In []) is currently empty.

s3.console.aws.amazon.com/s3/upload/kiktitanic?region=us-east-1

aws Services Search for services, features, blogs, docs, and more [Alt+S]

Upload succeeded View details below.

Upload: status

The information below will no longer be available after you navigate away from this page.

Summary

Destination	Succeeded	Failed
s3://kiktitanic	1 file, 59.8 KB (100.00%)	0 files, 0 B (0%)

Files and folders Configuration

Files and folders (1 Total, 59.8 KB)

Name	Type	Size	Status	Error
titanic.csv	text/csv	59.8 KB	Succeeded	-

jupyter Untitled Last Checkpoint: 19 minutes ago (autosaved)

File Edit View Insert Cell Kernel Widgets Help Trusted | PySpark O

In [1]: `from pyspark.sql import SparkSession`

Starting Spark application

ID	YARN Application ID	Kind	State	Spark UI	Driver log	User	Current session?
0	application_1664014599155_0001	pyspark	idle	Link	Link	None	✓

SparkSession available as 'spark'.

In [2]: `spark = SparkSession.builder \
 .appName("ETL") \
 .getOrCreate()`

In [3]: `bucket = "s3://kiktitanic"`

In [5]: `df = spark.read.csv(bucket)`

jupyter Untitled Last Checkpoint: 20 minutes ago (autosaved)

File Edit View Insert Cell Kernel Widgets Help Trusted PySpark

In [3]: `bucket = "s3://kiktitanic"`

In [5]: `df = spark.read.csv(bucket)`

▼ Spark Job Progress

▼ Job [0]: csv at NativeMethodAccessorImpl.java:0

Progress for csv at NativeMethodAccessorImpl.j... Job Progress: 1/1 Tasks... 

Stage [ID]: name at [source]	Status	Task Progress	Elapsed Time (s...)	Failed Task...
Stage [0]: csv at NativeMet...	COMP...	1/1 	5.601	

In [7]: `df = spark.read.option('header', 'true').csv(bucket)`

In [7]: `df = spark.read.option('header', 'true').csv(bucket)`

► Spark Job Progress

In [8]: `df.show()`

► Spark Job Progress

```
+-----+-----+-----+-----+-----+-----+-----+-----+
|PassengerId|Survived|Pclass|          Name|  Sex|  Age|SibSp|Parch|      Ticket|  Fare|Ca
|bin|Embarked|
+-----+-----+-----+-----+-----+-----+-----+-----+
|       1|     0|     3|Braund, Mr. Owen ...| male|  22|    1|    0| A/5 21171|  7.25| n
|       1|     1|     1|Cumings, Mrs. Joh...|female|  38|    1|    0| PC 17599|71.2833|
|       2|     1|     1|          ...|       |       |       |       |       |       |
+-----+-----+-----+-----+-----+-----+-----+-----+
```

jupyter Untitled (unsaved changes)

File Edit View Insert Cell Kernel Widgets Help Trusted | PySpark

only showing top 20 rows

In [9]: `df.select('Age', 'Survived')`

DataFrame[Age: string, Survived: string]

In [10]: `df.select('Age', 'Survived').show()`

▶ Spark Job Progress

Age	Survived
22	0
38	1
26	1
35	1
35	0
null	0
54	0
2	0
27	1
..	..

```
In [12]: df.createOrReplaceTempView("titanic")
```

```
In [14]: spark.sql("""  
    select  
        Age  
        , Survived  
    from  
        titanic  
""").show()
```

▶ Spark Job Progress

Age	Survived
22	0
38	1
26	1
35	1
35	0
null	0
54	0

jupyter Untitled Last Checkpoint: 32 minutes ago (unsaved changes)

File Edit View Insert Cell Kernel Widgets Help Trusted PySpark

2	0
null	1
31	0
null	1
+---+---+
only showing top 20 rows

```
In [15]: result = spark.sql("""
    select
        Age
        , Survived
    from
        titanic
""")
```

```
In [16]: result.show(3)
```

▶ Spark Job Progress

```
+---+---+
|Age|Survived|
+---+---+
| 22 | 0 |
| 38 | 1 |
| 26 | 1 |
+---+---+
only showing top 3 rows
```

```
In [17]: result.write.mode("overwrite").csv("s3://kiktitanic/cleaned")
```

▶ Spark Job Progress

aws Services Search for services, features, blogs, docs, and more [Alt+S] Global vodlabs/user1591029=peeyapak.somvitoon@g.swu.ac.th @ 0748-312...

Amazon S3

Amazon S3 > Buckets > kiktitanic > cleaned/

Cleaned/

Objects (2)

Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory [to get a list of all objects in your bucket.](#) For others to access your objects, you'll need to explicitly grant them permissions. [Learn more](#)

Actions

Upload

Find objects by prefix

Name	Type	Last modified	Size	Storage class
_SUCCESS	-	September 24, 2022, 18:13:28 (UTC+07:00)	0 B	Standard
part-00000-7815f5c1-f264-4755-9ea6-dbe4f450a12cf-c000.csv	csv	September 24, 2022, 18:13:28 (UTC+07:00)	4.3 KB	Standard

AWS Marketplace for S3

jupyter Untitled Last Checkpoint: 37 minutes ago (unsaved changes)

File Edit View Insert Cell Kernel Widgets Help

titanic

In [16]: `result.show(3)`

▶ Spark Job Progress

```
+---+-----+
|Age|Survived|
+---+-----+
| 22|      0|
| 38|      1|
| 26|      1|
+---+-----+
only showing top 3 rows
```

In [17]: `result.write.mode("overwrite").csv("s3://kiktitanic/cleaned")`

▶ Spark Job Progress

In [18]: `result.write.mode("overwrite").parquet("s3://kiktitanic/cleaned-parquet")`

▶ Spark Job Progress

In []:

s3.console.aws.amazon.com/s3/buckets/kiktitanic?region=us-east-1&tab=objects

poll_form.php Canny edge detect... Simple guide on ho... A Neural Network P... Learner Lab Canny edge detect... Machine Learning u... TensorFlow Lite Co...

aws Services Search for services, features, blogs, docs, and more [Alt+S]

Amazon S3 Buckets kiktitanic

kiktitanic info

Objects (3) Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more

Actions Copy S3 URI Copy URL Download Open Delete Create folder Upload

Find objects by prefix

Name	Type	Last modified	Size	Storage class
cleaned-parquet/	Folder	-	-	-
cleaned/	Folder	-	-	-
titanic.csv	csv	September 24, 2022, 17:49:40 (UTC+07:00)	59.8 KB	Standard

AWS Marketplace for S3

Course: | DS525: | swu-ds: | DS525: | DS525: | Gitpod | dh Sou x +

← → C dunnhumby.com/source-files/ 🔍 ⭐ 🧩 📁 🌐 ⋮

poll_form.php Canny edge detecti... Simple guide on ho... A Neural Network P... aws Learner Lab »

dunnhumby

- PROMOTIONAL EFFECTIVENESS ANALYSIS
- Comparing/contrasting results across products, categories, store groupings, or geographies

How should I use it?

Check back soon for example exercises, case studies, and other helpful info from our professor partners at Notre Dame.

[Download 'Breakfast at the Frat' →](#)



Professor and Department Chair, Department of Applied and Computational Mathematics and Statistics, University of Notre Dame

Timothy Gilbride

inavscript: 

EXPLORER workshop.ipynb U product-lookup-table.csv U store-lookup-table.csv U transaction-data-table.csv U

OPEN EDITORS workshop.ipynb 04-building-a-data-lake... U product-lookup-table.csv 04-building-a... U store-lookup-table.csv 04-building-a-dat... U transaction-data-table.csv 04-building-a... U

SWU-DS525 stores store-lookup-table.csv transactions transaction-data-table.csv docker-compose.yml workshop.ipynb

OUTLINE No symbols found in document 'transaction-data-table.csv'

TIMELINE transaction-data-table.csv No timeline information was provided.

workshop-pyspark-notebook-1 | [2022-10-01 10:35:29.260 ServerApp] Use Control-C to stop this server and shut down all kernels (twice to skip confirmation).

workshop-pyspark-notebook-1 | [C 2022-10-01 10:35:29.264 ServerApp]

workshop-pyspark-notebook-1 | To access the server, open this file in a browser:

workshop-pyspark-notebook-1 | file:///home/jovyan/.local/share/jupyter/runtime/jpserver-25-open.html

workshop-pyspark-notebook-1 | Or copy and paste one of these URLs:

workshop-pyspark-notebook-1 | http://c4749dc8605d:8888/lab?token=60fcf77228bcebae61f235de501111b1685e774cc1d5134

workshop-pyspark-notebook-1 | or http://127.0.0.1:8888/lab?token=60fcf77228bcebae61f235de501111b1685e774cc1d5134

workshop-pyspark-notebook-1 | [I 2022-10-01 10:41:32.420 ServerApp] 302 GET / (192.168.48.138) 0.61 ms

workshop-pyspark-notebook-1 | [I 2022-10-01 10:41:32.624 LabApp] 302 GET /lab? (192.168.48.138) 0.7 2ms

workshop-pyspark-notebook-1 | [I 2022-10-01 10:42:05.615 ServerApp] 302 POST /login?next=%2Flab%3F(192.168.48.138) 1.19ms

workshop-pyspark-notebook-1 | [I 2022-10-01 10:42:11.971 LabApp] Build is up to date

Gitpod main* Share

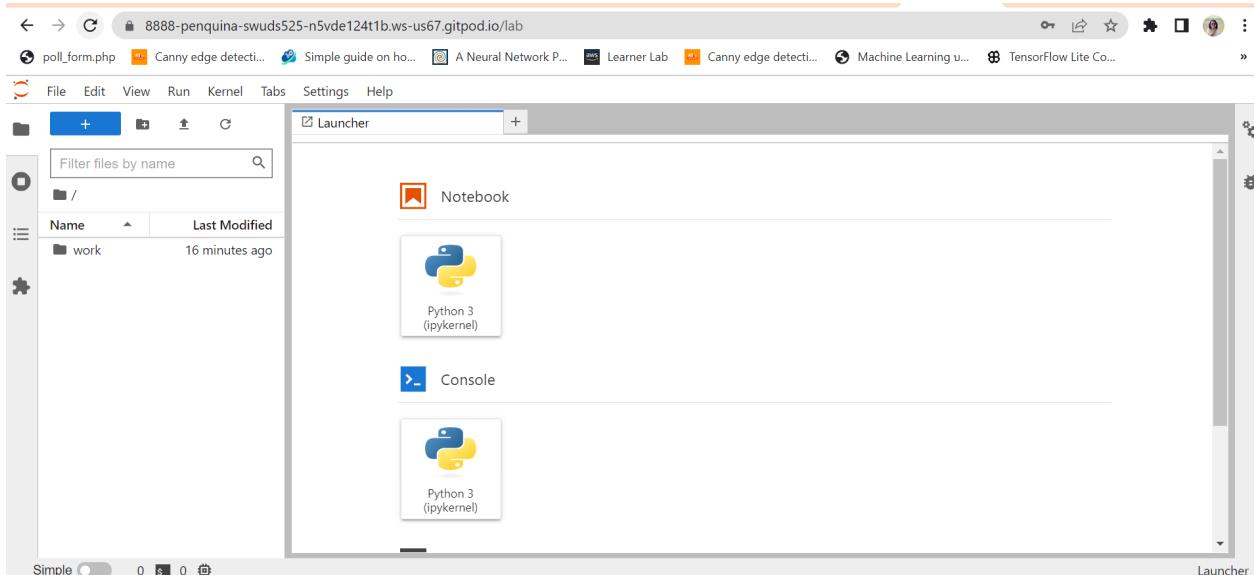
swu-ds525-main.zip Lab3_Lab.html Example_Lab (1).jpg Show all

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

```
token=60fcfc77228bcebae61f235de501111b1685e774ccc1d5134
workshop-pyspark-notebook-1 | [I 2022-10-01 10:35:29.260 ServerApp] Use Control-C to stop this server and shut down all kernels (twice to skip confirmation).
workshop-pyspark-notebook-1 | [C 2022-10-01 10:35:29.264 ServerApp]
workshop-pyspark-notebook-1 |
workshop-pyspark-notebook-1 | To access the server, open this file in a browser:
workshop-pyspark-notebook-1 |   file:///home/jovyan/.local/share/jupyter/runtime/jpserver-25-
open.html
workshop-pyspark-notebook-1 |   Or copy and paste one of these URLs:
workshop-pyspark-notebook-1 |     http://c4749dc8605d:8888/lab?token=60fcfc77228bcebae61f235de501111b1685e774ccc1d5134
workshop-pyspark-notebook-1 |   or http://127.0.0.1:8888/lab?token=60fcfc77228bcebae61f235de501111b1685e774ccc1d5134
workshop-pyspark-notebook-1 | [I 2022-10-01 10:41:32.420 ServerApp] 302 GET / (192.168.48.138) 0.61 ms
workshop-pyspark-notebook-1 | [I 2022-10-01 10:41:32.624 LabApp] 302 GET /lab? (192.168.48.138) 0.72ms
workshop-pyspark-notebook-1 | [I 2022-10-01 10:42:05.615 ServerApp] 302 POST /login?next=%2Flab%3F(192.168.48.138) 1.19ms
workshop-pyspark-notebook-1 | [I 2022-10-01 10:42:11.971 LabApp] Build is up to date
```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

Port	Address
3000	https://3000-penguinawswuds525-n5vde124t1b.ws-us67.gitpod.io
4040	https://4040-penguinawswuds525-n5vde124t1b.ws-us67.gitpod.io
4041	https://4041-penguinawswuds525-n5vde124t1b.ws-us67.gitpod.io
8888	https://8888-penguinawswuds525-n5vde124t1b.ws-us67.gitpod.io



← → ⌂ 8888-penguina-swuds525-n5vde124t1b.ws-us67.gitpod.io/lab/tree/work/workshop.ipynb

poll_form.php Canny edge detect... Simple guide on ho... A Neural Network P... Learner Lab Canny edge detect... Machine Learning u... TensorFlow Lite Co...

File Edit View Run Kernel Tabs Settings Help

Launcher workshop.ipynb Python 3 (ipykernel)

Filter files by name / work /

Name Last Modified

- dataset 18 minutes ago
- docker-compose.yml 19 minutes ago
- workshop.ipynb** 19 minutes ago

To identify outliers, it is suggested to look at

- The ratio of units vs. number of visits
- The ratio of visits vs. number of households
- Some items that may be out-of-stock or discontinued for a store

Source: <https://www.dunnhumby.com/source-files/>

```
[ ]: from pyspark.sql import Row, SparkSession
[ ]: spark = SparkSession.builder \
    .appName("breakfast") \
    .getOrCreate()
[ ]: product_data_folder = "dataset/products"
store_data_folder = "dataset/stores"
transaction_data_folder = "dataset/transactions"
```

1. What is the range of prices offered on products?

2. What is the impact on units/visit of promotions by geographies?

3. Which products would you lower the price to increase sales?

```
[4]: product_df = spark\
    .read\
    .option("header", True).\
    csv(product_data_folder)

[5]: product_df.show(1)
```

UPC	DESCRIPTION	MANUFACTURER	CATEGORY	SUB_CATEGORY	PRODUCT_SIZE
1111009477	PL MINI TWIST PRE...	PRIVATE LABEL	BAG SNACKS	PRETZELS	15 OZ

only showing top 1 row

[]:

Launcher × workshop.ipynb ● + Python 3 (ipykernel) ○

only showing top 20 rows

```
[12]: store_df.show(1)
```

	STORE_ID	STORE_NAME	ADDRESS_CITY_NAME	ADDRESS_STATE_PROV_CODE	MSA_CODE	SEG_VALUE_NAME	PARKING_SPACE_QTY	SALES_AREA_SIZE_NUM	AVG_WEEKLY_BASKETS	
408	389	SILVERLAKE	ERLANGER	KY	17140	MAINSTREAM		46073	24767	

only showing top 1 row

```
product_df.createOrReplaceTempView("products")
transaction_df.createOrReplaceTempView("transactions")
```



```
#Range of price
spark.sql("""
    select
        products.upc
        , min(price)
        , max(price)
        , description
        , category

    from transactions
    join products
    on
        transactions.upc = products.upc
    group by
        1, 4, 5
""").show(3)
```

upc	min(price)	max(price)	description	category
1111009477	0.89	1.83	PL MINI TWIST PRE...	BAG SNACKS
1111009497	0.86	1.69	PL PRETZEL STICKS	BAG SNACKS
1111009507	0.8	1.69	PL TWIST PRETZELS	BAG SNACKS

only showing top 3 rows

```
[27]: spark.sql("""
    select
        upc
        , product_size
        , case
            when contains(product_size, 'OZ') then 'yes'
        end as is_oz

    from products
""").show(5)
```

	upc	product_size	is_oz
1	1111009477	15 OZ	yes
2	1111009497	15 OZ	yes
3	1111009507	15 OZ	yes
4	1111035398	1.5 LT	null
5	1111038078	500 ML	null

only showing top 5 rows

The screenshot shows a Jupyter Notebook environment with several tabs at the top: poll_form.php, Canny edge detect..., Simple guide on ho..., A Neural Network P..., Learner Lab, Canny edge detect..., Machine Learning u..., TensorFlow Lite Co... . Below the tabs, there are three tabs for datasets: workshop.ipynb U, product-lookup-table.csv U, store-lookup-table.csv U, and transaction-data-table.csv U.

The main area displays a terminal session:

```
gitpod /workspace/swu-ds525 (main) $ c
d 04-building-a-data-lake/
gitpod /workspace/swu-ds525/04-building-a-data-lake (main) $ ls -ltr
total 0
drwxr-xr-x 3 gitpod gitpod 69 Oct 1 10:30 workshop
gitpod /workspace/swu-ds525/04-building-a-data-lake (main) $ cd workshop/
total 8
-rw-r--r-- 1 gitpod gitpod 3083 Oct 1 10:30 workshop.ipynb
-rw-r--r-- 1 gitpod gitpod 188 Oct 1 10:30 docker-compose.yml
drwxr-xr-x 5 gitpod gitpod 56 Oct 1 10:31 dataset
gitpod /workspace/swu-ds525/04-building-a-data-lake/workshop (main) $ docker-compose up
[+] Running 1/0
  => Container workshop-pyspark-notebook-1 Recreated 0.0s
Attaching to workshop-pyspark-notebook-1
workshop-pyspark-notebook-1 | Entered start.sh with args: jupyter lab
workshop-pyspark-notebook-1 | /usr/local/bin/start.sh: running hooks in /usr/local/bin/before-notebook.d as uid / gid: 1000 / 100
workshop-pyspark-notebook-1 | /usr/local/bin/start.sh: running script /usr/local/bin/before-notebook.d/spark-config.sh
workshop-pyspark-notebook-1 | /usr/local/bin/start.sh: done running hooks in /usr/local/bin/before-notebook.d
workshop-pyspark-notebook-1 | Executing the command: jupyter lab
workshop-pyspark-notebook-1 | [I:2022-10-01 11:26:38.380 ServerApp] jupyterlab | extension was successfully linked.
workshop-pyspark-notebook-1 | [I:2022-10-01 11:26:38.390 ServerApp] nbclassic | extension was successfully linked.
```

On the right side, there are two panes: "Gitpod Task ..." and "bash". The bottom status bar shows: Ln 1, Col 1 Spaces: 4 UTF-8 CRLF Plain Text Layout: US No open ports.