

04-building-a-data-lake

Actors

```
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)
```

Filter files by name

/ actors /

Name	Last Modified
_SUCCESS	seconds ago
part-00000-3df535d...	seconds ago
part-00001-3df535d...	seconds ago

etl_local.ipynb

part-00000-3df535d5-33f8-4 X

workshc

Delimiter: ,

	evilgaoshu	23487963576	https://api.github.com/...
1	gurram47	23487963624	https://api.github.com/...
2	afbeltranr	23487963529	https://api.github.com/...
3	karla-vm	23487963558	https://api.github.com/...
4	hsluoyz	23487963581	https://api.github.com/...
5	mnw1020	23487963532	https://api.github.com/...
6	ikjo93	23487963524	https://api.github.com/...
7	Gabe616	23487963526	https://api.github.com/...
8	BadProfessor	23487963492	https://api.github.com/...
9	allanrg4	23487963504	https://api.github.com/...
10	QGarchery	23487963536	https://api.github.com/...
11	Diyouf	23487963495	https://api.github.com/...
12	tiltingpenguin	23487963522	https://api.github.com/...
13	igrek-ovs	23487963444	https://api.github.com/...
14	cchanmi	23487963462	https://api.github.com/...
15	mvashishtha	23487963480	https://api.github.com/...
16	na4zagin3	23487963457	https://api.github.com/...
17	xsidc	23487963413	https://api.github.com/...

Repos

```
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)
```

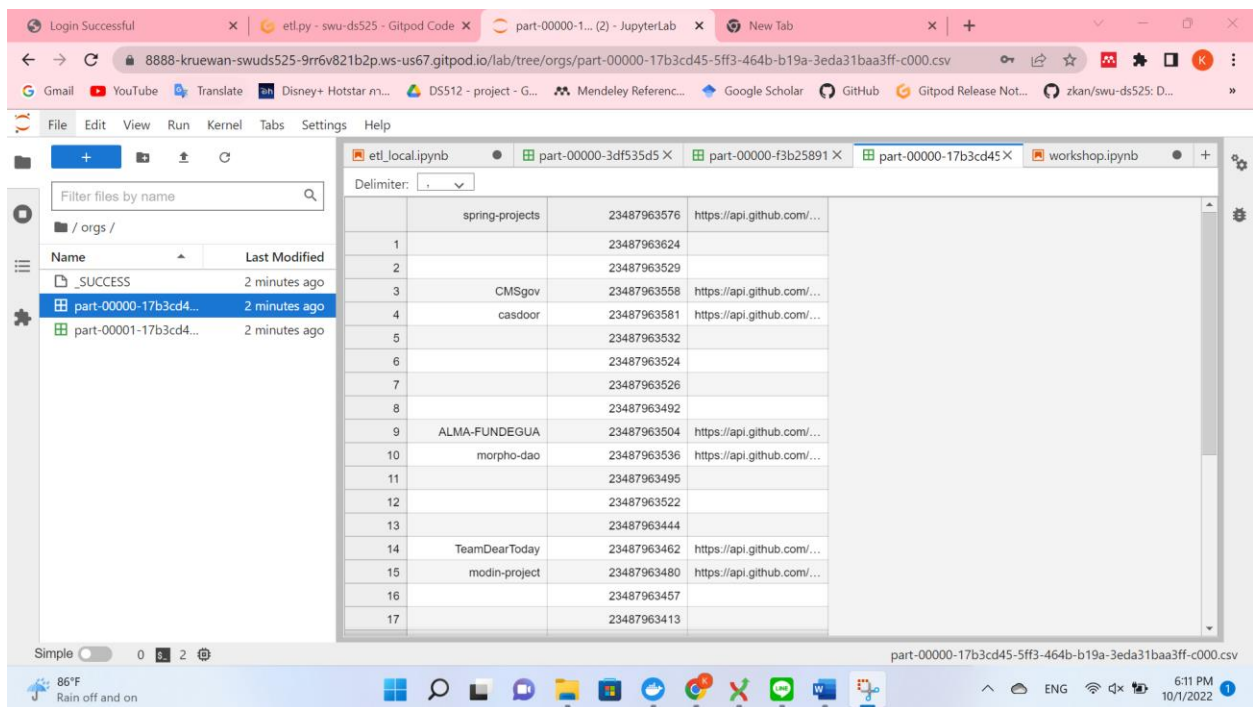
The screenshot shows a JupyterLab environment. The top bar includes tabs for 'Login Successful', 'etl.py - swu-ds525 - Gitpod Code', 'part-00000-4... (2) - JupyterLab', and 'New Tab'. The address bar shows the URL '8888-kruewan-swuds525-9rr6v821b2p.ws-us67.gitpod.io/lab/tree/repos/part-00000-4c3f1cee-0c45-4736-ae7b-f3f5aa21bbba-c000.csv'. The left sidebar contains a file browser for the '/repos/' directory, listing files like '_SUCCESS' and 'part-00000-4c3f1cee...' with their last modified times. The main area displays a data table with columns for repository names, IDs, and URLs. The table has 17 rows of data. The bottom status bar shows the temperature as 86°F, rain status, and the time as 6:12 PM on 10/1/2022.

	Repo Name	ID	URL
	spring-projects/spring...	23487963576	https://api.github.com/...
1	gurram47/AP2011001...	23487963624	https://api.github.com/...
2	afbeltranr/Agrilab2	23487963529	https://api.github.com/...
3	CMSgov/cms-carts-s...	23487963558	https://api.github.com/...
4	casdoor/casdoor-chro...	23487963581	https://api.github.com/...
5	mnw1020/obsidian	23487963532	https://api.github.com/...
6	ikjo93/Data-Structure	23487963524	https://api.github.com/...
7	Gabe616/ObbyCreato...	23487963526	https://api.github.com/...
8	BadProfessor/INL	23487963492	https://api.github.com/...
9	ALMA-FUNDEGUA/v...	23487963504	https://api.github.com/...
10	morpho-dao/morpho...	23487963536	https://api.github.com/...
11	Diyouf/newpage.githu...	23487963495	https://api.github.com/...
12	tiltingpenguin/uyuni	23487963522	https://api.github.com/...
13	igrek-ovs/igrek-ovs.git...	23487963444	https://api.github.com/...
14	TeamDearToday/Dear...	23487963462	https://api.github.com/...
15	modin-project/modin	23487963480	https://api.github.com/...
16	na4zagin3/satyrograp...	23487963457	https://api.github.com/...
17	xsaid/monmiaio	23487963413	https://api.github.com/...

Orgs

```
table = spark.sql("""
    select
        org.login
        , id as event_id
        , org.url as org_url

    from
        staging_events
""")
destination = "../orgs"
table.write.mode("overwrite").csv(destination)
```



The screenshot shows a JupyterLab interface with a file browser on the left and a data table in the center. The file browser shows a directory named 'orgs' with files like '_SUCCESS', 'part-00000-17b3cd4...', and 'part-00001-17b3cd4...'. The data table has columns for row number, org name, ID, and URL.

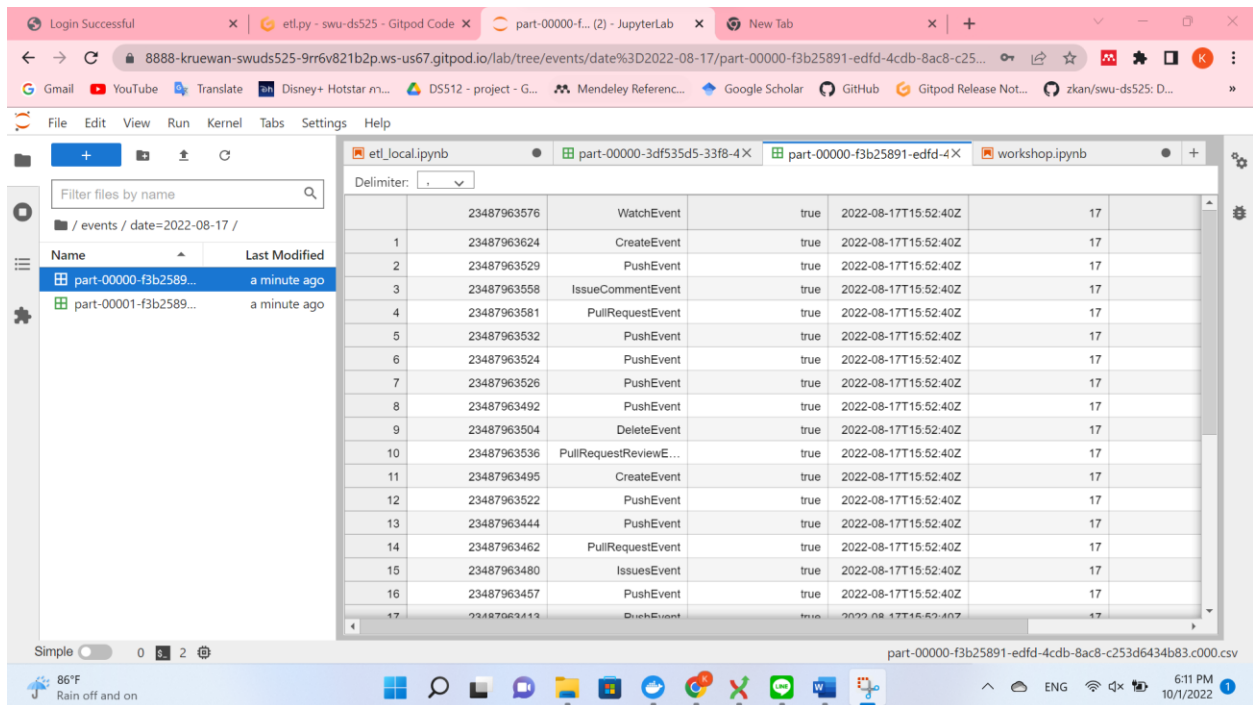
	org.login	id	org.url
1	spring-projects	23487963576	https://api.github.com/...
2		23487963624	
3	CMSgov	23487963529	https://api.github.com/...
4	casdoor	23487963558	https://api.github.com/...
5		23487963581	
6		23487963532	
7		23487963524	
8		23487963526	
9	ALMA-FUNDEGUA	23487963492	https://api.github.com/...
10	morpho-dao	23487963504	https://api.github.com/...
11		23487963536	
12		23487963495	
13		23487963522	
14	TeamDearToday	23487963444	https://api.github.com/...
15	modin-project	23487963462	https://api.github.com/...
16		23487963480	
17		23487963457	
18		23487963413	

Events

```
table = spark.sql("""
select
    id
  , type
  , public
  , created_at
  , to_date(created_at) as date
  , day(created_at) as day
  , month(created_at) as month
  , year(created_at) as year

from
    staging_events

""")
destination = "../events"
table.write.partitionBy("date").mode("overwrite").csv(destination)
```



The screenshot shows a JupyterLab interface with a file explorer on the left and a data table in the center. The file explorer shows a directory 'events' with a subdirectory 'date=2022-08-17'. The data table has columns for event ID, type, public status, date, day, month, and year. The table contains 17 rows of event data.

	id	type	public	date	day	month	year
1	23487963576	WatchEvent	true	2022-08-17T15:52:40Z	17		
2	23487963624	CreateEvent	true	2022-08-17T15:52:40Z	17		
3	23487963529	PushEvent	true	2022-08-17T15:52:40Z	17		
4	23487963558	IssueCommentEvent	true	2022-08-17T15:52:40Z	17		
5	23487963581	PullRequestEvent	true	2022-08-17T15:52:40Z	17		
6	23487963532	PushEvent	true	2022-08-17T15:52:40Z	17		
7	23487963524	PushEvent	true	2022-08-17T15:52:40Z	17		
8	23487963526	PushEvent	true	2022-08-17T15:52:40Z	17		
9	23487963492	PushEvent	true	2022-08-17T15:52:40Z	17		
10	23487963504	DeleteEvent	true	2022-08-17T15:52:40Z	17		
11	23487963536	PullRequestReviewEvent	true	2022-08-17T15:52:40Z	17		
12	23487963495	CreateEvent	true	2022-08-17T15:52:40Z	17		
13	23487963522	PushEvent	true	2022-08-17T15:52:40Z	17		
14	23487963444	PushEvent	true	2022-08-17T15:52:40Z	17		
15	23487963462	PullRequestEvent	true	2022-08-17T15:52:40Z	17		
16	23487963480	IssuesEvent	true	2022-08-17T15:52:40Z	17		
17	23487963457	PushEvent	true	2022-08-17T15:52:40Z	17		