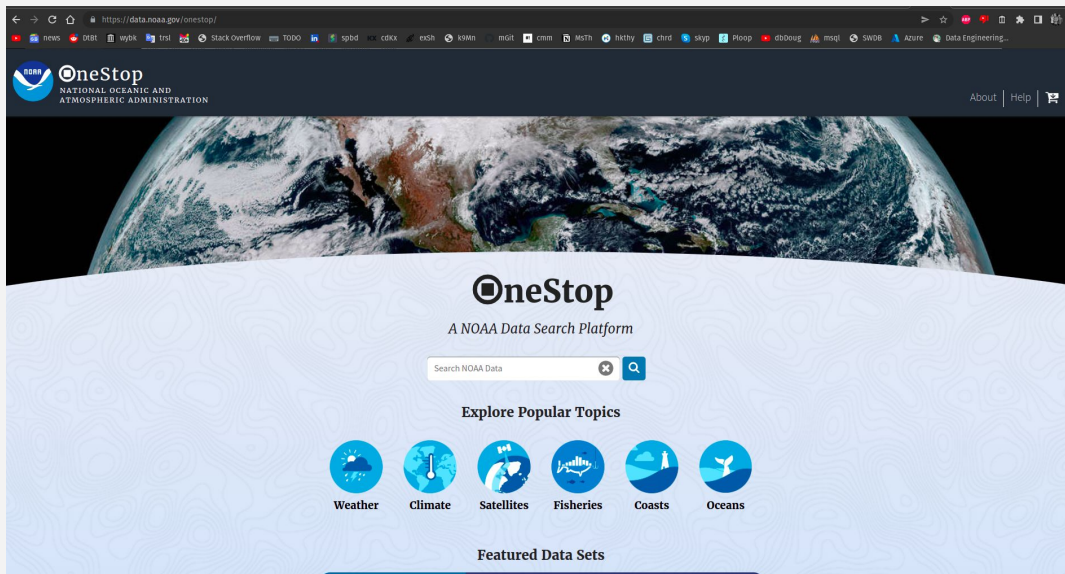


Final Submission

Capstone

Background



Index of /pub/data/swdi/stormevents/csvfiles

Name	Last modified	Size	Description
Parent Directory	-		
Storm-Data-Bulk-csv-Format.pdf	2020-07-17 13:10	161K	
Storm-Data-Export-Format.pdf	2020-07-17 09:17	163K	
StormEvents_details-ftp_v1.0_d1950_c20210803.csv.gz	2021-08-05 09:53	10K	
StormEvents_details-ftp_v1.0_d1951_c20210803.csv.gz	2021-08-05 09:56	12K	
StormEvents_details-ftp_v1.0_d1952_c20210803.csv.gz	2021-08-05 09:56	12K	
StormEvents_details-ftp_v1.0_d1953_c20210803.csv.gz	2021-08-05 09:56	21K	
StormEvents_details-ftp_v1.0_d1954_c20210803.csv.gz	2021-08-05 09:56	26K	
StormEvents_details-ftp_v1.0_d1955_c20210803.csv.gz	2021-08-05 09:56	52K	
StormEvents_details-ftp_v1.0_d1956_c20210803.csv.gz	2021-08-05 09:56	62K	
StormEvents_details-ftp_v1.0_d1957_c20210803.csv.gz	2021-08-05 09:56	80K	
StormEvents_details-ftp_v1.0_d1958_c20210803.csv.gz	2021-08-05 09:56	69K	
StormEvents_details-ftp_v1.0_d1959_c20210803.csv.gz	2021-08-05 09:56	66K	
StormEvents_details-ftp_v1.0_d1960_c20210803.csv.gz	2021-08-05 09:56	70K	
StormEvents_details-ftp_v1.0_d1961_c20210803.csv.gz	2021-08-05 09:56	81K	
StormEvents_details-ftp_v1.0_d1962_c20210803.csv.gz	2021-08-05 09:56	83K	
StormEvents_details-ftp_v1.0_d1963_c20210803.csv.gz	2021-08-05 09:56	70K	
StormEvents_details-ftp_v1.0_d1964_c20210803.csv.gz	2021-08-05 09:56	84K	
StormEvents_details-ftp_v1.0_d1965_c20210803.csv.gz	2021-08-05 09:56	102K	
StormEvents_details-ftp_v1.0_d1966_c20210803.csv.gz	2021-08-05 09:56	81K	
StormEvents_details-ftp_v1.0_d1967_c20210803.csv.gz	2021-08-05 09:56	95K	
StormEvents_details-ftp_v1.0_d1968_c20210803.csv.gz	2021-08-05 09:56	112K	
StormEvents_details-ftp_v1.0_d1969_c20210803.csv.gz	2021-08-05 09:56	100K	
StormEvents_details-ftp_v1.0_d1970_c20210803.csv.gz	2021-08-05 09:56	112K	
StormEvents_details-ftp_v1.0_d1971_c20210803.csv.gz	2021-08-05 09:56	123K	
StormEvents_details-ftp_v1.0_d1972_c20220425.csv.gz	2022-04-25 15:06	80K	
StormEvents_details-ftp_v1.0_d1973_c20220425.csv.gz	2022-04-25 15:06	157K	
StormEvents_details-ftp_v1.0_d1974_c20220425.csv.gz	2022-04-25 15:06	183K	
StormEvents_details-ftp_v1.0_d1975_c20220425.csv.gz	2022-04-25 15:06	172K	
StormEvents_details-ftp_v1.0_d1976_c20220425.csv.gz	2022-04-25 15:06	133K	

Alternative Stacks Explored

Initially explored data in q/KDB+ (instead of MySQL).

Speed using my laptop was orders of magnitude faster (~20 secs to copy full dataset from gzip'd files into tables, ½ that being decompression)

unfortunately, licensing costs squanders commercial feasibility for all but the most lucrative ventures

```
//DECOMPRESS FILES
t2::z.p
system "gzip -kd gzipped/**"
t2::z.p
t2::t2l-t2o

//INGEST DETAILS
t0::z.p
details:asc hsym each '$' ~/.home/conner/testgz/gzipped/,/: system "ls gzipped | grep -v gz | grep details"
details: (./) {(51#"";enlist ",") 0: x) each defiles
t1::z.p

//CREATE DETAILS SUBSET TABLE AND CAST COLUMN TYPES
det::select BEGIN_YEARMONTH,BEGIN_DAY,END_YEARMONTH,END_DAY,"I"$EPISODE_ID,"I"$EVENT_ID,EVENT_TYPE,
"$STATE,"I"$INJURIES_DIRECT,"I"$INJURIES_INDIRECT,"I"$DEATHS_DIRECT,"I"$DEATHS_INDIRECT,"F"$BEGIN_LAT,"F"$BEGIN_LON from details
update BEGIN_DATE: (BEGIN_YEARMONTH, BEGIN_DAY) from `det where not l=count each BEGIN_DAY;
update BEGIN_DATE: (BEGIN_YEARMONTH, ("0", BEGIN_DAY)) from `det where l=count each BEGIN_DAY;
update END_DATE: (END_YEARMONTH, END_DAY) from `det where not l=count each END_DAY;
update END_DATE: (END_YEARMONTH, ("0", END_DAY)) from `det where l=count each END_DAY;

//CALC DETAILS ELAPSED TIMES
t2::z.p;td1:t1-t0;td2:t2-t1;td3:t2-t0;td4::z.p

//INGEST FATALITIES
fatfiles:asc hsym each '$' ~/.home/conner/testgz/gzipped/,/: system "ls gzipped | grep -v gz | grep fatalities"
fatalities: (./) {(11#"";enlist ",") 0: x) each fatfiles
t5::z.p

//CREATE FATALITIES SUBSET TABLE AND CAST COLUMN TYPES
fat::select FATALITY_DATE,FATALITY_ID,EVENT_ID,FATALITY_TYPE,FATALITY_AGE,FATALITY_SEX,FATALITY_LOCATION from fatalities
update "I"$FATALITY_ID,"I"$EVENT_ID,"I"$FATALITY_TYPE,"I"$FATALITY_AGE,"F"$FATALITY_SEX from `fat;
update "D"$ID#FATALITY_DATE from `fat;

//CALC FATALITIES ELAPSED TIMES
t6::z.p;td4:t5-t4;td5:t6-t5;td6:t6-t4;td7:t6-t0;show ""

//PRINT SCRIPT TOTAL ELAPSED TIME
show (enlist "$UNZIPPING TIME: ")|enlist "$((-6.8_string t2t), " secs")
show ""

//PRINT DETAILS SUMMARY DICT
show ("TABLE: " | "$ROWS: " | "$COLS: " | "$COPY: " | "$CAST: " | "$TOTAL: ")|
details,(' $string count details),(' $string count cols details), '$((-6.8_string value each `td1`td2`td3), \: " secs"
show ""

//PRINT FATALITIES SUMMARY DICT
show ("TABLE: " | "$ROWS: " | "$COLS: " | "$COPY: " | "$CAST: " | "$TOTAL: ")|
fatalities,(' $string count fatalities),(' $string count cols fatalities), '$((-6.8_string value each `td4`td5`td6), \: " secs"
show ""

//PRINT SCRIPT TOTAL ELAPSED TIME
show (enlist "$FULL SCRIPT RUN ELAPSED TIME: ")|enlist "$((-6.8_string td7), " secs")
show ""
\\
```

```
conner@Linuxclevo:~/testgz$ q ingest_all_gz.q
KDB+ 4.0 2022.05.11 Copyright (C) 1993-2022 Kx Systems
164/ 8(24)core 23735MB conner linuxclevo 127.0.1.1 EXPIRE 2023.06.30 connermcnicholas@gmail.com KXCE #72902

()
""
UNZIPPING TIME:| 08.057 secs
""
TABLE:| details
ROWS:| 1740597
COLS:| 51
COPY:| 10.097 secs
CAST:| 03.174 secs
TOTAL:| 13.272 secs
""
TABLE:| fatalities
ROWS:| 20368
COLS:| 11
COPY:| 00.029 secs
CAST:| 00.007 secs
TOTAL:| 00.036 secs
""
FULL SCRIPT RUN ELAPSED TIME:| 21.366 secs
""
conner@Linuxclevo:~/testgz$ |
```

Alternative Stacks Explored

Apache Airflow as ELT orchestration framework (instead of Azure Data Factory)

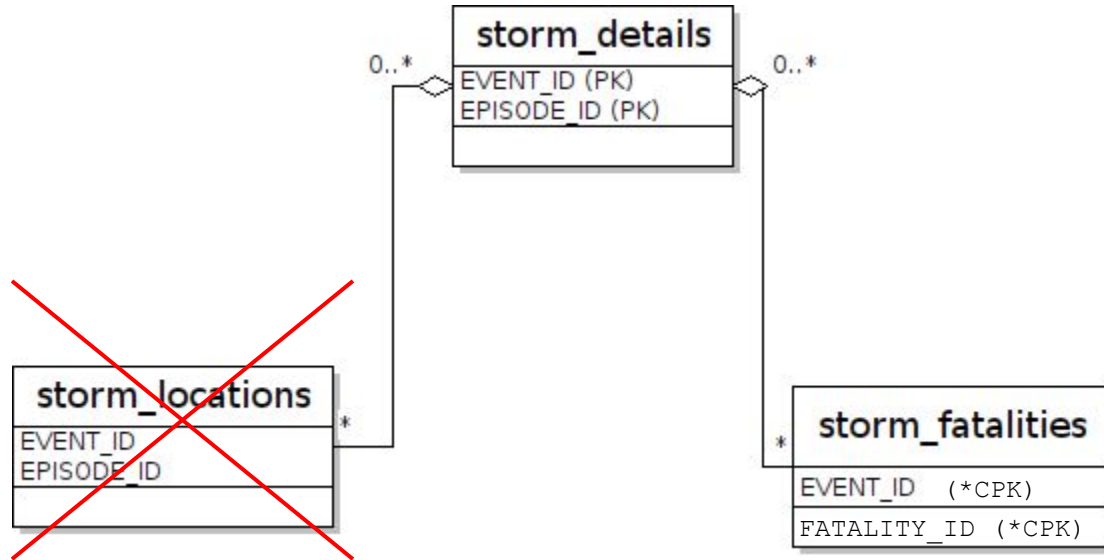
Spark as datastore and analytics engine via Databricks cluster (instead of MySQL+/Workbench)

Azure MS SQL Server+/SandDance (instead of MySQL+/Workbench)

Azure Batch as compute runtime (instead of Databricks cluster)

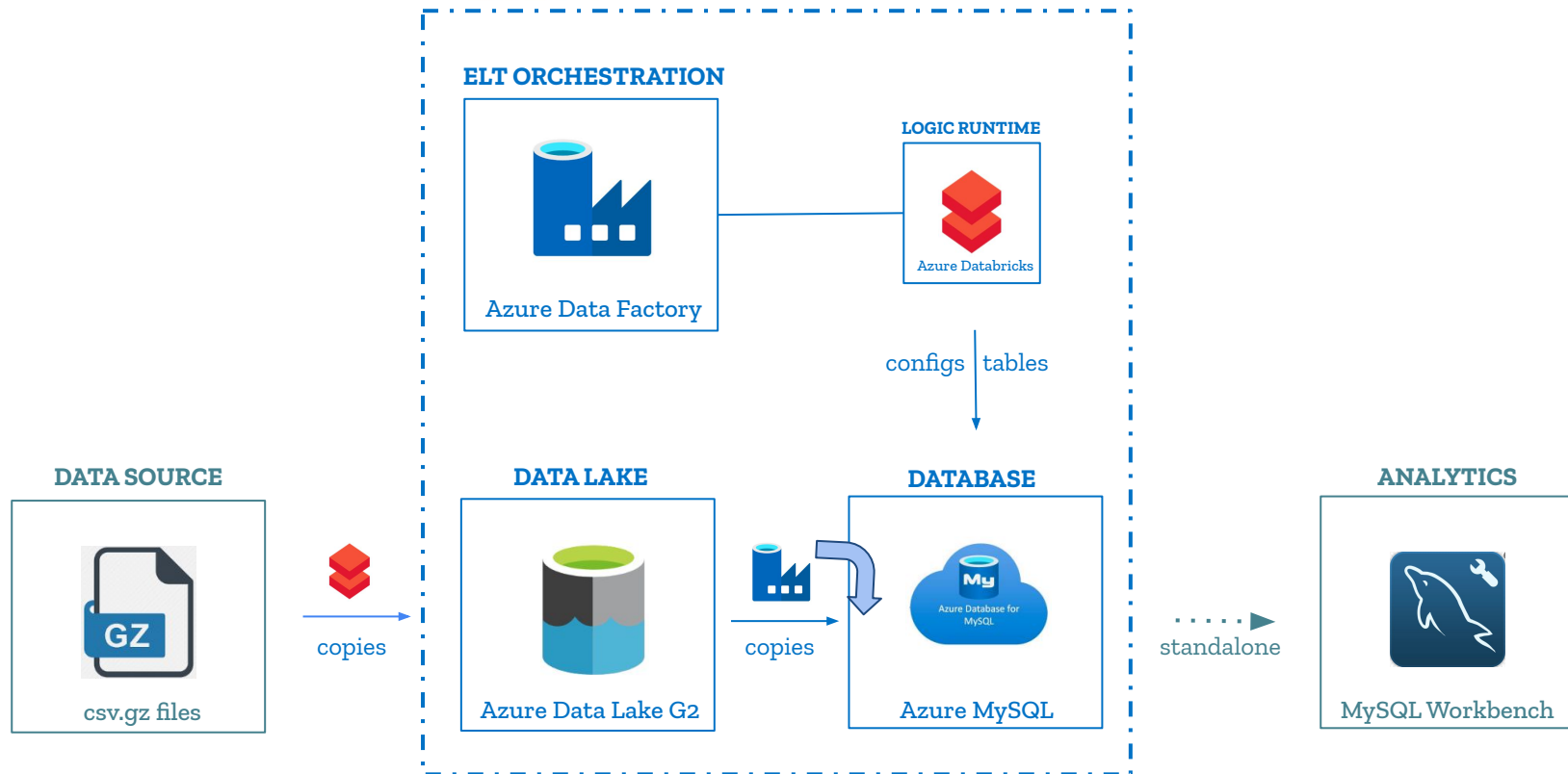
Azure Synapse Analytics as analytics engine (instead of MySQL+/Workbench)

Data Model



ALL LOCATIONS DATA REPRODUCED IN DETAILS

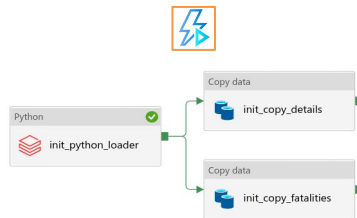
Architecture



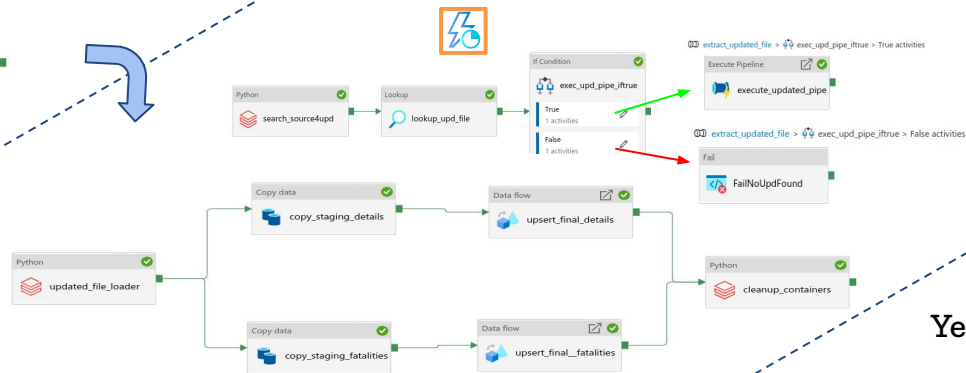
ETL Overview

Three Azure Data Factory pipelines tailored for different use case based on source data release cadence

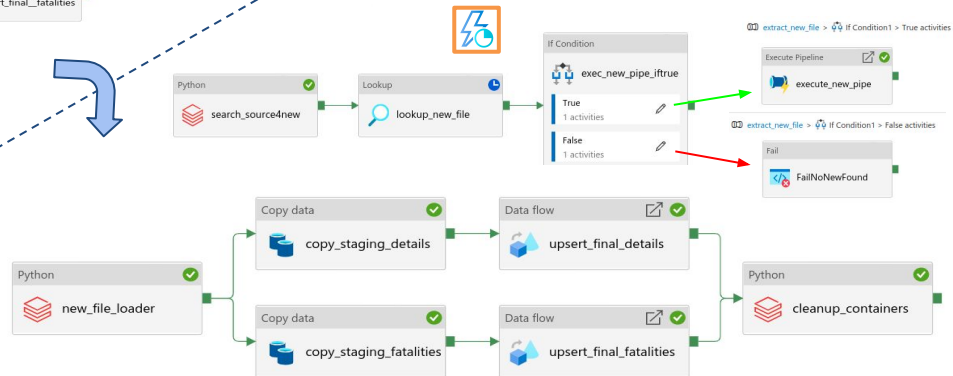
Initial Load



Monthly Update



Yearly New



Deep Dive: “Yearly New” ETL

1) New File Trigger

- Annually on April 17th Triggers “Extract New File” Pipeline

2) “Extract New File” Pipeline

- check source url for new file drop
 - pull if available
 - success: triggers “Load+Transform New File” Pipeline
 - failure: triggers retry in 1 day

3) “Load+Transform New File” Pipeline

Trigger configuration form:

- Name: yearly_new_tumbling
- Description:
- Type: TumblingWindowTrigger
- Start Date (UTC): 4/17/23 00:00:00
- Recurrence: Every 12 Months
- Specify an end date: ☐
- Advanced: ☐ Trigger, ☐ Offset, ☐ Window size
- Delay: 00:00:00
- Max concurrency: 1
- Retry policy: count 13
- Retry policy: interval in seconds 86400
- Annotations:

Microsoft Azure | Data Factory | datafactoryausins

Would you like to try preview updates to Azure Data Factory Studio? Open settings to learn more and opt in

Pipeline runs

Triggered Debug Rerun Cancel Refresh Edit columns List Gantt

Filter by name ID or name Local time: Custom (7/22/22, 7:30 PM - 7/22/22, 8:12 PM) Pipeline name: All Status: All Runs: All runs Triggered by: All Add filter Copy filters Export to CSV

Pipeline name	Run start	Run end	Duration	Activity	Status	Error	Run ID
2 "Extract New File" Pipeline				1 "Yearly New File" Scheduled Trigger (retries daily until success)			
<input type="checkbox"/> pull_new_trigger_on_success	Jul 22, 2022, 7:45:00 pm	Jul 22, 2022, 7:46:24 pm	00:01:23	Pull_New	✓ Succeeded		6baa09fc-5c34-454b-bdc2-2f0d51cd2318
<input type="checkbox"/> new_blob_triggered	Jul 22, 2022, 7:46:22 pm	Jul 22, 2022, 7:52:17 pm	00:05:55	8a109a36-42ee-4616-8845-c6d0c70c5331	✓ Succeeded		6fa1cd32-b1b5-4596-921f-b66f569117f0
3 "Load+Transform New File" Pipeline							

Last refreshed 0 minutes ago

id of the "Execute Pipeline" Activity that completes the "Extract New File" Pipeline

"Extract New File" Pipeline

Databricks Python Activity:

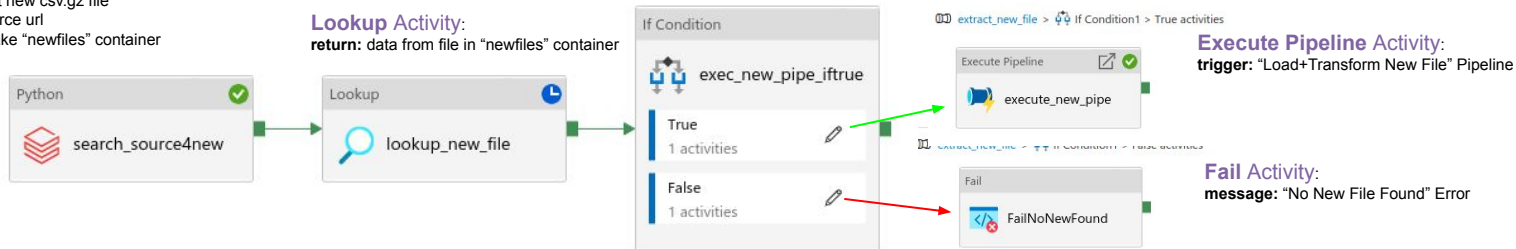
spark cluster: initialized with environment variables for auth
runs: `python new_file_to_blob_only.py`
if: source url has dropped a new filename for new year
then: extract new csv.gz file
from: source url
to: data lake "newfiles" container

Lookup Activity:

return: data from file in "newfiles" container

If Condition Activity:

if: # of rows from Lookup'd data > 0
then: trigger "Load+Transform New File" Pipeline
else: trigger "Fail" of pipeline, enabling retry in 1 day



"Load+Transform New File" Pipeline

Copy Data Activity

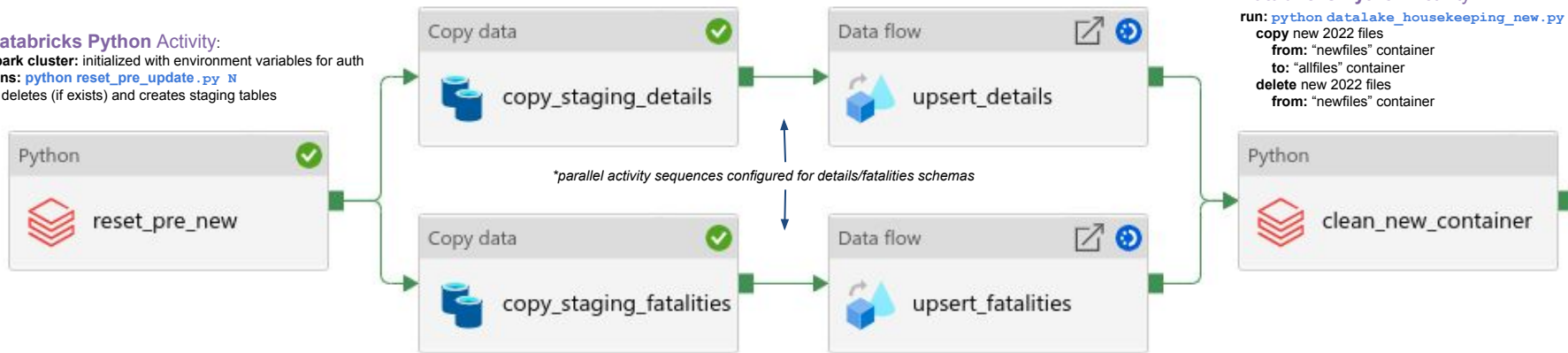
copy: all data from new year csvgz file
from: data lake "newfiles" container
to: mysql staging table

Data Flow Activity

insert: new rows based on primary key (equates to all)
from: mysql staging table
to: mysql final table

Databricks Python Activity:

spark cluster: initialized with environment variables for auth
runs: `python reset_pre_update.py`
 deletes (if exists) and creates staging tables



Databricks Python Activity

run: `python dataLake_housekeeping_new.py`
copy new 2022 files
from: "newfiles" container
to: "allfiles" container
delete new 2022 files
from: "newfiles" container

Testing

General Tests verify each year has a file in Data Lake for both table

Pipeline Tests verify each line from source files have rows in both MySQL tables

- *8 Total Tests = 2 General Tests + 6 Pipeline Tests*
 - *2 General Tests = 1 General Test x 2 Tables*
 - *6 Pipeline Tests = 3 Pipeline Tests x 2 Tables*

```
conner@Linuxclevo: ~/SevereWeatherDB/Step9/testing
conner@Linuxclevo:~/SevereWeatherDB/Step9/testing$ pytest -v runinitialtests.py runupdatetests.py runnewtests.py -W ignore::DeprecationWarning
===== test session starts =====
platform linux -- Python 3.10.5, pytest-6.2.5, py-1.10.0, pluggy-1.0.0 -- /usr/local/bin/python
cachedir: .pytest_cache
rootdir: /home/conner/SevereWeatherDB/Step9/testing
plugins: anyio-3.6.1
collected 8 items

runinitialtests.py::test_details_uploaded PASSED [ 12%]
runinitialtests.py::test_fatalities_uploaded PASSED [ 25%]
runinitialtests.py::test_details_count PASSED [ 37%]
runinitialtests.py::test_fatalities_count PASSED [ 50%]
runupdatetests.py::test_details_updated PASSED [ 62%]
runupdatetests.py::test_fatalities_updated PASSED [ 75%]
runnewtests.py::test_details_new PASSED [ 87%]
runnewtests.py::test_fatalities_new PASSED [100%]

===== 8 passed in 101.62s (0:01:41) =====
```

Deployment

Deploys Azure resources via Docker container image:

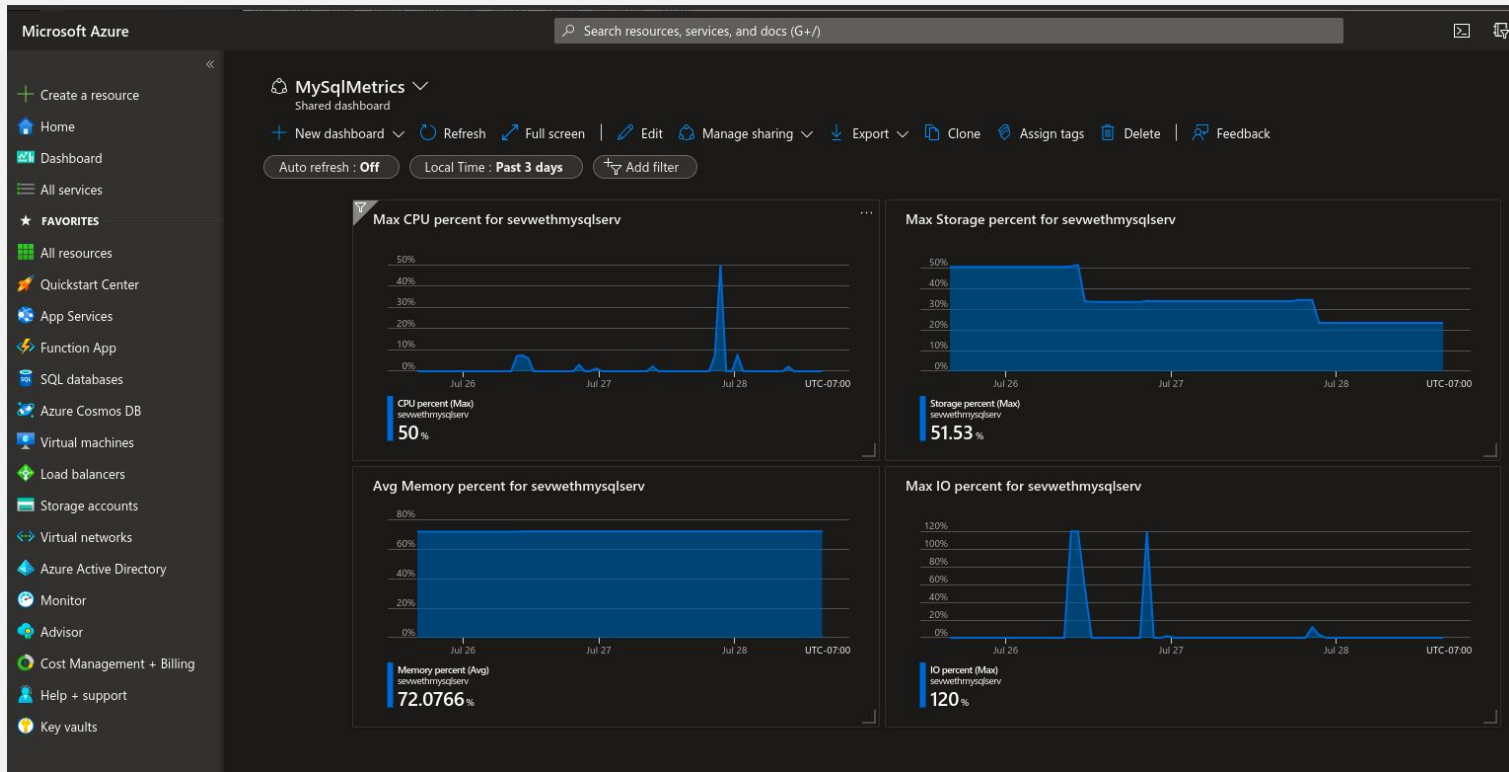
- data lake blob storage
- mysql database
- databricks
- data factory
 - iii. creates pipelines based on json configs in DATAFACTORY_pipelines
 - iv. runs init pipeline to ingest all available data at source
- Requirement: Docker
 - iii. to start azure cli container, run from local terminal:
- ./start.sh
 - iii. now from inside azure cli container shell, run:
- bash-5.1# ./create_resources.sh
 - iii. login to az by entering given code at <https://microsoft.com/devicelogin>
- json metadata describing created resources will print to stdout (execution_log.txt)

```
conner@LinuxLeve:~/SevereWeatherDB/ProjectJourney/Step9_Deploy_To_Production/AZURE_resources$ ./start.sh
Unable to find image 'mcr.microsoft.com/azure-cli:latest' locally
latest: Pulling from azure-cli
408cc74d12b: Pull complete
8f22aa6a21a6: Pull complete
44cc866f118a: Pull complete
3624af7d529: Pull complete
ae78d2f3e6f: Pull complete
637e76a679fb: Pull complete
4b51ef8683b: Pull complete
c9768a15b64: Pull complete
005d675e4066: Pull complete
aa062aab2311: Pull complete
1abb35879ae: Pull complete
Digest: sha256:df5911c7d58978a94bcd9653368ffa6e4683cee840be3a126cea8df48d4db
Status: Downloaded newer image for mcr.microsoft.com/azure-cli:latest
bash-5.1# ./create_resources.sh
az login
To sign in, use a web browser to open the page https://microsoft.com/devicelogin and enter the code F0X56ERY4 to authenticate.
{
  "cloudName": "AzureCloud",
  "homeTenantId": "2f3a3629-3599-4272-ac5e-cd4c5a76d972",
  "id": "b5f807fa-c5eb-4a4d-8357-76a176efc946",
  "isDefault": true,
  "managedByTenants": [
    {
      "tenantId": "2f4a9838-26b7-47ee-be68-ccc1fdec5953"
    }
  ],
  "name": "SparkPipelineSub",
  "state": "Enabled",
  "tenantId": "2f3a3629-3599-4272-ac5e-cd4c5a76d972",
  "user": {
    "name": "connermcnicholas@gmail.com",
    "type": "user"
  }
}
```

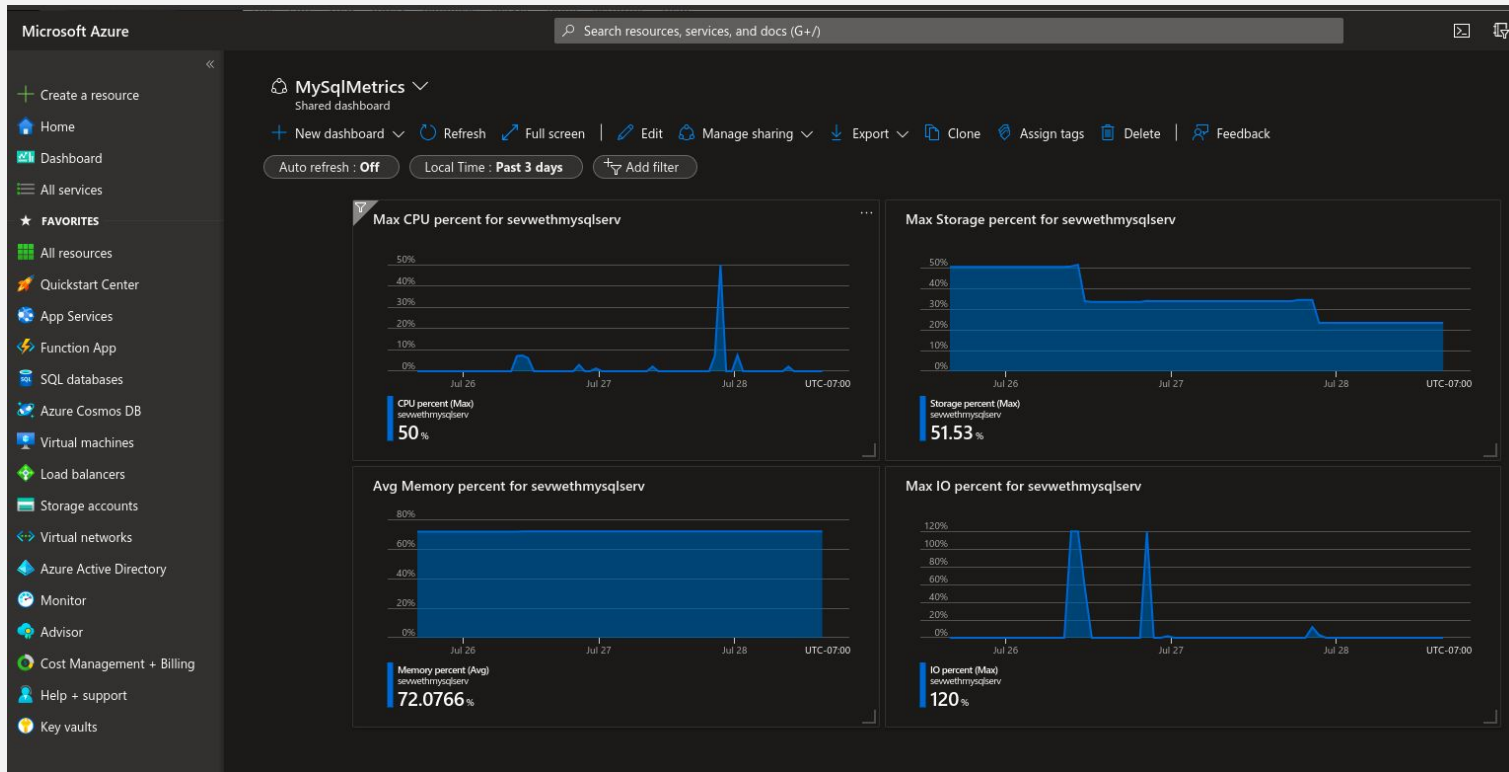
```
{
  "apiProfile": null,
  "kind": "Storage",
  "targetNetworks": null,
  "lastGeoFailoverTime": null,
  "location": "australiaeast",
  "minimumOverSize": "15.0",
  "name": "adbstorageaccount",
  "networkRuleSet": {
    "type": "AzureDefault",
    "defaultAction": "Allow",
    "rules": [
      {
        "resourceAccessTypes": null,
        "virtualNetworkRules": [
        ]
      }
    ]
  },
  "targetEndpoints": {
    "blob": "https://adbstorageaccount.blob.core.windows.net/",
    "dfs": "https://adbstorageaccount.dfs.core.windows.net/",
    "file": "https://adbstorageaccount.file.core.windows.net/",
    "unconnectedEndpoints": null,
    "microsoftEndpoints": null,
    "queue": "https://adbstorageaccount.queue.core.windows.net/",
    "table": "https://adbstorageaccount.table.core.windows.net/",
    "web": "https://adbstorageaccount.web.core.windows.net/"
  },
  "targetLocation": "australiaeast",
  "privateEndpointConnections": [
  ],
  "provisioningState": "Succeeded",
  "publicNetworkAccess": null,
  "resourceGroup": "adbstoragecrgrp",
  "routingPreference": null,
  "security": null,
  "secondaryEndpoints": {
    "blob": "https://adbstorageaccount-secondary.blob.core.windows.net/",
    "dfs": "https://adbstorageaccount-secondary.dfs.core.windows.net/",
    "file": null,
    "unconnectedEndpoints": null,
    "microsoftEndpoints": null,
    "queue": "https://adbstorageaccount-secondary.queue.core.windows.net/",
    "table": "https://adbstorageaccount-secondary.table.core.windows.net/",
    "web": "https://adbstorageaccount-secondary.web.core.windows.net/"
  },
  "secondaryLocation": "australiasoutheast",
  "sku": {
    "name": "Standard_RAGRS",
    "tier": "Standard"
  },
  "statusOffyTeam": "available",
  "statusOffSecondary": "available",
  "storageAccount3huConversionStatus": null,
  "tags": {
  },
  "type": "Microsoft.Storage/storageAccounts"
}

-- az storage account hns-migration start --type validation -n adbstorageaccount -g adbstoragecrgrp
-- az storage account hns-migration start --type upgrade -n adbstorageaccount -g adbstoragecrgrp
-- STARTING ...
```

Monitoring



Monitoring



Query

The screenshot shows a SQL query editor with multiple tabs. The active tab is 'SQL File 14*', which contains the following SQL query:

```
1 • SELECT
2   EVENT_TYPE,
3   SUM(DEATHS_DIRECT+DEATHS_INDIRECT) AS DEATHS,
4   SUM(INJURIES_DIRECT+INJURIES_INDIRECT) AS INJURIES FROM details where BINARY EVENT_TYPE != BINARY upper(EVENT_TYPE)
5   GROUP BY 1 ORDER BY 2 DESC
```

Below the query editor is a 'Result Grid' showing the results of the query. The grid has columns for '#', 'EVENT_TYPE', 'DEATHS', and 'INJURIES'. The results are sorted by 'DEATHS' in descending order.

#	EVENT_TYPE	DEATHS	INJURIES
24	Ice Storm	184	1821
25	Hurricane	137	369
26	Debris Flow	125	244
27	Marine Strong Wind	68	54
28	Dust Storm	61	755
29	Coastal Flood	58	15
30	Marine Thunderst...	53	66
31	Storm Surge/Tide	36	42
32	Tsunami	33	150
33	Hail	26	1565
34	Lake-Effect Snow	23	36
35	Sneakerwave	22	13
36	Freezing Fog	20	59
37	Marine Dense Fog	13	4
38	Marine Tropical St...	9	0
39	Tropical Depression	8	4
40	Marine High Wind	8	7
41	Frost/Freeze	6	16
42	Waterspout	5	3
43	Sleet	4	19
44	Dust Devil	4	82

The bottom of the screenshot shows a 'Result 4' tab, indicating that the query results are displayed in the fourth result set.