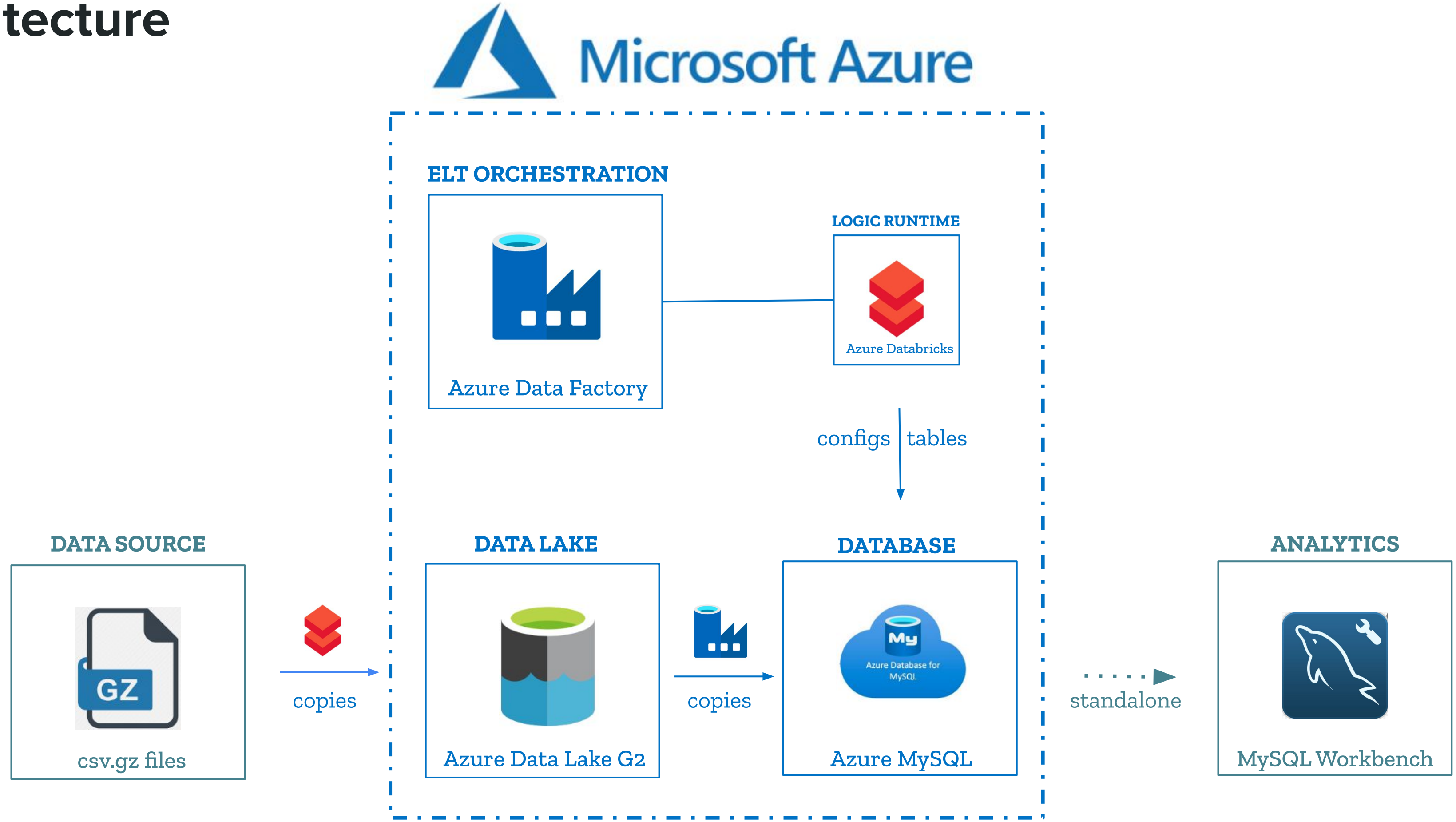
Final Submission

Capstone

Background

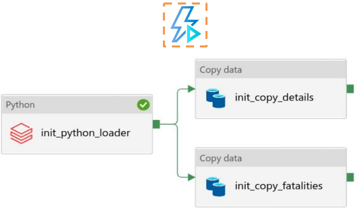
Architecture



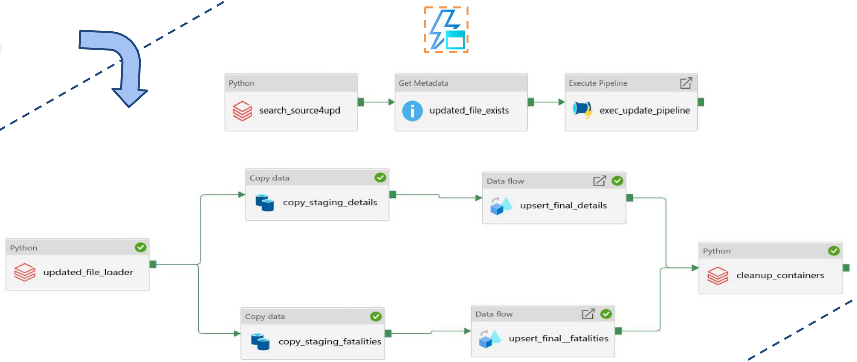
ETL Overview

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

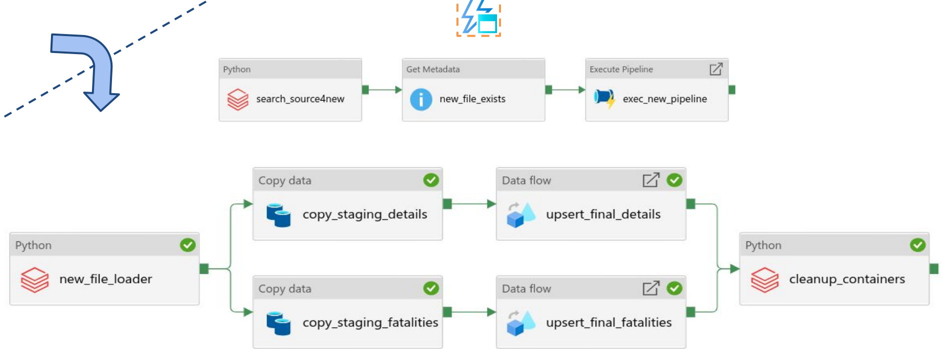
Initial Load



Monthly Update

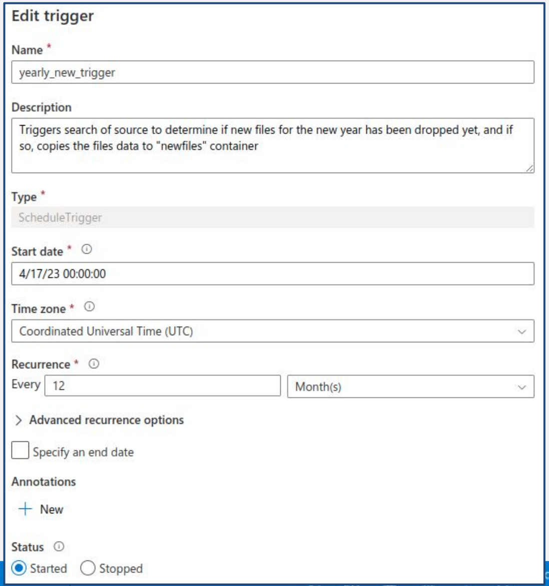


Yearly New



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



Edit trigger

Name *
yearly_new_trigger

Description
Triggers search of source to determine if new files for the new year has been dropped yet, and if so, copies the files data to "newfiles" container

Type *
ScheduleTrigger

Start date *
4/17/23 00:00:00

Time zone *
Coordinated Universal Time (UTC)

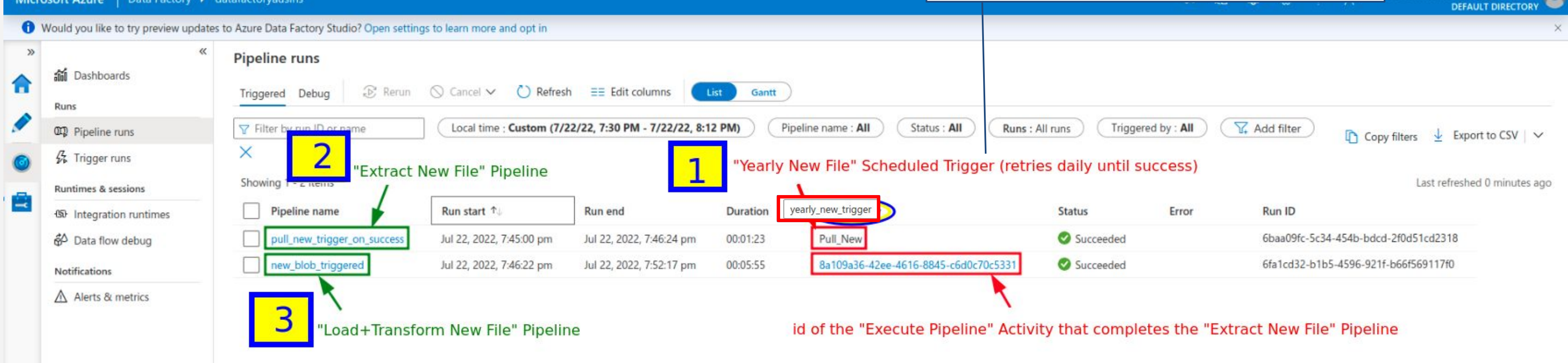
Recurrence *
Every 12 Month(s)

> Advanced recurrence options

□ Specify an end date

Annotations
+ New

Status
● Started ○ Stopped



Microsoft Azure | Data Factory ▸ datafactoryausins
connermcnicholas@gmail.com
DEFAULT DIRECTORY

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

Pipeline runs

Filter by run 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

Showing 1 - 2 items

Last refreshed 0 minutes ago

□	Pipeline name	Run start ↑↓	Run end	Duration	yearly_new_trigger	Status	Error	Run ID
□	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-bdcd-2f0d51cd2318
□	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

1 "Yearly New File" Scheduled Trigger (retries daily until success)

2 "Extract New File" Pipeline

3 "Load+Transform New File" Pipeline

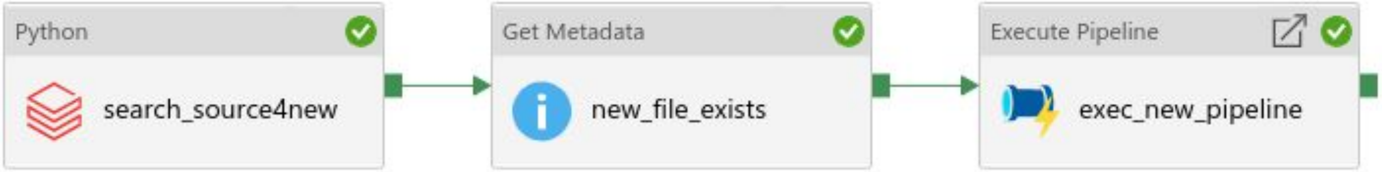
id of the "Execute Pipeline" Activity that completes the "Extract New File" Pipeline

"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

Get Metadata Activity:
if: file exists in "newfiles" container
  then: continue

Execute Pipeline Activity:
trigger: "Load+Transform New File" Pipeline



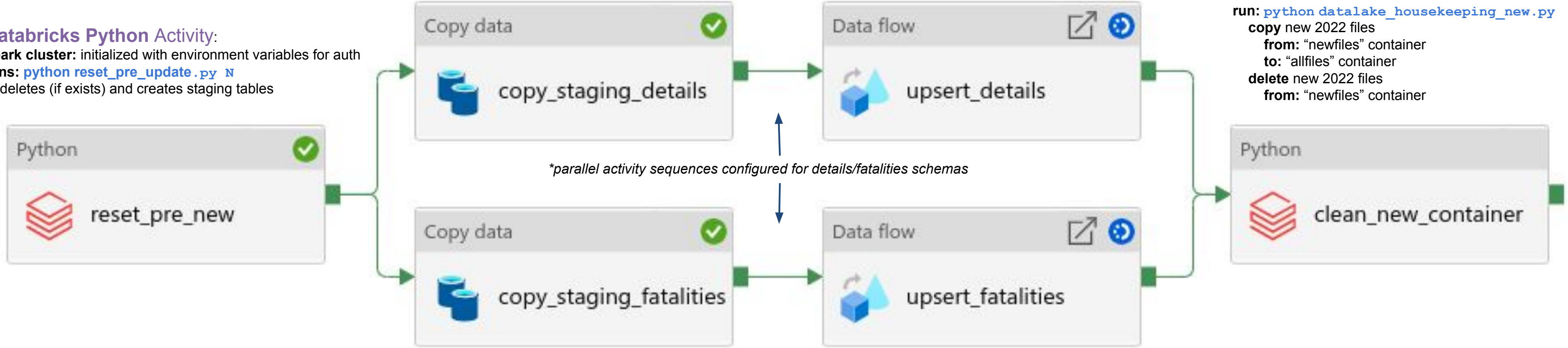
"Load+Transform New File" Pipeline

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 N`
  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



**parallel activity sequences configured for details/fatalities schemas*

Development Process

Testing

Deployment

Monitoring

Lessons Learned