# Benchmark TPC-H on GCP BigQuery and Azure SQL Data Warehouse

Data volume: 100GB, 1TB and 10TB

www.tpc.org

## Benchmark Preparation

### Prepare VM

#### GCP

* Create VM with 4-8 cores and 30GB ram (it was overutilized for 1TB and 10 TB datasets generation)
* Add and mount proper persistent disk with enough space
  + <https://cloud.google.com/compute/docs/disks/add-persistent-disk>

### Build DBGen Program

### Generating Data

#### GCP

#### Azure

### Loading Data into Buckets

#### GCP

#### Azure

<https://docs.microsoft.com/en-us/azure/storage/common/storage-use-azcopy-linux>

wget -O azcopy.tar.gz https://aka.ms/downloadazcopylinux64

tar -xf azcopy.tar.gz

sudo ./install.sh

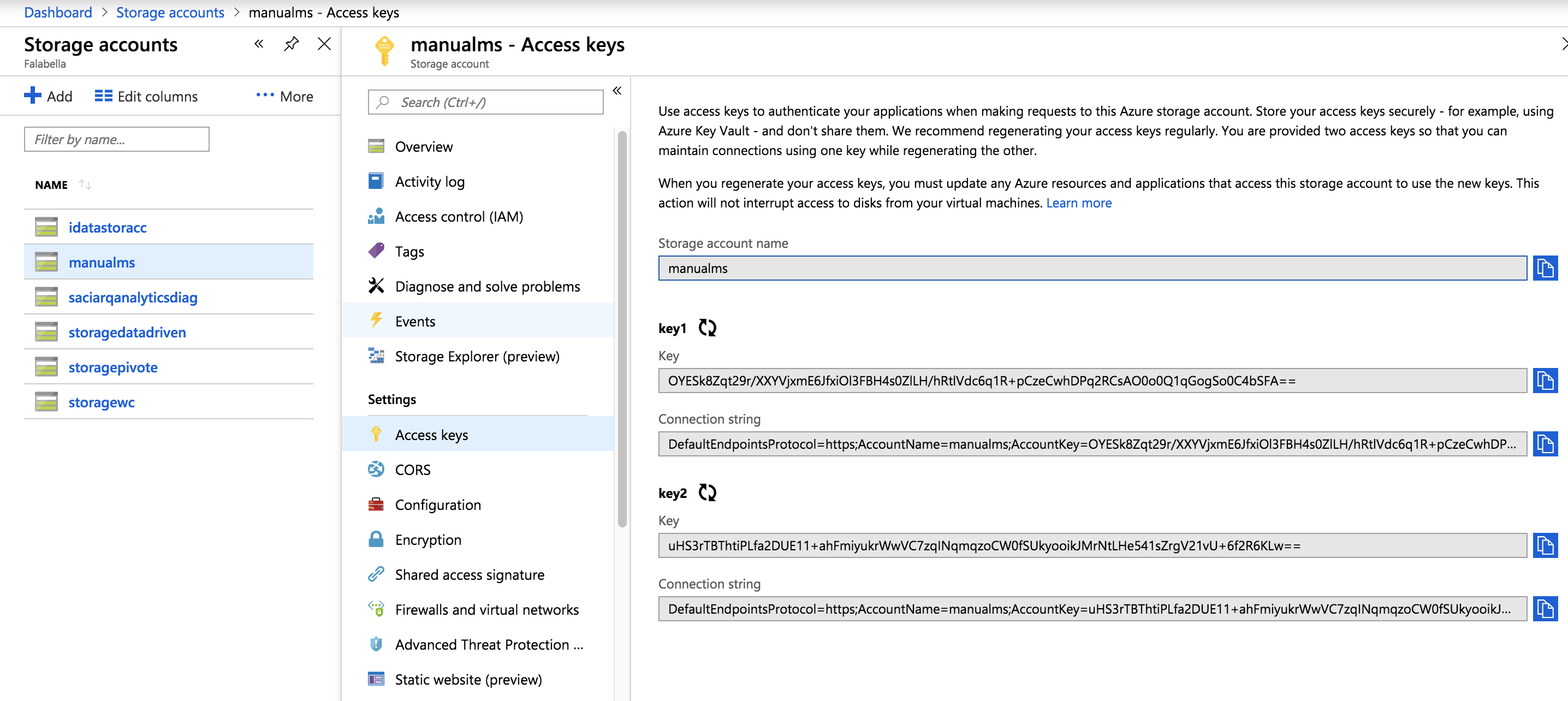
azcopy \

--source '/mnt/disks/disk-1/lineitem.tbl' \

--destination 'https://manualms.blob.core.windows.net/gendata-tpch100v2/lineitem.tbl' \

--dest-key 'OYESk8Zqt29r/XXYVjxmE6JfxiOl3FBH4s0ZlLH/hRtlVdc6q1R+pCzeCwhDPq2RCsAO0o0Q1qGogSo0C4bSFA=='

Para la key:



Comment: Crear grupo de recurso de Blobs no es natural

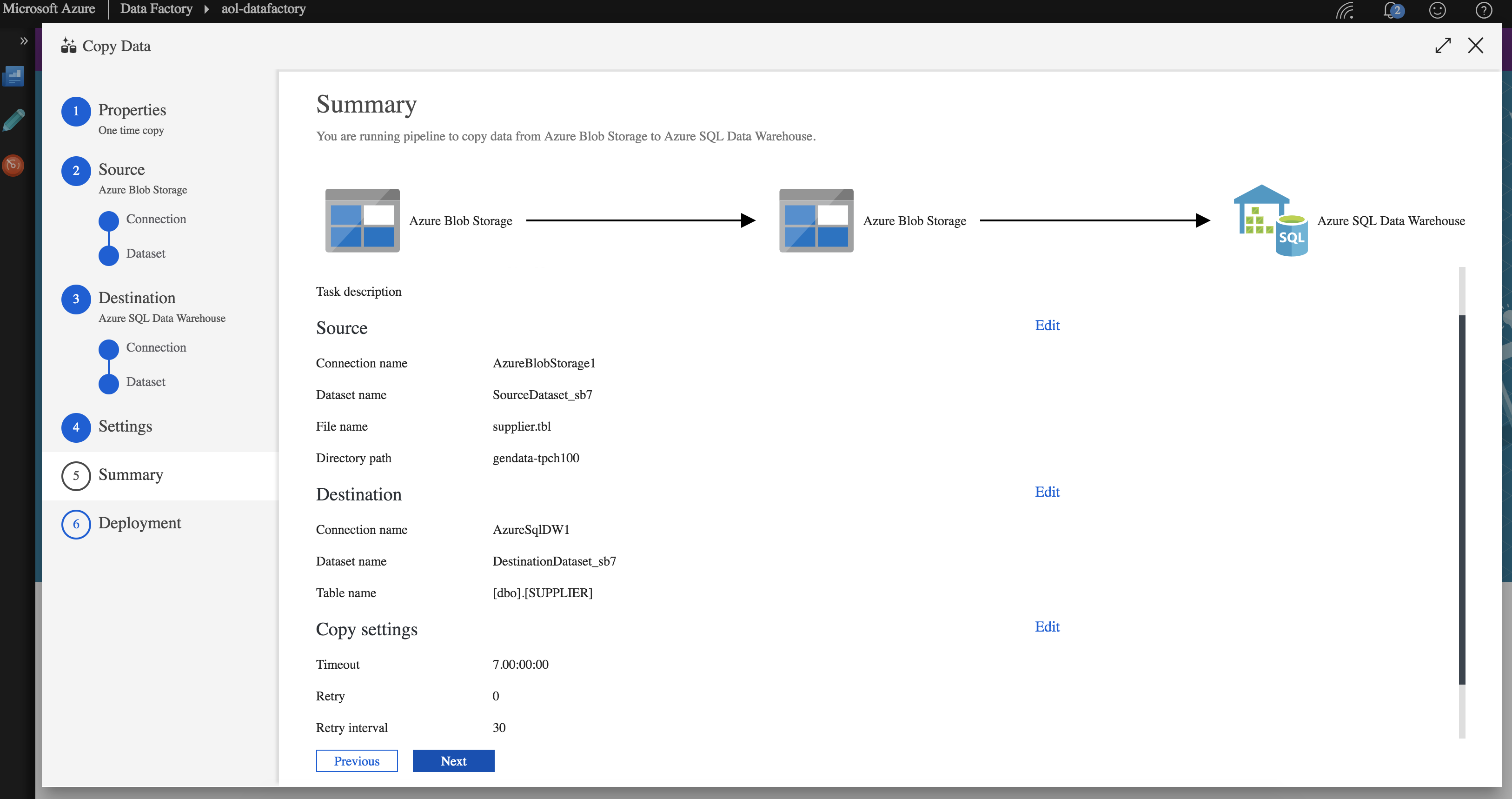
Comment: Storage account es separado de lugar donde estan los blob. Es engorroso tener la informacion tan dispersa

### Loading data into Data bases

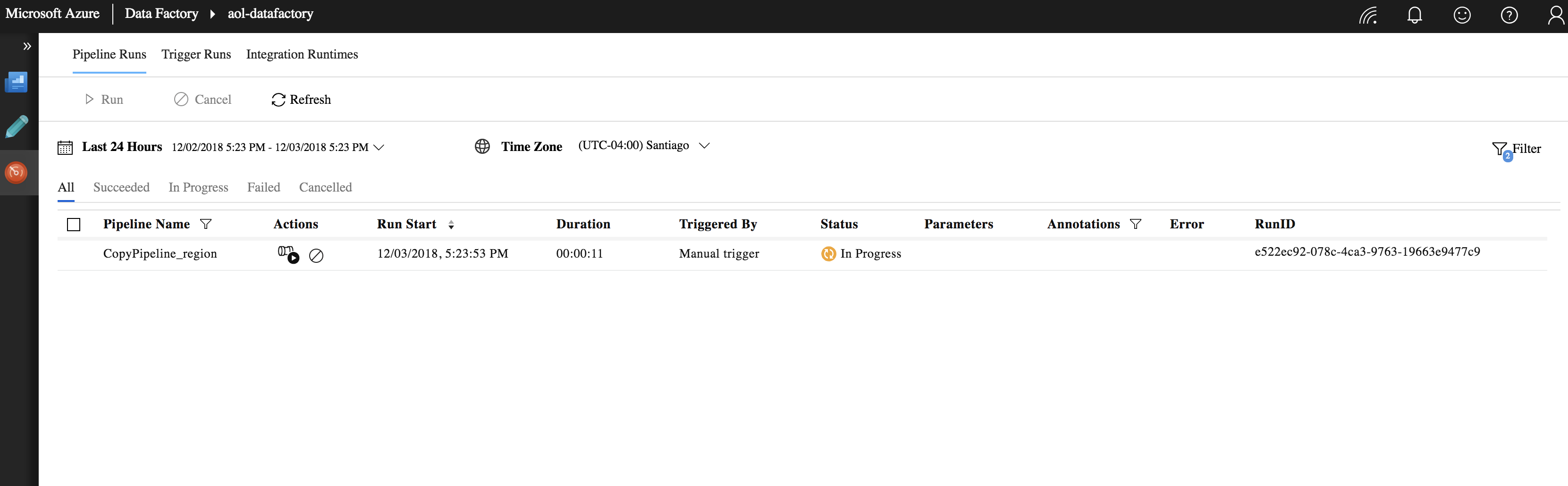
#### GCP

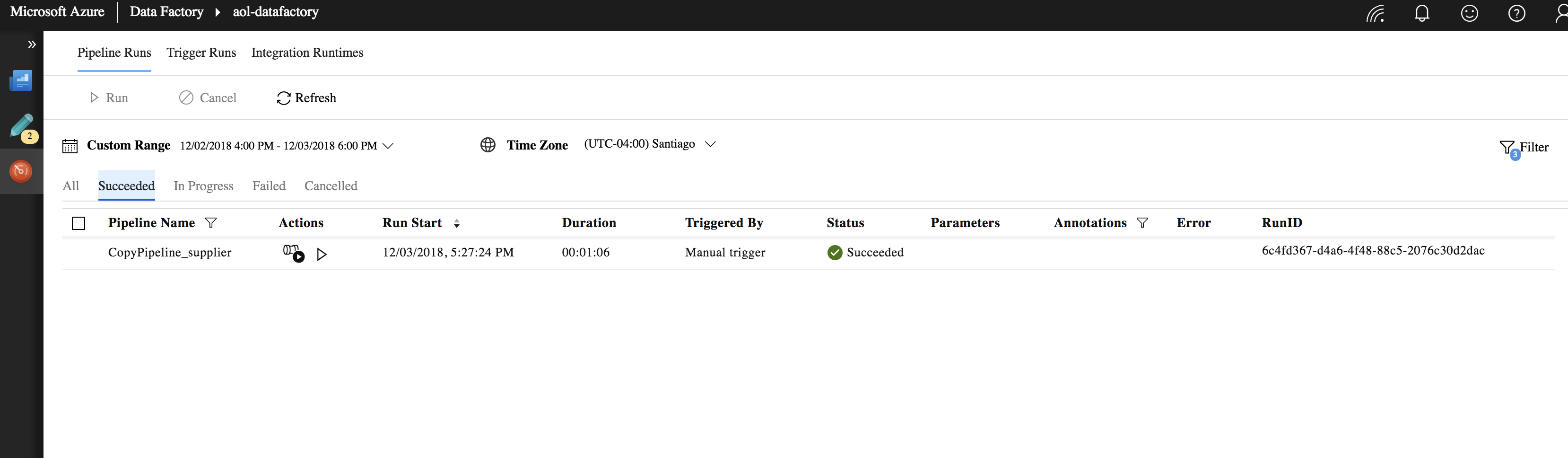
#### Azure

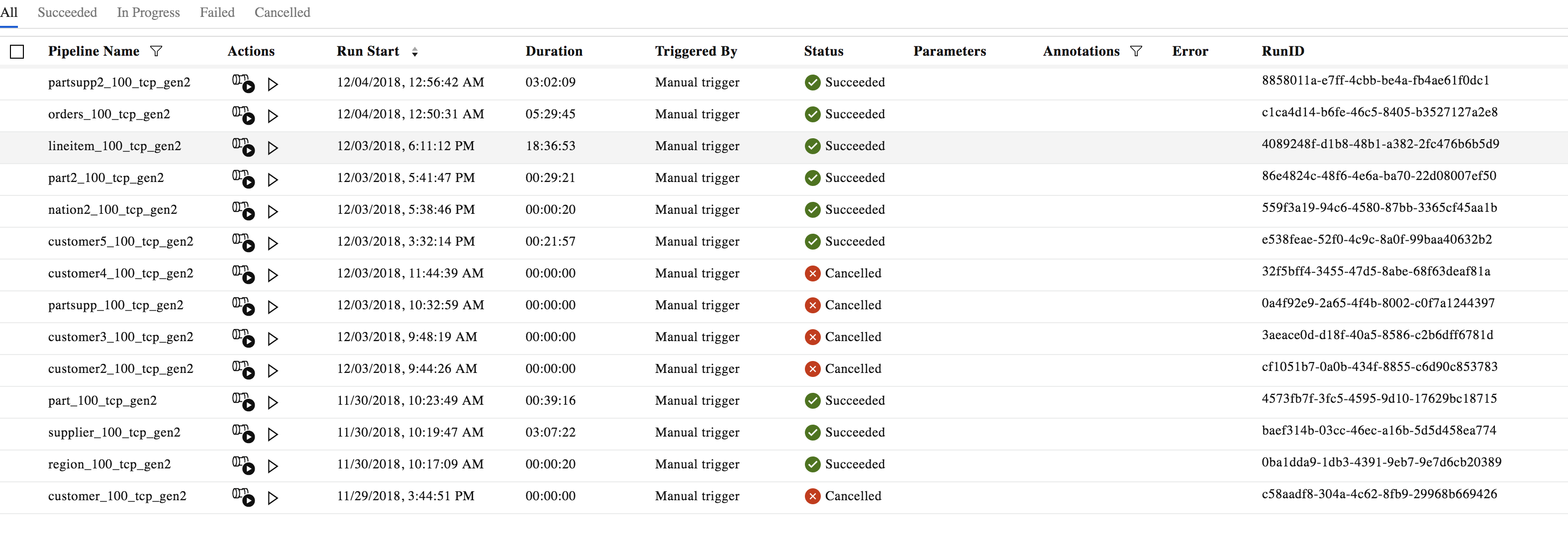


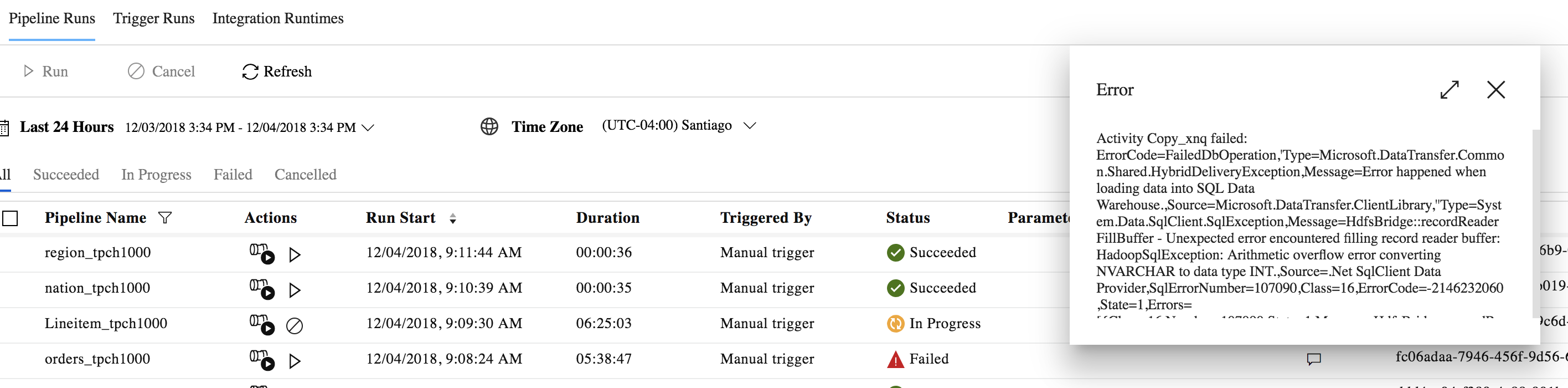


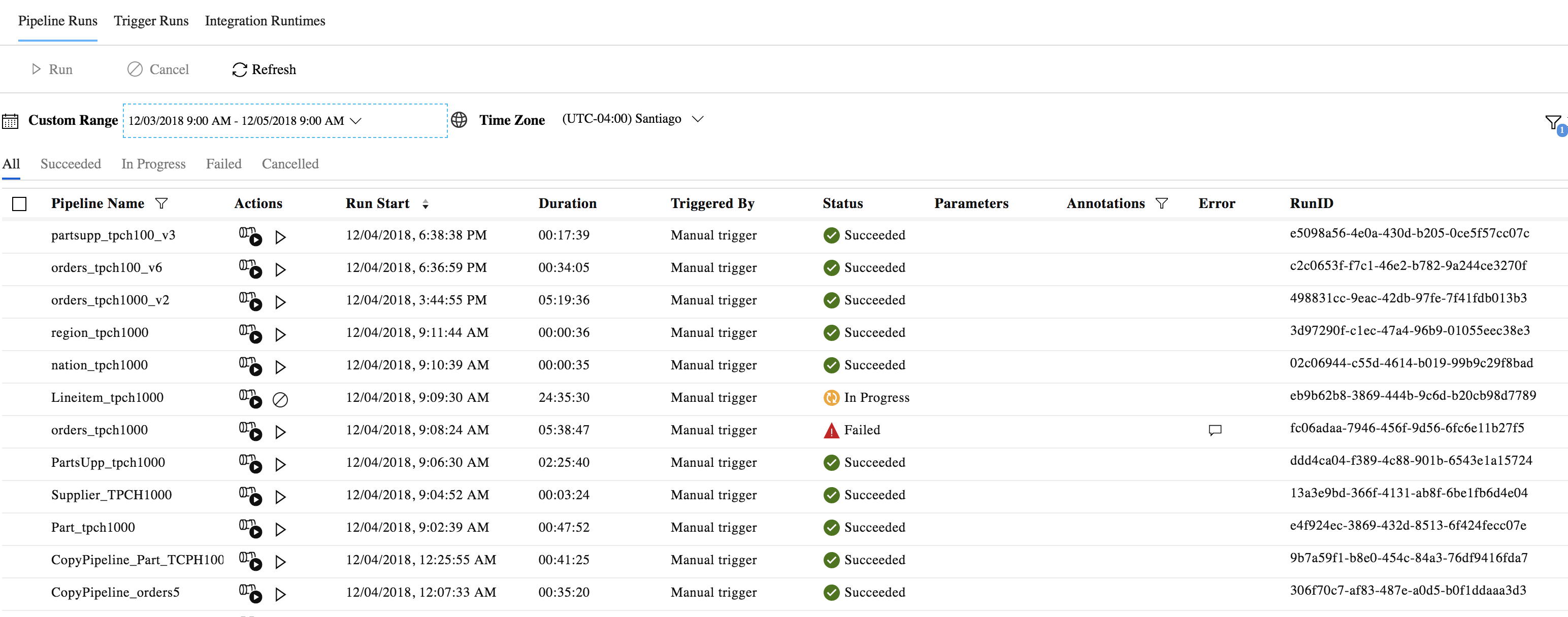
Comment: Not that easy to monitor:

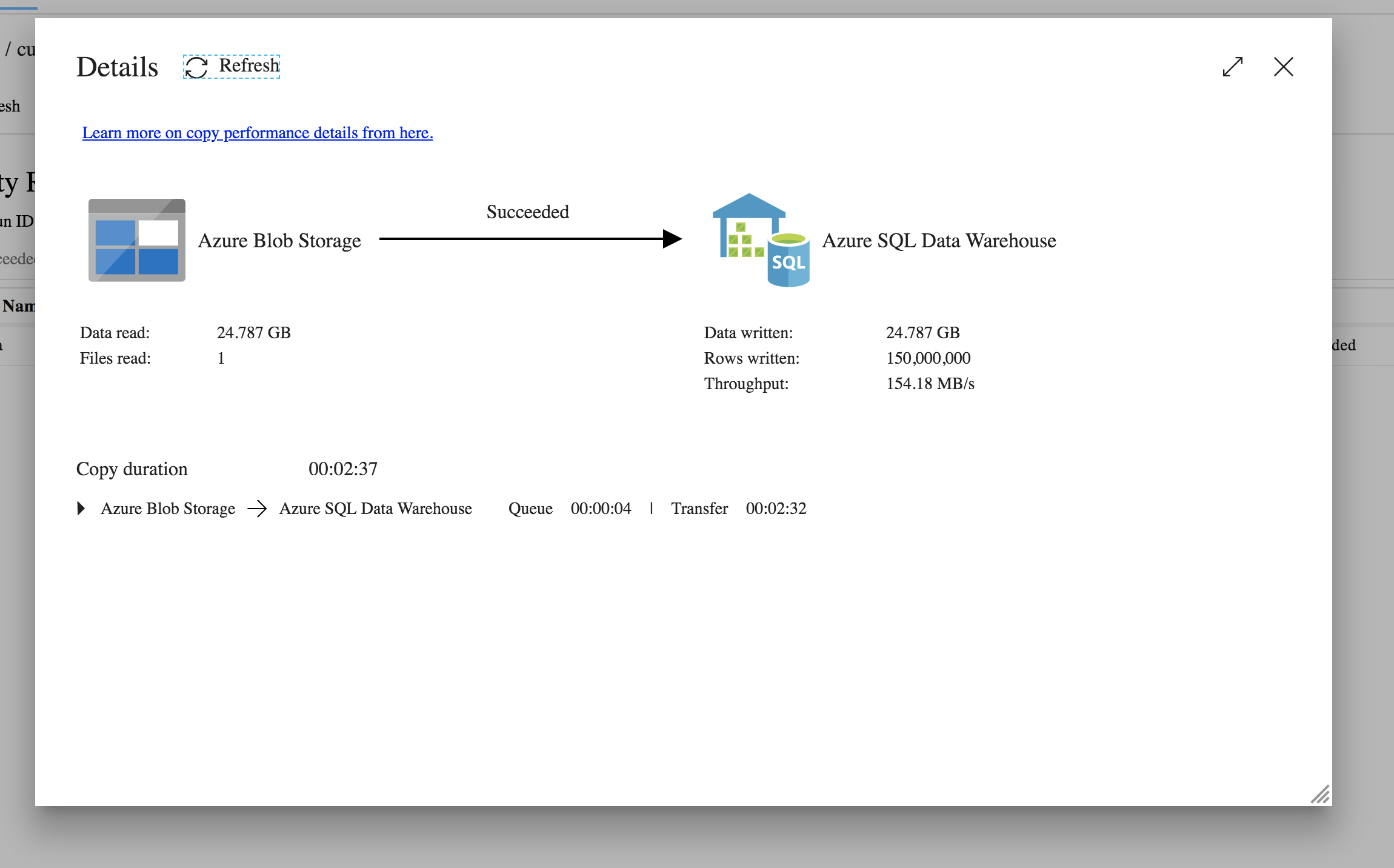




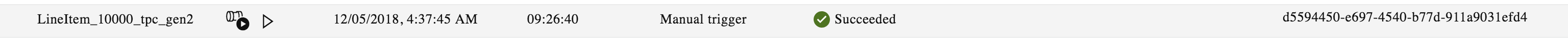








<https://docs.microsoft.com/en-us/azure/data-factory/v1/data-factory-load-sql-data-warehouse#step-1-configure-data-loading-schedule>



### Queries Generation

* 22 Queries generated using same TPC-H script ‘qgen.sh’
* Then, each query is adapted for each Data base syntax
* BigQuery required more modifications in queries, since types are quite different regarding to the accepted syntax
* SQL Data Warehouse did not need many modifications

### Tool - JMeter

#### GCP

<https://cloud.google.com/bigquery/docs/reference/rest/v2/jobs/query>

Query Jobs can be seen easily in the GUI



Api is largely flexible, due to cache, response timeout, jobid for see results, asyncronus call,

#### Azure

It is not straightforward to know which queries are being executed in the UI if multiple are send via JDBC

JDBC and parameters are less flexible. No able no configure cache or results, etc.