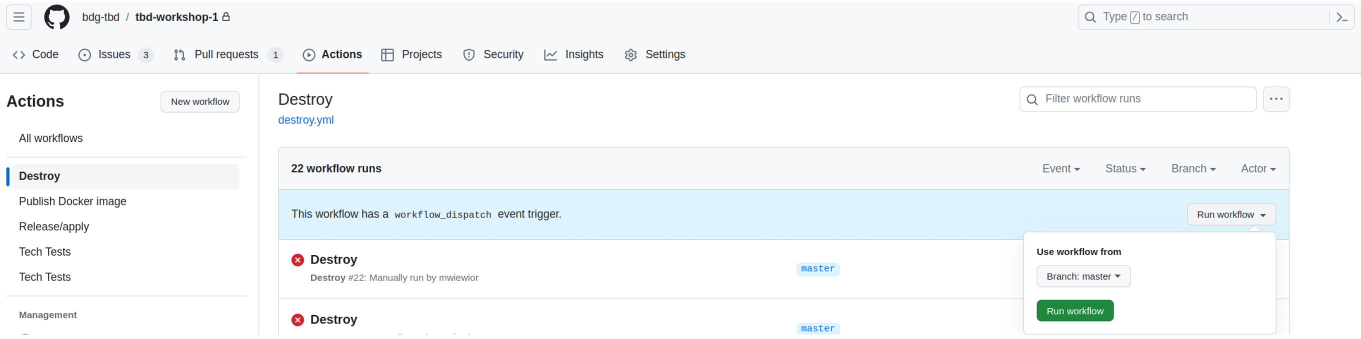
IMPORTANT ❗❗❗ Please remember to destroy all the resources after each work session. You can recreate infrastructure by creating new PR and merging it to master.



22 workflow runs

This workflow has a `workflow_dispatch` event trigger.

Destroy
Destroy #22: Manually run by mwiewior

Destroy

0. The goal of this phase is to create infrastructure, perform benchmarking/scalability tests of sample three-tier lakehouse solution and analyze the results using:

- TPC-DI benchmark
- dbt - data transformation tool
- GCP Composer - managed Apache Airflow
- GCP Dataproc - managed Apache Spark
- GCP Vertex AI Workbench - managed JupyterLab

Worth to read:

- https://docs.getdbt.com/docs/introduction
- https://airflow.apache.org/docs/apache-airflow/stable/index.html
- https://spark.apache.org/docs/latest/api/python/index.html
- https://medium.com/snowflake/loading-the-tpc-di-benchmark-dataset-into-snowflake-96011e2c26cf
- https://www.databricks.com/blog/2023/04/14/how-we-performed-etl-one-billion-records-under-1-delta-live-tables.html

2. Authors:

   Z7

   - Jakub Kokoszka: 304154
   - Jonatan Kasperczak: 341208
   - Będkowski Patryk: 310603

   Link to forked repo https://github.com/j-kokoszka/tbd-workshop-1/tree/master

3. Sync your repo with https://github.com/bdg-tbd/tbd-workshop-1.

4. Provision your infrastructure.

   a) setup Vertex AI Workbench `pyspark` kernel as described in point 8

   b) upload tpc-di-setup.ipynb to the running instance of your Vertex AI Workbench

5. In `tpc-di-setup.ipynb` modify cell under section ***Clone tbd-tpc-di repo***:

   a)first, fork https://github.com/mwiewior/tbd-tpc-di.git to your github organization.

b)create new branch (e.g. 'notebook') in your fork of tbd-tpc-di and modify profiles.yaml by commenting following lines:

```
      #"spark.driver.port": "30000"
      #"spark.blockManager.port": "30001"
      #"spark.driver.host": "10.11.0.5"  #FIXME: Result of the command
  (kubectl get nodes -o json |  jq -r
  '.items[0].status.addresses[0].address')
      #"spark.driver.bindAddress": "0.0.0.0"
```

This lines are required to run dbt on airflow but have to be commented while running dbt in notebook.

c)update git clone command to point to ***your fork***.

6. Access Vertex AI Workbench and run cell by cell notebook `tpc-di-setup.ipynb`.

   a) in the first cell of the notebook replace: `%env DATA_BUCKET=tbd-2023z-9910-data` with your data bucket.

   b) in the cell: `%%bash mkdir -p git && cd git git clone https://github.com/mwiewior/tbd-tpc-di.git cd tbd-tpc-di git pull` replace repo with your fork. Next checkout to 'notebook' branch.

   c) after running first cells your fork of `tbd-tpc-di` repository will be cloned into Vertex AI enviroment (see git folder).

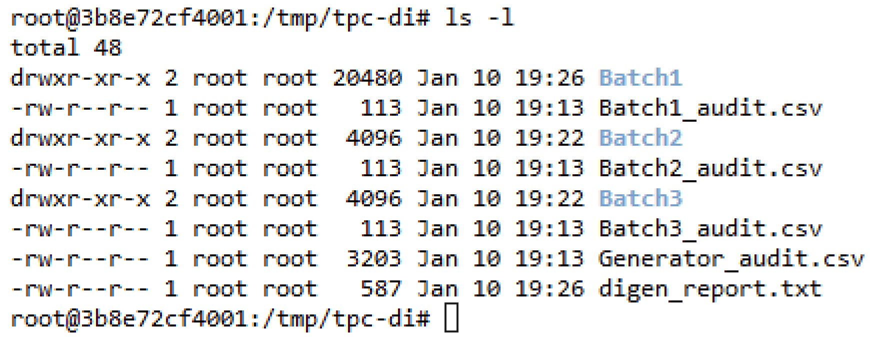
   d) take a look on `git/tbd-tpc-di/profiles.yaml`. This file includes Spark parameters that can be changed if you need to increase the number of executors and

```
  server_side_parameters:
      "spark.driver.memory": "2g"
      "spark.executor.memory": "4g"
      "spark.executor.instances": "2"
      "spark.hadoop.hive.metastore.warehouse.dir":
"hdfs:///user/hive/warehouse/"
```

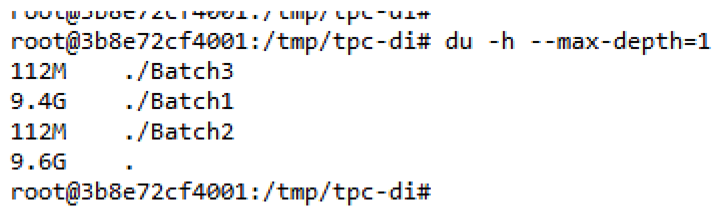
7. Explore files created by generator and describe them, including format, content, total size.

   Generator umieścił dane ścieżce /tmp/tpc-di. Formaty plików to : .txt, .xml, .csv

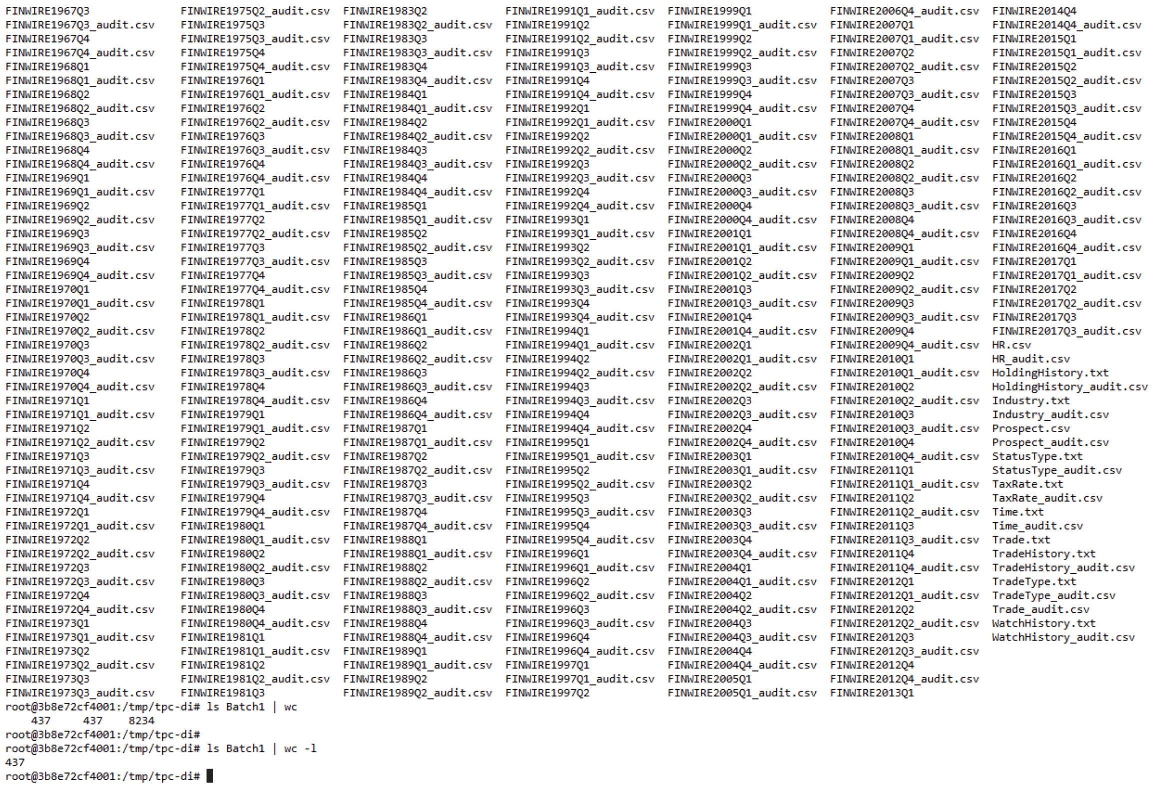
   Batch1 wygenerował prawie 10GB plików, Batch2 i Batch3 razem około 200MB



```
root@3b8e72cf4001:/tmp/tpc-di# ls -l
total 48
drwxr-xr-x 2 root root 20480 Jan 10 19:26 Batch1
-rw-r--r-- 1 root root   113 Jan 10 19:13 Batch1_audit.csv
drwxr-xr-x 2 root root  4096 Jan 10 19:22 Batch2
-rw-r--r-- 1 root root   113 Jan 10 19:13 Batch2_audit.csv
drwxr-xr-x 2 root root  4096 Jan 10 19:22 Batch3
-rw-r--r-- 1 root root   113 Jan 10 19:13 Batch3_audit.csv
-rw-r--r-- 1 root root  3203 Jan 10 19:13 Generator_audit.csv
-rw-r--r-- 1 root root   587 Jan 10 19:26 digen_report.txt
root@3b8e72cf4001:/tmp/tpc-di# 
```

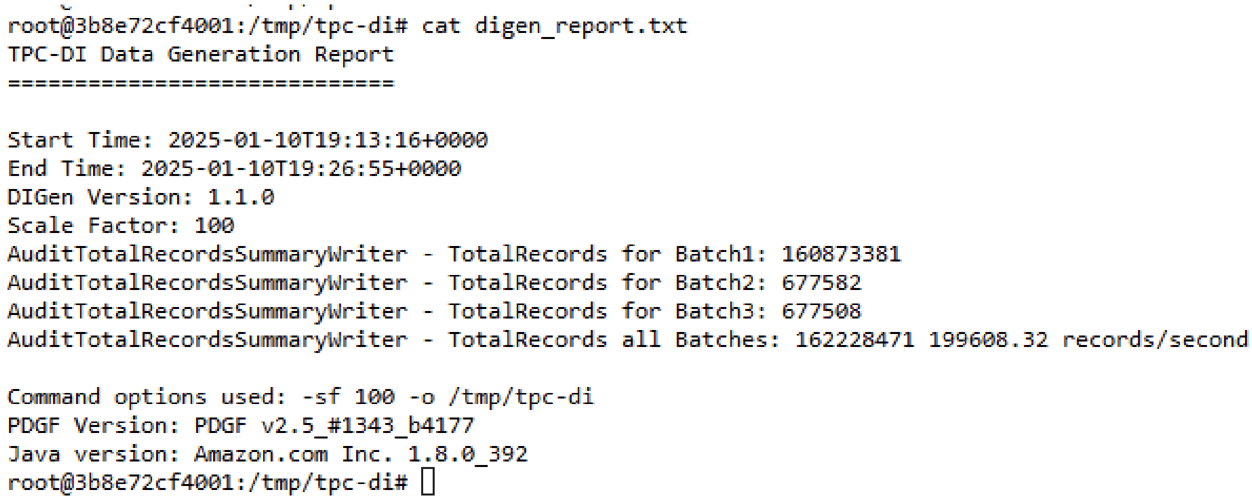


```
root@3b8e72cf4001:/tmp/tpc-di#
root@3b8e72cf4001:/tmp/tpc-di# du -h --max-depth=1
112M    ./Batch3
9.4G    ./Batch1
112M    ./Batch2
9.6G    .
root@3b8e72cf4001:/tmp/tpc-di#
```



```
FINWIRE1967Q3            FINWIRE1975Q2_audit.csv  FINWIRE1983Q2            FINWIRE1991Q1_audit.csv  FINWIRE1999Q1            FINWIRE2006Q4_audit.csv  FINWIRE2014Q4
FINWIRE1967Q3_audit.csv  FINWIRE1975Q3            FINWIRE1983Q2_audit.csv  FINWIRE1991Q2            FINWIRE1999Q1_audit.csv  FINWIRE2007Q1            FINWIRE2014Q4_audit.csv
FINWIRE1967Q4            FINWIRE1975Q3_audit.csv  FINWIRE1983Q3            FINWIRE1991Q2_audit.csv  FINWIRE1999Q2            FINWIRE2007Q1_audit.csv  FINWIRE2015Q1
FINWIRE1967Q4_audit.csv  FINWIRE1975Q4            FINWIRE1983Q3_audit.csv  FINWIRE1991Q3            FINWIRE1999Q2_audit.csv  FINWIRE2007Q2            FINWIRE2015Q1_audit.csv
FINWIRE1968Q1            FINWIRE1975Q4_audit.csv  FINWIRE1983Q4            FINWIRE1991Q3_audit.csv  FINWIRE1999Q3            FINWIRE2007Q2_audit.csv  FINWIRE2015Q2
FINWIRE1968Q1_audit.csv  FINWIRE1976Q1            FINWIRE1983Q4_audit.csv  FINWIRE1991Q4            FINWIRE1999Q3_audit.csv  FINWIRE2007Q3            FINWIRE2015Q2_audit.csv
FINWIRE1968Q2            FINWIRE1976Q1_audit.csv  FINWIRE1984Q1            FINWIRE1991Q4_audit.csv  FINWIRE1999Q4            FINWIRE2007Q3_audit.csv  FINWIRE2015Q3
FINWIRE1968Q2_audit.csv  FINWIRE1976Q2            FINWIRE1984Q1_audit.csv  FINWIRE1992Q1            FINWIRE1999Q4_audit.csv  FINWIRE2007Q4            FINWIRE2015Q3_audit.csv
FINWIRE1968Q3            FINWIRE1976Q2_audit.csv  FINWIRE1984Q2            FINWIRE1992Q1_audit.csv  FINWIRE2000Q1            FINWIRE2007Q4_audit.csv  FINWIRE2015Q4
FINWIRE1968Q3_audit.csv  FINWIRE1976Q3            FINWIRE1984Q2_audit.csv  FINWIRE1992Q2            FINWIRE2000Q1_audit.csv  FINWIRE2008Q1            FINWIRE2015Q4_audit.csv
FINWIRE1968Q4            FINWIRE1976Q3_audit.csv  FINWIRE1984Q3            FINWIRE1992Q2_audit.csv  FINWIRE2000Q2            FINWIRE2008Q1_audit.csv  FINWIRE2016Q1
FINWIRE1968Q4_audit.csv  FINWIRE1976Q4            FINWIRE1984Q3_audit.csv  FINWIRE1992Q3            FINWIRE2000Q2_audit.csv  FINWIRE2008Q2            FINWIRE2016Q1_audit.csv
FINWIRE1969Q1            FINWIRE1976Q4_audit.csv  FINWIRE1984Q4            FINWIRE1992Q3_audit.csv  FINWIRE2000Q3            FINWIRE2008Q2_audit.csv  FINWIRE2016Q2
FINWIRE1969Q1_audit.csv  FINWIRE1977Q1            FINWIRE1984Q4_audit.csv  FINWIRE1992Q4            FINWIRE2000Q3_audit.csv  FINWIRE2008Q3            FINWIRE2016Q2_audit.csv
FINWIRE1969Q2            FINWIRE1977Q1_audit.csv  FINWIRE1985Q1            FINWIRE1992Q4_audit.csv  FINWIRE2000Q4            FINWIRE2008Q3_audit.csv  FINWIRE2016Q3
FINWIRE1969Q2_audit.csv  FINWIRE1977Q2            FINWIRE1985Q1_audit.csv  FINWIRE1993Q1            FINWIRE2000Q4_audit.csv  FINWIRE2008Q4            FINWIRE2016Q3_audit.csv
FINWIRE1969Q3            FINWIRE1977Q2_audit.csv  FINWIRE1985Q2            FINWIRE1993Q1_audit.csv  FINWIRE2001Q1            FINWIRE2008Q4_audit.csv  FINWIRE2016Q4
FINWIRE1969Q3_audit.csv  FINWIRE1977Q3            FINWIRE1985Q2_audit.csv  FINWIRE1993Q2            FINWIRE2001Q1_audit.csv  FINWIRE2009Q1            FINWIRE2016Q4_audit.csv
FINWIRE1969Q4            FINWIRE1977Q3_audit.csv  FINWIRE1985Q3            FINWIRE1993Q2_audit.csv  FINWIRE2001Q2            FINWIRE2009Q1_audit.csv  FINWIRE2017Q1
FINWIRE1969Q4_audit.csv  FINWIRE1977Q4            FINWIRE1985Q3_audit.csv  FINWIRE1993Q3            FINWIRE2001Q2_audit.csv  FINWIRE2009Q2            FINWIRE2017Q1_audit.csv
FINWIRE1970Q1            FINWIRE1977Q4_audit.csv  FINWIRE1985Q4            FINWIRE1993Q3_audit.csv  FINWIRE2001Q3            FINWIRE2009Q2_audit.csv  FINWIRE2017Q2
FINWIRE1970Q1_audit.csv  FINWIRE1978Q1            FINWIRE1985Q4_audit.csv  FINWIRE1993Q4            FINWIRE2001Q3_audit.csv  FINWIRE2009Q3            FINWIRE2017Q2_audit.csv
FINWIRE1970Q2            FINWIRE1978Q1_audit.csv  FINWIRE1986Q1            FINWIRE1993Q4_audit.csv  FINWIRE2001Q4            FINWIRE2009Q3_audit.csv  FINWIRE2017Q3
FINWIRE1970Q2_audit.csv  FINWIRE1978Q2            FINWIRE1986Q1_audit.csv  FINWIRE1994Q1            FINWIRE2001Q4_audit.csv  FINWIRE2009Q4            FINWIRE2017Q3_audit.csv
FINWIRE1970Q3            FINWIRE1978Q2_audit.csv  FINWIRE1986Q2            FINWIRE1994Q1_audit.csv  FINWIRE2002Q1            FINWIRE2009Q4_audit.csv  HR.csv
FINWIRE1970Q3_audit.csv  FINWIRE1978Q3            FINWIRE1986Q2_audit.csv  FINWIRE1994Q2            FINWIRE2002Q1_audit.csv  FINWIRE2010Q1            HR_audit.csv
FINWIRE1970Q4            FINWIRE1978Q3_audit.csv  FINWIRE1986Q3            FINWIRE1994Q2_audit.csv  FINWIRE2002Q2            FINWIRE2010Q1_audit.csv  HoldingHistory.txt
FINWIRE1970Q4_audit.csv  FINWIRE1978Q4            FINWIRE1986Q3_audit.csv  FINWIRE1994Q3            FINWIRE2002Q2_audit.csv  FINWIRE2010Q2            HoldingHistory_audit.csv
FINWIRE1971Q1            FINWIRE1978Q4_audit.csv  FINWIRE1986Q4            FINWIRE1994Q3_audit.csv  FINWIRE2002Q3            FINWIRE2010Q2_audit.csv  Industry.txt
FINWIRE1971Q1_audit.csv  FINWIRE1979Q1            FINWIRE1986Q4_audit.csv  FINWIRE1994Q4            FINWIRE2002Q3_audit.csv  FINWIRE2010Q3            Industry_audit.csv
FINWIRE1971Q2            FINWIRE1979Q1_audit.csv  FINWIRE1987Q1            FINWIRE1994Q4_audit.csv  FINWIRE2002Q4            FINWIRE2010Q3_audit.csv  Prospect.csv
FINWIRE1971Q2_audit.csv  FINWIRE1979Q2            FINWIRE1987Q1_audit.csv  FINWIRE1995Q1            FINWIRE2002Q4_audit.csv  FINWIRE2010Q4            Prospect_audit.csv
FINWIRE1971Q3            FINWIRE1979Q2_audit.csv  FINWIRE1987Q2            FINWIRE1995Q1_audit.csv  FINWIRE2003Q1            FINWIRE2010Q4_audit.csv  StatusType.txt
FINWIRE1971Q3_audit.csv  FINWIRE1979Q3            FINWIRE1987Q2_audit.csv  FINWIRE1995Q2            FINWIRE2003Q1_audit.csv  FINWIRE2011Q1            StatusType_audit.csv
FINWIRE1971Q4            FINWIRE1979Q3_audit.csv  FINWIRE1987Q3            FINWIRE1995Q2_audit.csv  FINWIRE2003Q2            FINWIRE2011Q1_audit.csv  TaxRate.txt
FINWIRE1971Q4_audit.csv  FINWIRE1979Q4            FINWIRE1987Q3_audit.csv  FINWIRE1995Q3            FINWIRE2003Q2_audit.csv  FINWIRE2011Q2            TaxRate_audit.csv
FINWIRE1972Q1            FINWIRE1979Q4_audit.csv  FINWIRE1987Q4            FINWIRE1995Q3_audit.csv  FINWIRE2003Q3            FINWIRE2011Q2_audit.csv  Time.txt
FINWIRE1972Q1_audit.csv  FINWIRE1980Q1            FINWIRE1987Q4_audit.csv  FINWIRE1995Q4            FINWIRE2003Q3_audit.csv  FINWIRE2011Q3            Time_audit.csv
FINWIRE1972Q2            FINWIRE1980Q1_audit.csv  FINWIRE1988Q1            FINWIRE1995Q4_audit.csv  FINWIRE2003Q4            FINWIRE2011Q3_audit.csv  Trade.txt
FINWIRE1972Q2_audit.csv  FINWIRE1980Q2            FINWIRE1988Q1_audit.csv  FINWIRE1996Q1            FINWIRE2003Q4_audit.csv  FINWIRE2011Q4            TradeHistory.txt
FINWIRE1972Q3            FINWIRE1980Q2_audit.csv  FINWIRE1988Q2            FINWIRE1996Q1_audit.csv  FINWIRE2004Q1            FINWIRE2011Q4_audit.csv  TradeHistory_audit.csv
FINWIRE1972Q3_audit.csv  FINWIRE1980Q3            FINWIRE1988Q2_audit.csv  FINWIRE1996Q2            FINWIRE2004Q1_audit.csv  FINWIRE2012Q1            TradeType.txt
FINWIRE1972Q4            FINWIRE1980Q3_audit.csv  FINWIRE1988Q3            FINWIRE1996Q2_audit.csv  FINWIRE2004Q2            FINWIRE2012Q1_audit.csv  TradeType_audit.csv
FINWIRE1972Q4_audit.csv  FINWIRE1980Q4            FINWIRE1988Q3_audit.csv  FINWIRE1996Q3            FINWIRE2004Q2_audit.csv  FINWIRE2012Q2            Trade_audit.csv
FINWIRE1973Q1            FINWIRE1980Q4_audit.csv  FINWIRE1988Q4            FINWIRE1996Q3_audit.csv  FINWIRE2004Q3            FINWIRE2012Q2_audit.csv  WatchHistory.txt
FINWIRE1973Q1_audit.csv  FINWIRE1981Q1            FINWIRE1988Q4_audit.csv  FINWIRE1996Q4            FINWIRE2004Q3_audit.csv  FINWIRE2012Q3            WatchHistory_audit.csv
FINWIRE1973Q2            FINWIRE1981Q1_audit.csv  FINWIRE1989Q1            FINWIRE1996Q4_audit.csv  FINWIRE2004Q4            FINWIRE2012Q3_audit.csv
FINWIRE1973Q2_audit.csv  FINWIRE1981Q2            FINWIRE1989Q1_audit.csv  FINWIRE1997Q1            FINWIRE2004Q4_audit.csv  FINWIRE2012Q4
FINWIRE1973Q3            FINWIRE1981Q2_audit.csv  FINWIRE1989Q2            FINWIRE1997Q1_audit.csv  FINWIRE2005Q1            FINWIRE2012Q4_audit.csv
FINWIRE1973Q3_audit.csv  FINWIRE1981Q3            FINWIRE1989Q2_audit.csv  FINWIRE1997Q2            FINWIRE2005Q1_audit.csv  FINWIRE2013Q1
root@3b8e72cf4001:/tmp/tpc-di# ls Batch1 | wc
    437     437    8234
root@3b8e72cf4001:/tmp/tpc-di#
root@3b8e72cf4001:/tmp/tpc-di# ls Batch1 | wc -l
437
root@3b8e72cf4001:/tmp/tpc-di# 
```

W pliku digen_report.txt dostajemy informacje ile rekordów jest wygenerowanych dla każdego batcha.

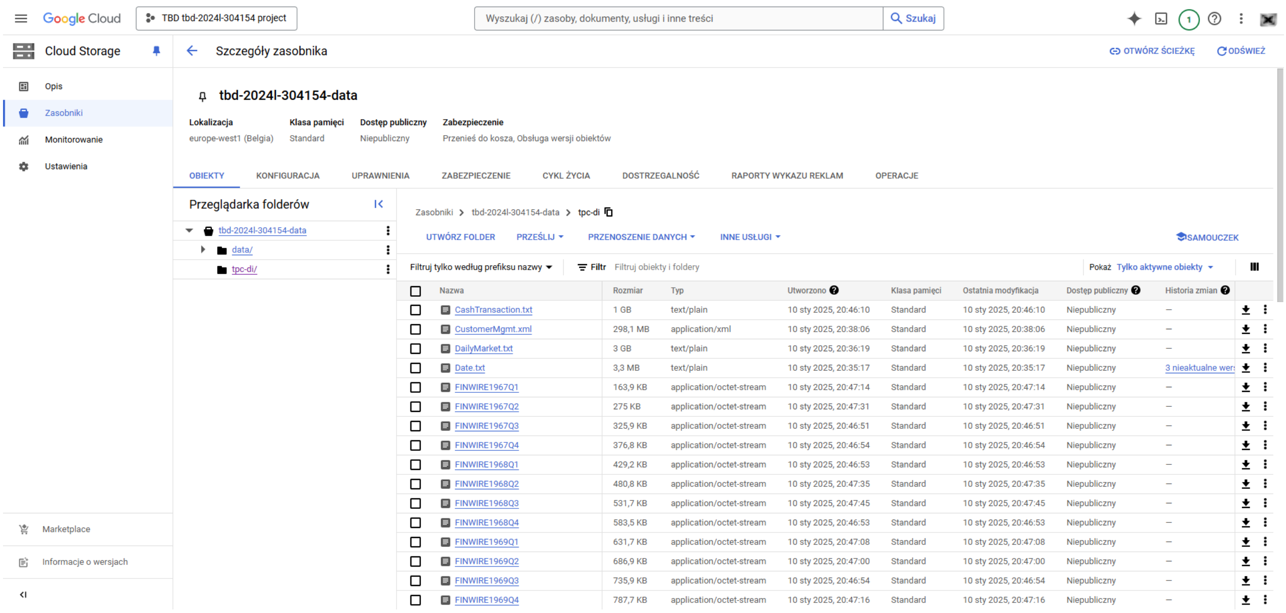


```
root@3b8e72cf4001:/tmp/tpc-di# cat digen_report.txt
TPC-DI Data Generation Report
=============================

Start Time: 2025-01-10T19:13:16+0000
End Time: 2025-01-10T19:26:55+0000
DIGen Version: 1.1.0
Scale Factor: 100
AuditTotalRecordsSummaryWriter - TotalRecords for Batch1: 160873381
AuditTotalRecordsSummaryWriter - TotalRecords for Batch2: 677582
AuditTotalRecordsSummaryWriter - TotalRecords for Batch3: 677508
AuditTotalRecordsSummaryWriter - TotalRecords all Batches: 162228471 199608.32 records/second

Command options used: -sf 100 -o /tmp/tpc-di
PDGF Version: PDGF v2.5_#1343_b4177
Java version: Amazon.com Inc. 1.8.0_392
root@3b8e72cf4001:/tmp/tpc-di# 
```

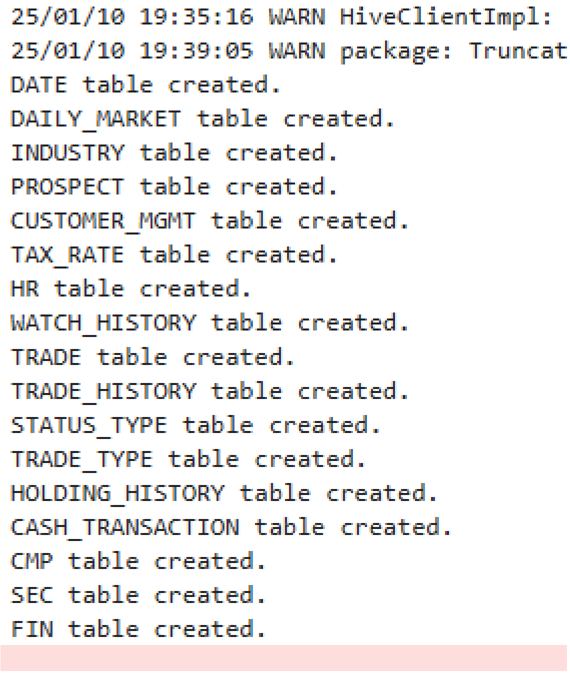
8. Analyze tpcdi.py. What happened in the loading stage?



tbd-2024l-304154-data

Lokalizacja	Klasa pamięci	Dostęp publiczny	Zabezpieczenie
europe-west1 (Belgia)	Standard	Niepubliczny	Przenieś do kosza, Obsługa wersji obiektów

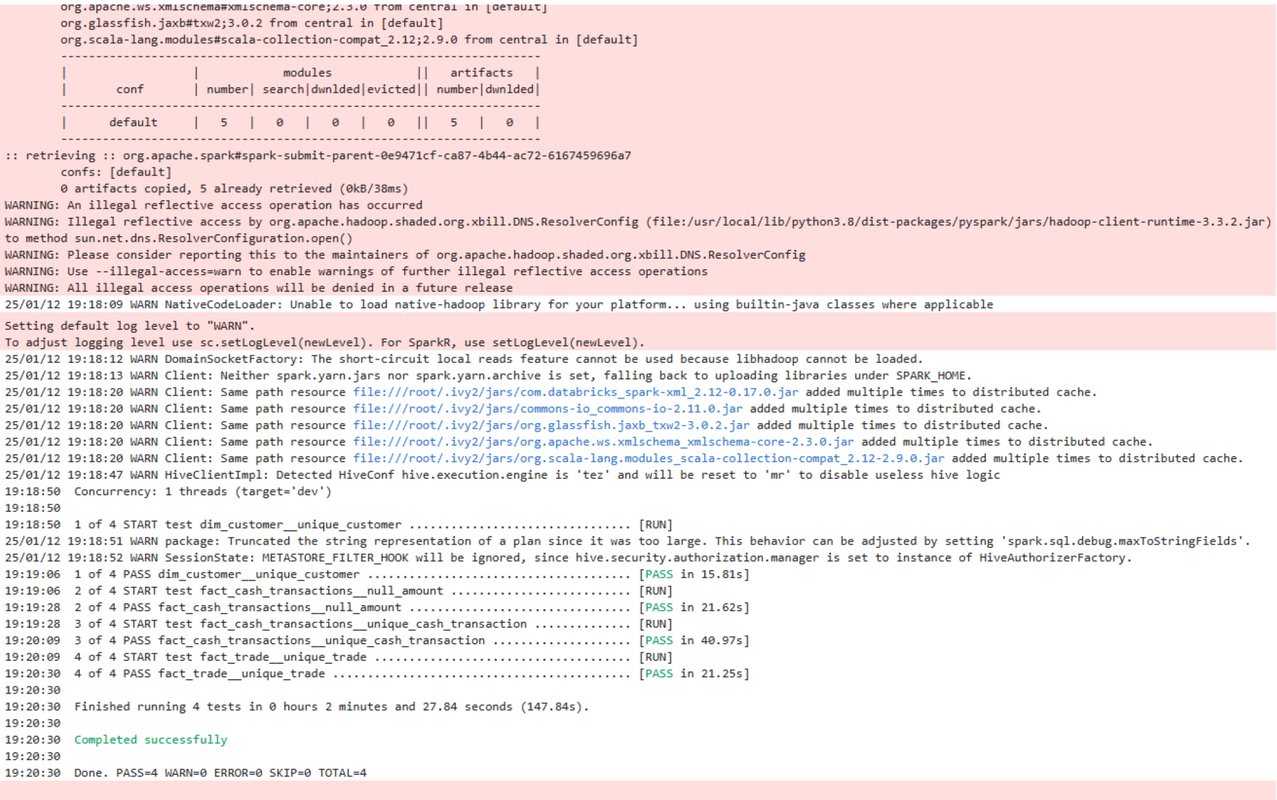
□	Nazwa	Rozmiar	Typ	Utworzono	Klasa pamięci	Ostatnia modyfikacja	Dostęp publiczny	Historia zmian
□	CashTransaction.txt	1 GB	text/plain	10 sty 2025, 20:46:10	Standard	10 sty 2025, 20:46:10	Niepubliczny	–
□	CustomerMgmt.xml	298,1 MB	application/xml	10 sty 2025, 20:38:06	Standard	10 sty 2025, 20:38:06	Niepubliczny	–
□	DailyMarket.txt	3 GB	text/plain	10 sty 2025, 20:36:19	Standard	10 sty 2025, 20:36:19	Niepubliczny	–
□	Date.txt	3,3 MB	text/plain	10 sty 2025, 20:35:17	Standard	10 sty 2025, 20:35:17	Niepubliczny	3 nieaktualne wer
□	FINWIRE1967Q1	163,9 KB	application/octet-stream	10 sty 2025, 20:47:14	Standard	10 sty 2025, 20:47:14	Niepubliczny	–
□	FINWIRE1967Q2	275 KB	application/octet-stream	10 sty 2025, 20:47:31	Standard	10 sty 2025, 20:47:31	Niepubliczny	–
□	FINWIRE1967Q3	325,9 KB	application/octet-stream	10 sty 2025, 20:46:51	Standard	10 sty 2025, 20:46:51	Niepubliczny	–
□	FINWIRE1967Q4	376,8 KB	application/octet-stream	10 sty 2025, 20:46:54	Standard	10 sty 2025, 20:46:54	Niepubliczny	–
□	FINWIRE1968Q1	429,2 KB	application/octet-stream	10 sty 2025, 20:46:53	Standard	10 sty 2025, 20:46:53	Niepubliczny	–
□	FINWIRE1968Q2	480,8 KB	application/octet-stream	10 sty 2025, 20:47:35	Standard	10 sty 2025, 20:47:35	Niepubliczny	–
□	FINWIRE1968Q3	531,7 KB	application/octet-stream	10 sty 2025, 20:47:45	Standard	10 sty 2025, 20:47:45	Niepubliczny	–
□	FINWIRE1968Q4	583,5 KB	application/octet-stream	10 sty 2025, 20:46:53	Standard	10 sty 2025, 20:46:53	Niepubliczny	–
□	FINWIRE1969Q1	631,7 KB	application/octet-stream	10 sty 2025, 20:47:08	Standard	10 sty 2025, 20:47:08	Niepubliczny	–
□	FINWIRE1969Q2	686,9 KB	application/octet-stream	10 sty 2025, 20:47:00	Standard	10 sty 2025, 20:47:00	Niepubliczny	–
□	FINWIRE1969Q3	735,9 KB	application/octet-stream	10 sty 2025, 20:46:54	Standard	10 sty 2025, 20:46:54	Niepubliczny	–
□	FINWIRE1969Q4	787,7 KB	application/octet-stream	10 sty 2025, 20:47:16	Standard	10 sty 2025, 20:47:16	Niepubliczny	–



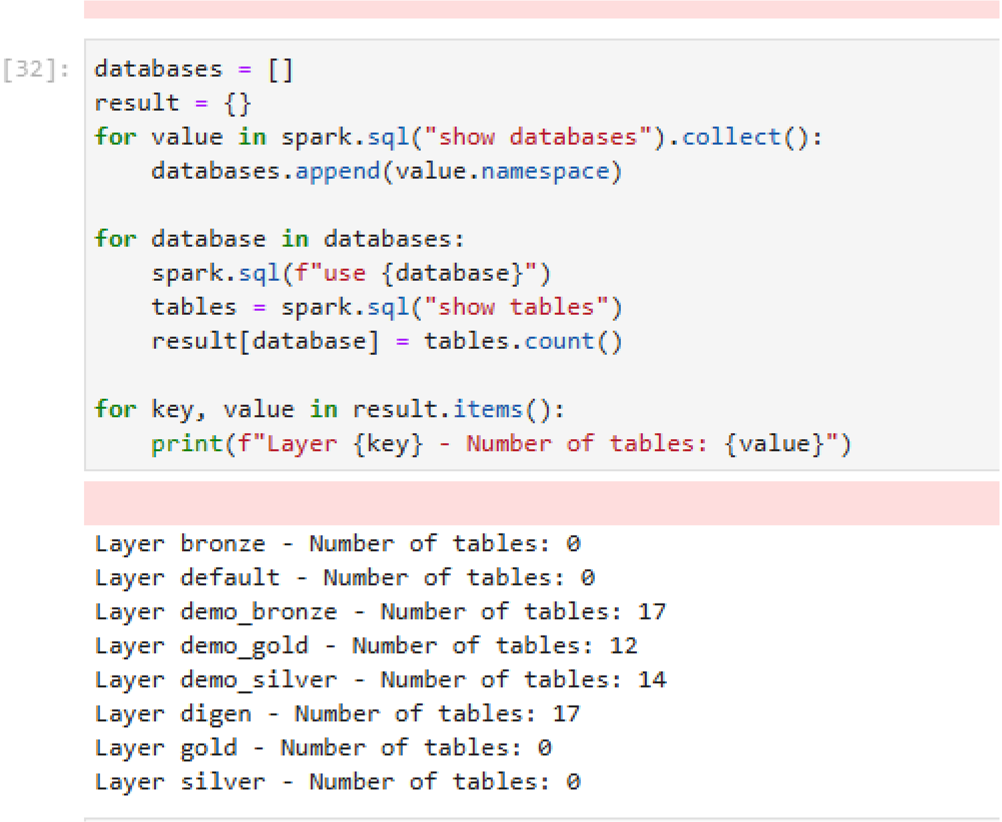
```
25/01/10 19:35:16 WARN HiveClientImpl:
25/01/10 19:39:05 WARN package: Truncat
DATE table created.
DAILY_MARKET table created.
INDUSTRY table created.
PROSPECT table created.
CUSTOMER_MGMT table created.
TAX_RATE table created.
HR table created.
WATCH_HISTORY table created.
TRADE table created.
TRADE_HISTORY table created.
STATUS_TYPE table created.
TRADE_TYPE table created.
HOLDING_HISTORY table created.
CASH_TRANSACTION table created.
CMP table created.
SEC table created.
FIN table created.
```

W tym kroku dane zostały przesłane do cloud storage.

9. Using SparkSQL answer: how many table were created in each layer?



```
        org.apache.ws.xmlschema#xmlschema-core;2.3.0 from central in [default]
        org.glassfish.jaxb#txw2;3.0.2 from central in [default]
        org.scala-lang.modules#scala-collection-compat_2.12;2.9.0 from central in [default]
        ---------------------------------------------------------------------
        |                  |            modules            ||   artifacts   |
        |       conf       | number| search|dwnlded|evicted|| number|dwnlded|
        ---------------------------------------------------------------------
        |      default     |   5   |   0   |   0   |   0   ||   5   |   0   |
        ---------------------------------------------------------------------
:: retrieving :: org.apache.spark#spark-submit-parent-0e9471cf-ca87-4b44-ac72-6167459696a7
        confs: [default]
        0 artifacts copied, 5 already retrieved (0kB/38ms)
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.shaded.org.xbill.DNS.ResolverConfig (file:/usr/local/lib/python3.8/dist-packages/pyspark/jars/hadoop-client-runtime-3.3.2.jar)
to method sun.net.dns.ResolverConfiguration.open()
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.shaded.org.xbill.DNS.ResolverConfig
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
25/01/12 19:18:09 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
25/01/12 19:18:12 WARN DomainSocketFactory: The short-circuit local reads feature cannot be used because libhadoop cannot be loaded.
25/01/12 19:18:13 WARN Client: Neither spark.yarn.jars nor spark.yarn.archive is set, falling back to uploading libraries under SPARK_HOME.
25/01/12 19:18:20 WARN Client: Same path resource file:///root/.ivy2/jars/com.databricks_spark-xml_2.12-0.17.0.jar added multiple times to distributed cache.
25/01/12 19:18:20 WARN Client: Same path resource file:///root/.ivy2/jars/commons-io_commons-io-2.11.0.jar added multiple times to distributed cache.
25/01/12 19:18:20 WARN Client: Same path resource file:///root/.ivy2/jars/org.glassfish.jaxb_txw2-3.0.2.jar added multiple times to distributed cache.
25/01/12 19:18:20 WARN Client: Same path resource file:///root/.ivy2/jars/org.apache.ws.xmlschema_xmlschema-core-2.3.0.jar added multiple times to distributed cache.
25/01/12 19:18:20 WARN Client: Same path resource file:///root/.ivy2/jars/org.scala-lang.modules_scala-collection-compat_2.12-2.9.0.jar added multiple times to distributed cache.
25/01/12 19:18:47 WARN HiveClientImpl: Detected HiveConf hive.execution.engine is 'tez' and will be reset to 'mr' to disable useless hive logic
19:18:50  Concurrency: 1 threads (target='dev')
19:18:50
19:18:50  1 of 4 START test dim_customer__unique_customer ............................... [RUN]
25/01/12 19:18:51 WARN package: Truncated the string representation of a plan since it was too large. This behavior can be adjusted by setting 'spark.sql.debug.maxToStringFields'.
25/01/12 19:18:52 WARN SessionState: METASTORE_FILTER_HOOK will be ignored, since hive.security.authorization.manager is set to instance of HiveAuthorizerFactory.
19:19:06  1 of 4 PASS dim_customer__unique_customer ..................................... [PASS in 15.81s]
19:19:06  2 of 4 START test fact_cash_transactions__null_amount ......................... [RUN]
19:19:28  2 of 4 PASS fact_cash_transactions__null_amount ............................... [PASS in 21.62s]
19:19:28  3 of 4 START test fact_cash_transactions__unique_cash_transaction ............. [RUN]
19:20:09  3 of 4 PASS fact_cash_transactions__unique_cash_transaction ................... [PASS in 40.97s]
19:20:09  4 of 4 START test fact_trade__unique_trade .................................... [RUN]
19:20:30  4 of 4 PASS fact_trade__unique_trade .......................................... [PASS in 21.25s]
19:20:30
19:20:30  Finished running 4 tests in 0 hours 2 minutes and 27.84 seconds (147.84s).
19:20:30
19:20:30  Completed successfully
19:20:30
19:20:30  Done. PASS=4 WARN=0 ERROR=0 SKIP=0 TOTAL=4
```



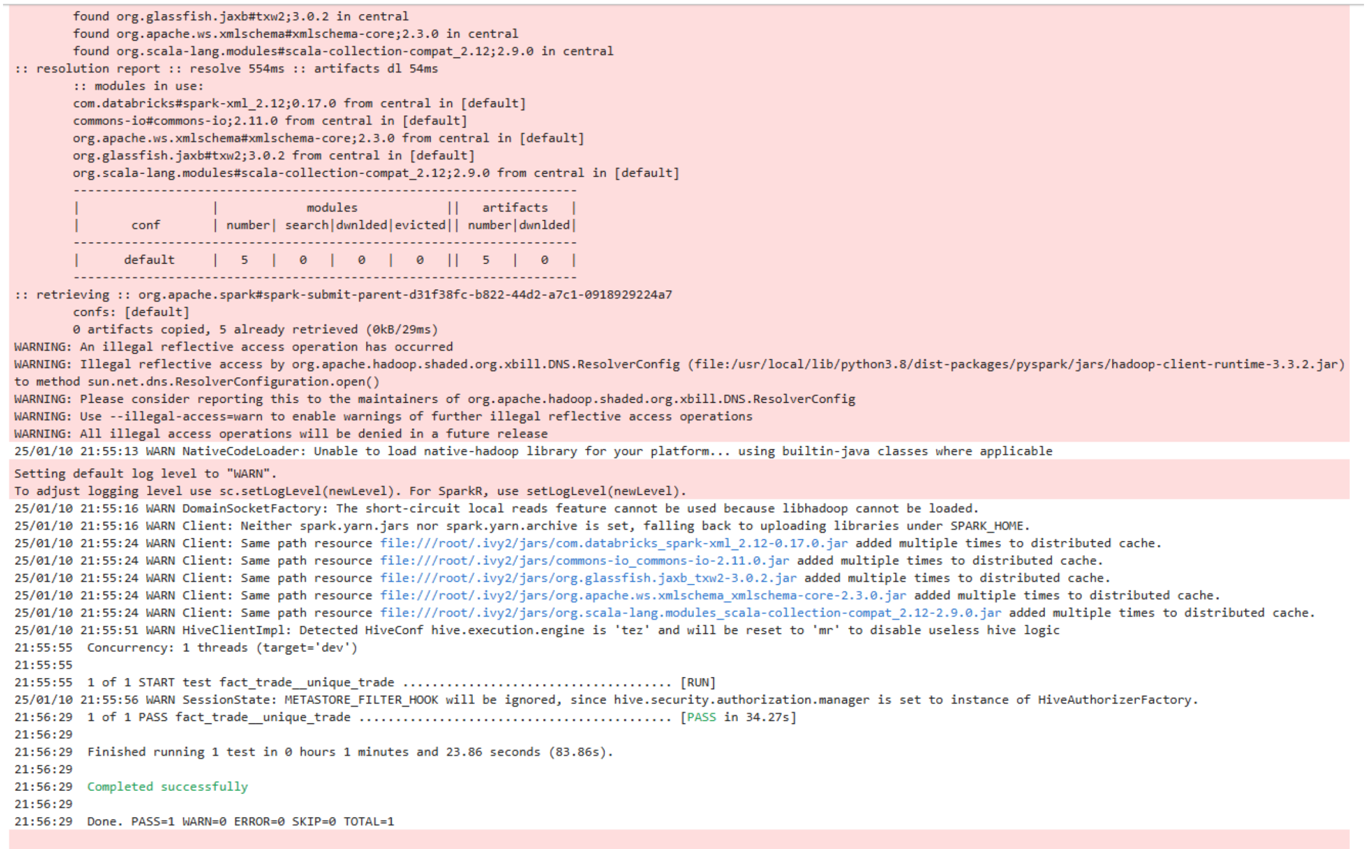
```
[32]: databases = []
      result = {}
      for value in spark.sql("show databases").collect():
          databases.append(value.namespace)

      for database in databases:
          spark.sql(f"use {database}")
          tables = spark.sql("show tables")
          result[database] = tables.count()

      for key, value in result.items():
          print(f"Layer {key} - Number of tables: {value}")
```

```
Layer bronze - Number of tables: 0
Layer default - Number of tables: 0
Layer demo_bronze - Number of tables: 17
Layer demo_gold - Number of tables: 12
Layer demo_silver - Number of tables: 14
Layer digen - Number of tables: 17
Layer gold - Number of tables: 0
Layer silver - Number of tables: 0
```

10. Add some 3 more dbt tests and explain what you are testing.



```
        found org.glassfish.jaxb#txw2;3.0.2 in central
        found org.apache.ws.xmlschema#xmlschema-core;2.3.0 in central
        found org.scala-lang.modules#scala-collection-compat_2.12;2.9.0 in central
:: resolution report :: resolve 554ms :: artifacts dl 54ms
        :: modules in use:
        com.databricks#spark-xml_2.12;0.17.0 from central in [default]
        commons-io#commons-io;2.11.0 from central in [default]
        org.apache.ws.xmlschema#xmlschema-core;2.3.0 from central in [default]
        org.glassfish.jaxb#txw2;3.0.2 from central in [default]
        org.scala-lang.modules#scala-collection-compat_2.12;2.9.0 from central in [default]
        ---------------------------------------------------------------------
        |                  |            modules            ||   artifacts   |
        |       conf       | number| search|dwnlded|evicted|| number|dwnlded|
        ---------------------------------------------------------------------
        |      default     |   5   |   0   |   0   |   0   ||   5   |   0   |
        ---------------------------------------------------------------------
:: retrieving :: org.apache.spark#spark-submit-parent-d31f38fc-b822-44d2-a7c1-0918929224a7
        confs: [default]
        0 artifacts copied, 5 already retrieved (0kB/29ms)
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.shaded.org.xbill.DNS.ResolverConfig (file:/usr/local/lib/python3.8/dist-packages/pyspark/jars/hadoop-client-runtime-3.3.2.jar)
to method sun.net.dns.ResolverConfiguration.open()
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.shaded.org.xbill.DNS.ResolverConfig
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
25/01/10 21:55:13 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
25/01/10 21:55:16 WARN DomainSocketFactory: The short-circuit local reads feature cannot be used because libhadoop cannot be loaded.
25/01/10 21:55:16 WARN Client: Neither spark.yarn.jars nor spark.yarn.archive is set, falling back to uploading libraries under SPARK_HOME.
25/01/10 21:55:24 WARN Client: Same path resource file:///root/.ivy2/jars/com.databricks_spark-xml_2.12-0.17.0.jar added multiple times to distributed cache.
25/01/10 21:55:24 WARN Client: Same path resource file:///root/.ivy2/jars/commons-io_commons-io-2.11.0.jar added multiple times to distributed cache.
25/01/10 21:55:24 WARN Client: Same path resource file:///root/.ivy2/jars/org.glassfish.jaxb_txw2-3.0.2.jar added multiple times to distributed cache.
25/01/10 21:55:24 WARN Client: Same path resource file:///root/.ivy2/jars/org.apache.ws.xmlschema_xmlschema-core-2.3.0.jar added multiple times to distributed cache.
25/01/10 21:55:24 WARN Client: Same path resource file:///root/.ivy2/jars/org.scala-lang.modules_scala-collection-compat_2.12-2.9.0.jar added multiple times to distributed cache.
25/01/10 21:55:51 WARN HiveClientImpl: Detected HiveConf hive.execution.engine is 'tez' and will be reset to 'mr' to disable useless hive logic
21:55:55  Concurrency: 1 threads (target='dev')
21:55:55
21:55:55  1 of 1 START test fact_trade__unique_trade .................................... [RUN]
25/01/10 21:55:56 WARN SessionState: METASTORE_FILTER_HOOK will be ignored, since hive.security.authorization.manager is set to instance of HiveAuthorizerFactory.
21:56:29  1 of 1 PASS fact_trade__unique_trade .......................................... [PASS in 34.27s]
21:56:29
21:56:29  Finished running 1 test in 0 hours 1 minutes and 23.86 seconds (83.86s).
21:56:29
21:56:29  Completed successfully
21:56:29
21:56:29  Done. PASS=1 WARN=0 ERROR=0 SKIP=0 TOTAL=1
```

airflow_monitoring – sprawdza ogólną kondycję środowiska Airflow (np. czy zadania startują i wykonują się zgodnie z harmonogramem).

composer_sample_dbt_task – demonstruje integrację z DBT (Data Build Tool), weryfikując, czy można poprawnie uruchamiać i monitorować procesy przetwarzania danych przy użyciu DBT w Airflow.

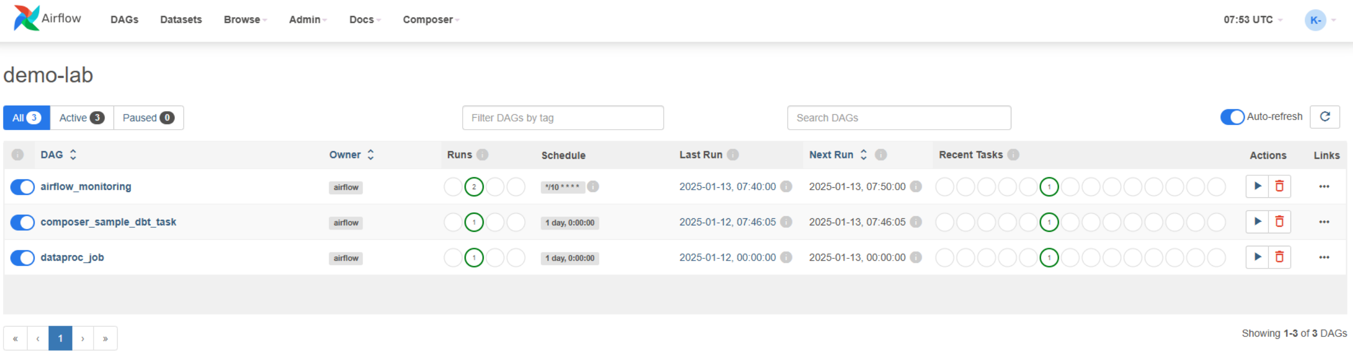
dataproc_job – testuje możliwość zlecania zadań do Dataproc, czyli czy Airflow potrafi prawidłowo nawiązać komunikację i uruchamiać joby w klastrze Dataproc.

11. In main.tf update

```
dbt_git_repo                = "https://github.com/mwiewior/tbd-tpc-di.git"
dbt_git_repo_branch         = "main"
```

so dbt_git_repo points to your fork of tbd-tpc-di.

12. Redeploy infrastructure and check if the DAG finished with no errors:



demo-lab

DAG	Owner	Runs	Schedule	Last Run	Next Run	Recent Tasks	Actions	Links
airflow_monitoring	airflow	2	*/10 * * * *	2025-01-13, 07:40:00	2025-01-13, 07:50:00	1	▶ 🗑	•••
composer_sample_dbt_task	airflow	1	1 day, 0:00:00	2025-01-12, 07:46:05	2025-01-13, 07:46:05	1	▶ 🗑	•••
dataproc_job	airflow	1	1 day, 0:00:00	2025-01-12, 00:00:00	2025-01-13, 00:00:00	1	▶ 🗑	•••

Showing 1-3 of 3 DAGs