idealista

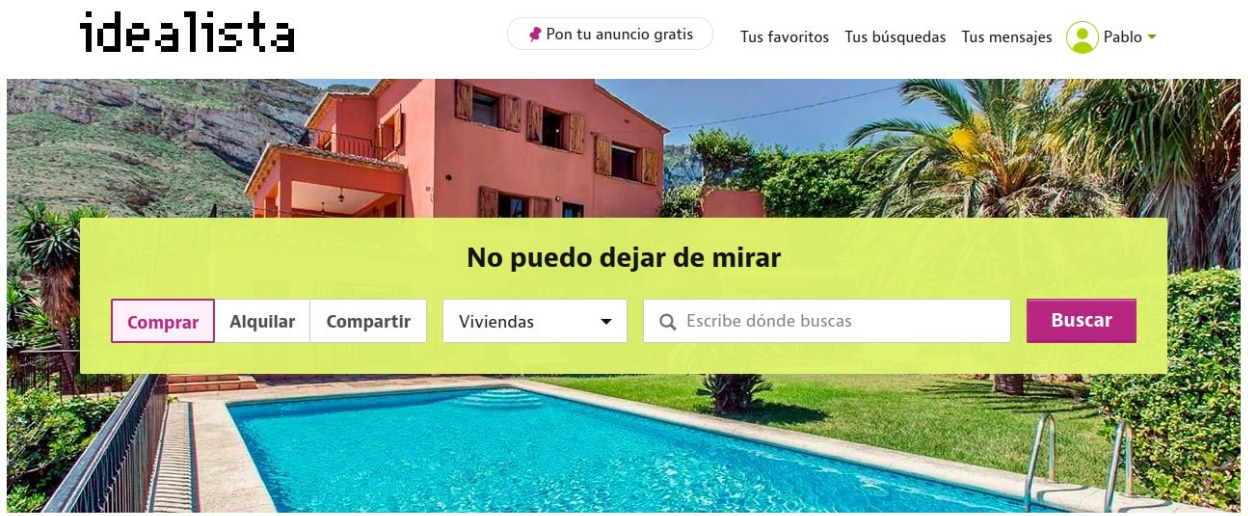
Great Expectations: A ring to rule them all

Pablo Lozano Santiuste

Royal Society - Benders


MIGUEL DE CERVANTES
DON QUIJOTE DE LA MANCHA
AUSTRALCÓMIC

idealista



Tus búsquedas

Viviendas en tu zona personalizada

hasta 220.000€

2 filtros más

Búsqueda guardada

Dibujar tu zona de búsqueda

Elegir exactamente la zona que estás buscando sobre un mapa.

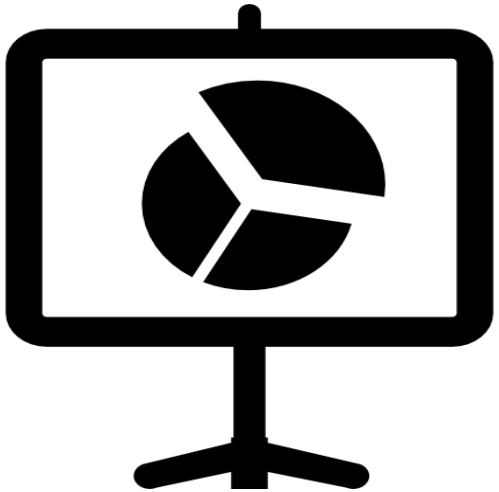
Empezar a dibujar tu búsqueda

Publicar tu inmueble gratis

Tus 2 primeros anuncios son gratis. Casas, habitaciones, oficinas... ¡Todo cabe!

Poner tu anuncio gratis

Talk Structure



+

https://github.com/idealista/great_expectations_talk

idealista

Data test

- movies_metadata
  - adult
  - belongs_to_collection
  - budget
  - genres
  - homepage
  - id
  - imdb_id
  - original_language
  - original_title
  - overview
  - popularity
  - poster_path
  - production_companies
  - production_countries
  - release_date
  - revenue
  - runtime
  - spoken_languages
  - status
  - tagline
  - title
  - video
  - vote_average
  - vote_count

Data Quality – Why I’ts important?

Data Quality focus on Data Assets regardless of how the data was generated.

- **Completeness**
- **Timeless**
- **Validity**
- **Integrity**
- **Uniqueness**
- **Consistency**

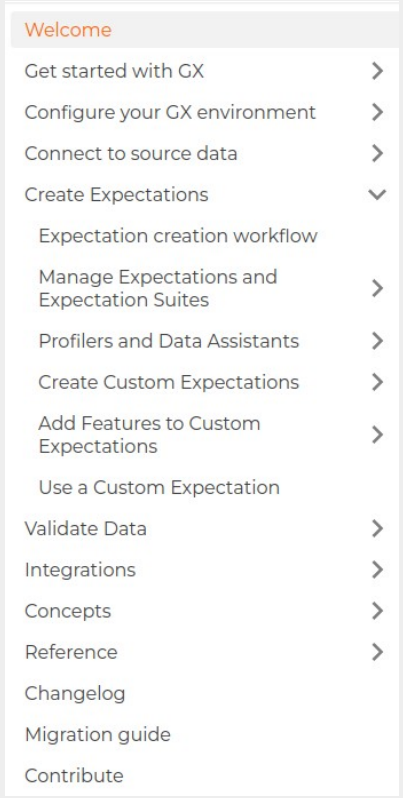
Use cases

- **Business: If we are buying data from external providers**
- **Business: If we are selling data to external providers**
- **If we are ingesting data into our domain**
- **If we depends strongly from other domains**
- **If we scrap websites :)**

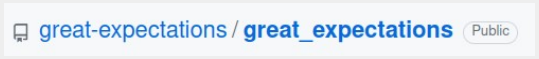
idealista

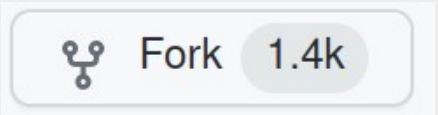
Great Expectations

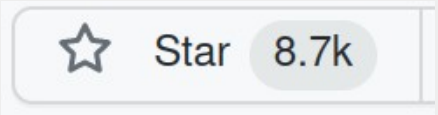
Documentation



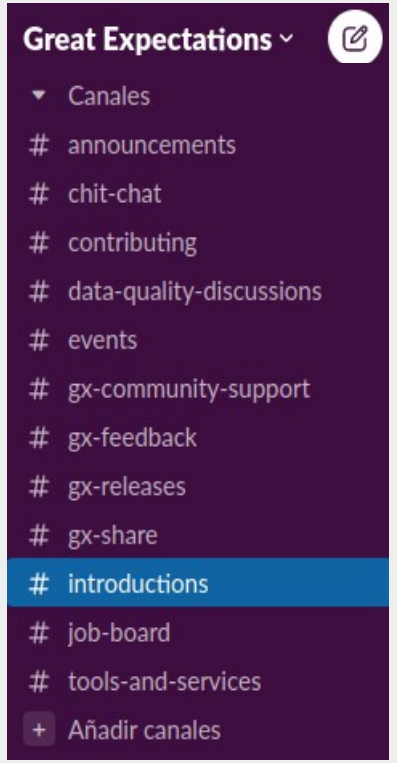
Github







Slack



idealista

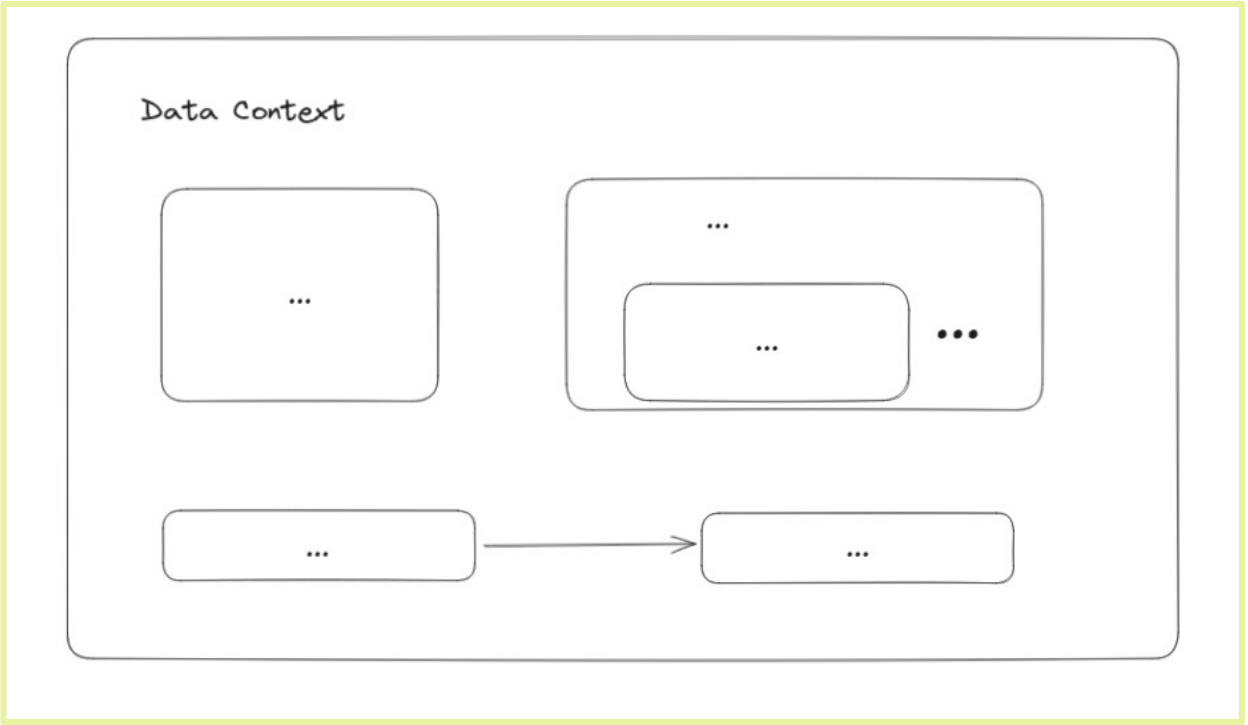
¿Como instalamos la herramienta?

pyproject.toml

```toml
[tool.poetry]
name = "great-expectations-talk"
version = "0.1.0"
description = ""
authors = ["Pablo Lozano Santiuste <plozano94@gmail.com>"]
readme = "README.md"

[tool.poetry.scripts]
accidentes_checkpoint = "great_expectations_talk.checkpoint.accidentes_checkpoint:run"

[tool.poetry.dependencies]
python = "^3.9"
great-expectations = "^0.17.14"
sqlalchemy = "^1"
clickhouse-sqlalchemy = "^0.2.4"
psycopg2 = "^2.9.7"
click = "^8.1.7"


[[tool.poetry.source]]
name = "id"
url = "http://nexus.int.sys.idealista/repository/pypi-idealista/"
priority = "default"

[build-system]
requires = ["poetry-core"]
build-backend = "poetry.core.masonry.api"
```

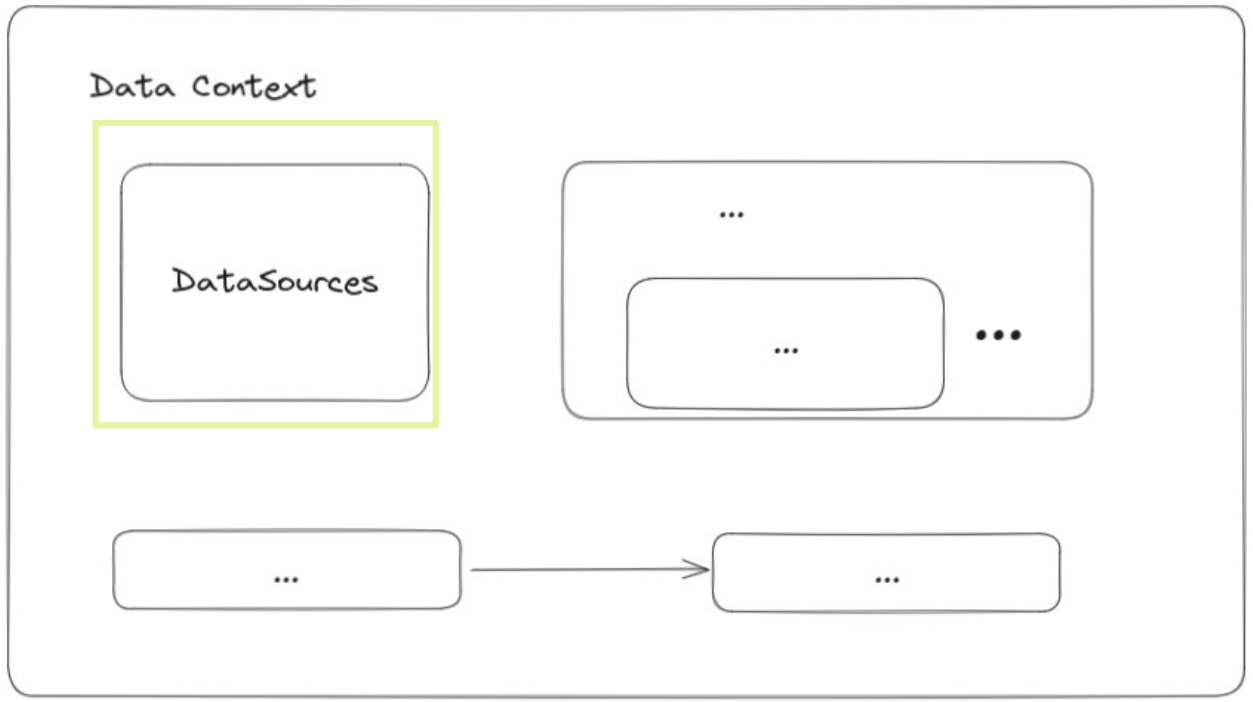
Glossary

- Data Context
  - EphemeralDataContext
  - FileDataContext
  - CloudDataContext



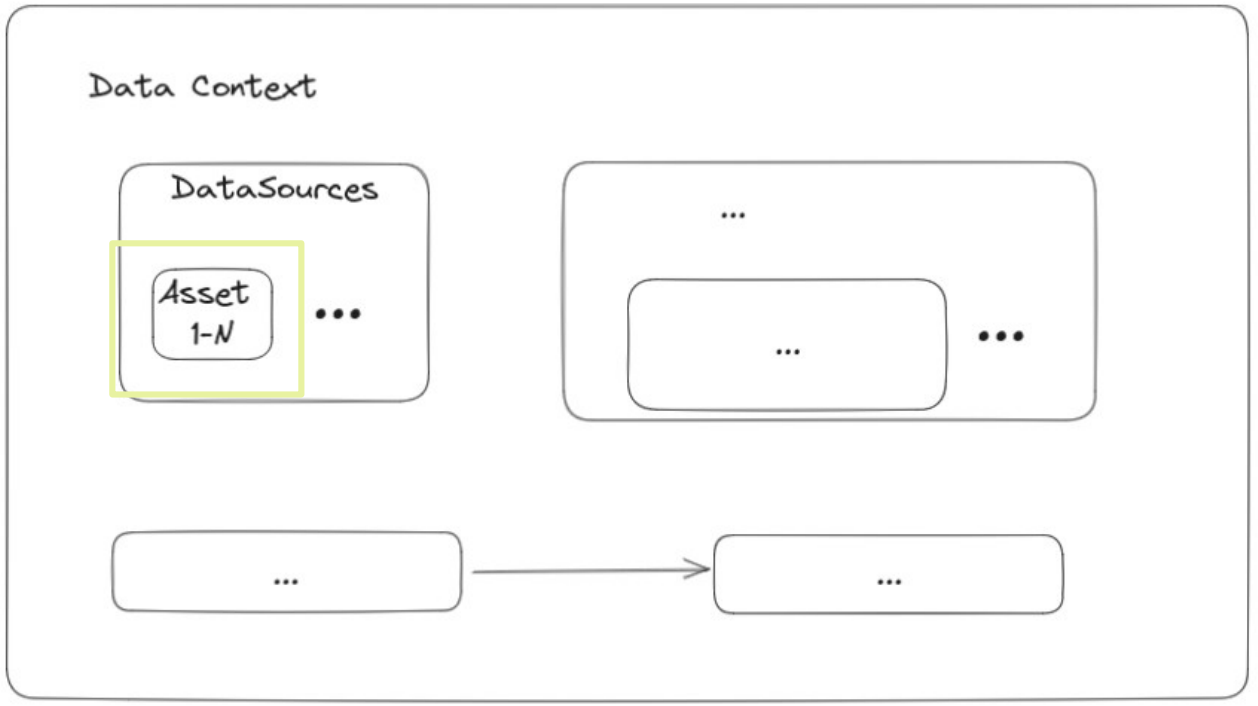
Glossary

- Data Sources
  - AWS Athena
  - BigQuery
  - MSSQL
  - MySQL
  - PostgreSQL
  - Redshift
  - Snowflake
  - SQLite
  - Trino
  - ClickHouse
  - Potentially Any SQLALchemy compatible



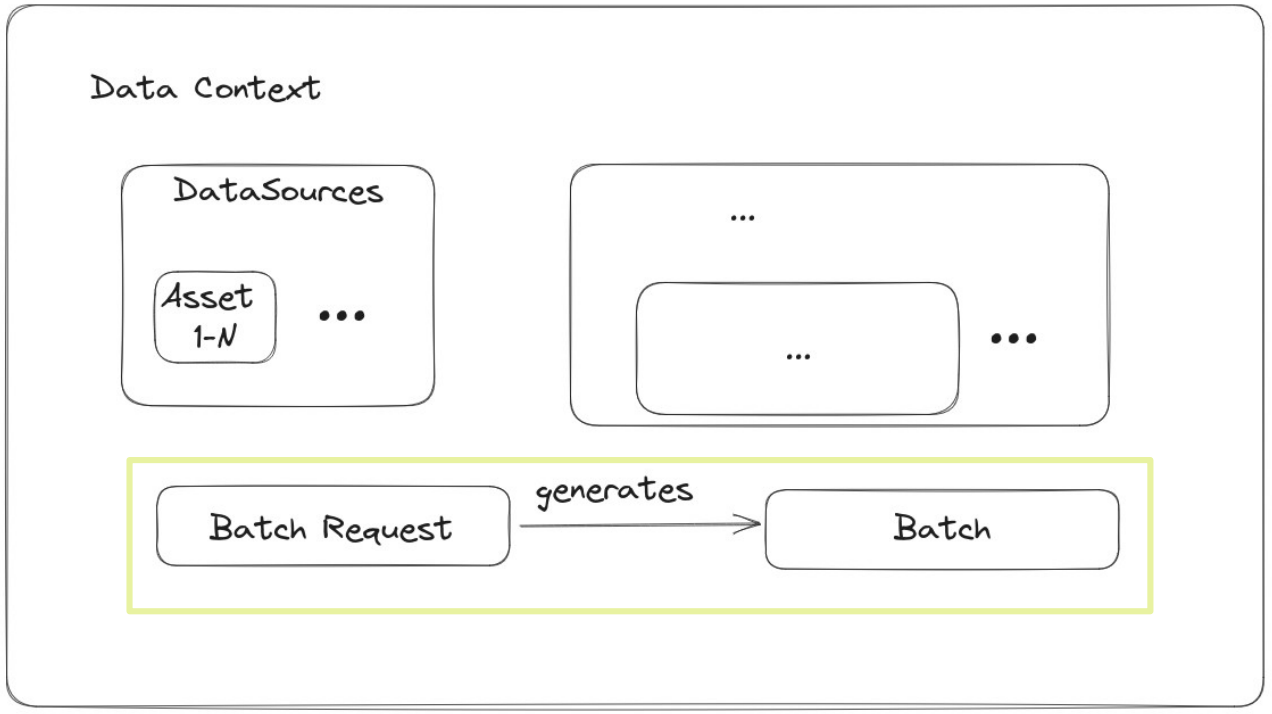
Glossary

- **Data Assets → Basically tables in GX world**
  - **Splitters**
  - **Sorters**



Glossary

- **Batch Requests:** A Batch Request contains all the necessary details to query the appropriate underlying data.
- You could fetch only last ingested data in an incremental daily table and validate only that part.
- Also you could only fetch a partition of the table in order to auto generate expectations (ALERT!!! SPOILER)

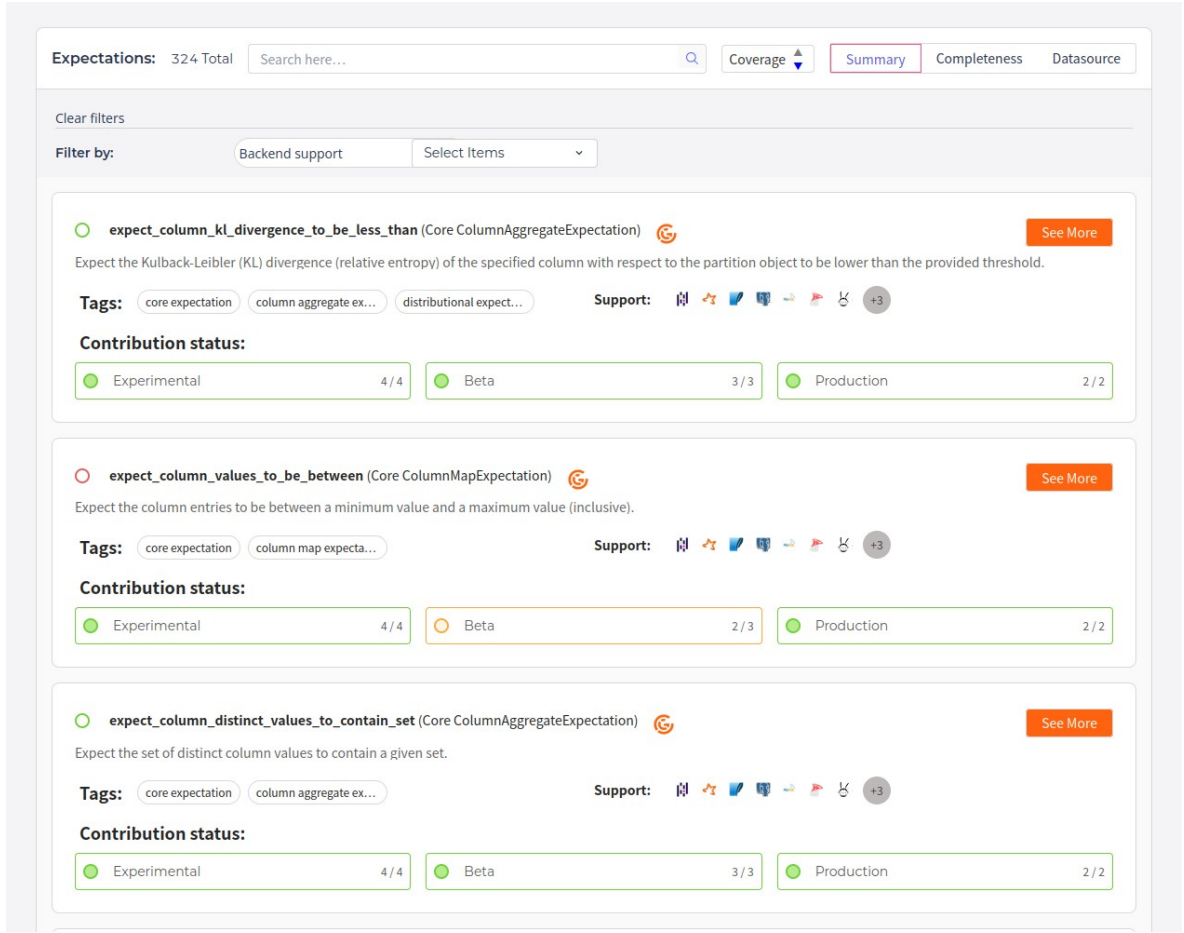


Glossary

- Expectations
  - JSON based
  - Common parameters:
    - Columns involved
    - Threshold values
    - Others

Glossary

Expectations Gallery



expect_column_distinct_values_to_contain_set

This expectation level is PRODUCTION

Contributors:

@great_expectations

Tags:

core expectation | column aggregate expectation

Metrics:

column.value_counts

Backend support:

Pandas | Spark | SQLite | PostgreSQL | MySQL | MSSQL | Trino | Redshift | BigQuery | Snowflake

Expectation Type: Core ColumnAggregateExpectation

Description

Expect the set of distinct column values to contain a given set.

expect_column_distinct_values_to_contain_set is a Column Aggregate Expectation.

Args:

- column (str): The column name.
- value_set (set-like): A set of objects used for comparison.

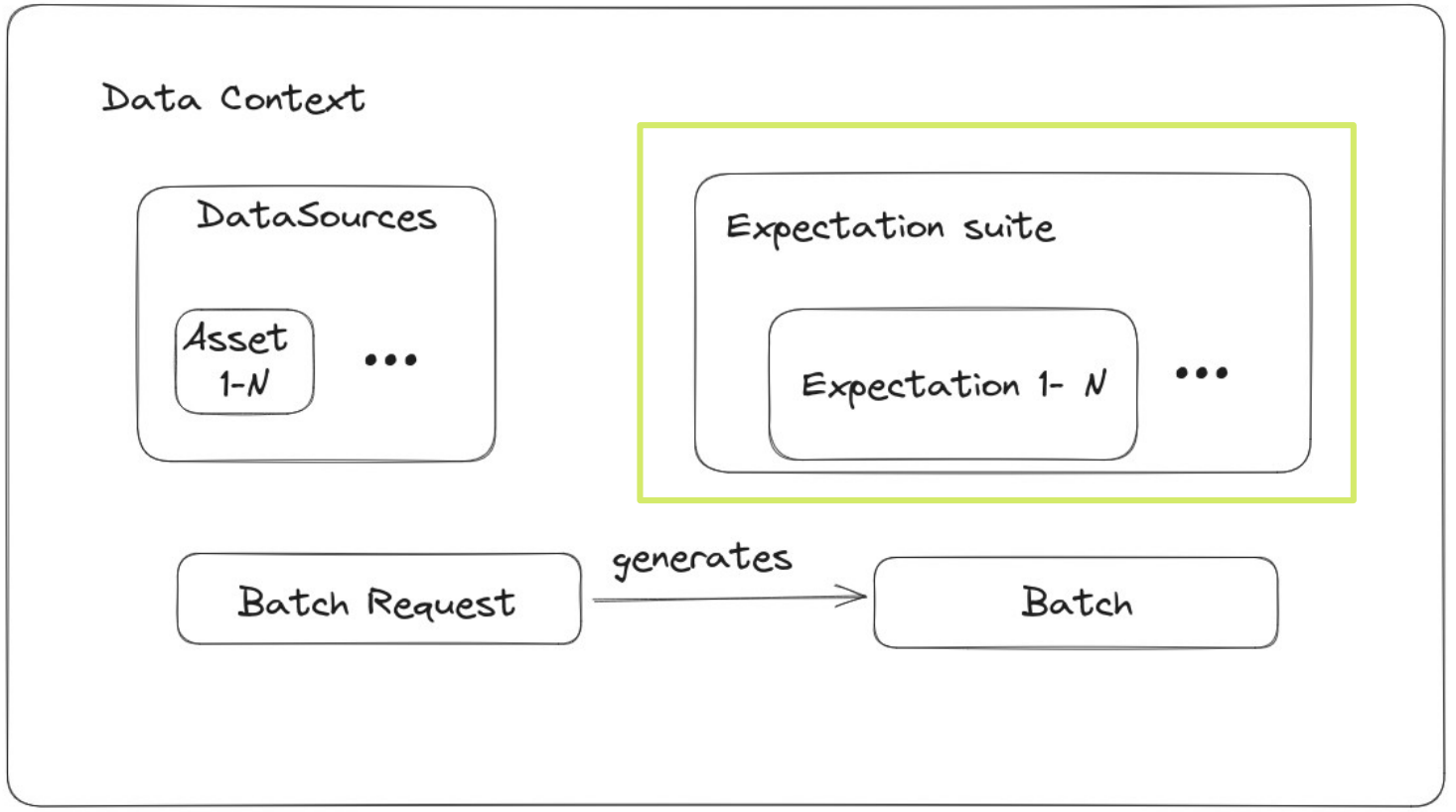
Glossary

Expectations Examples

- expect_column_mean_to_be_between
- expect_table_columns_to_match_set
- expect_column_value_lengths_to_be_between
- expect_multicolumn_sum_to_equal
- expect_compound_columns_to_be_unique
- expect_column_kl_divergence_to_be_less_than

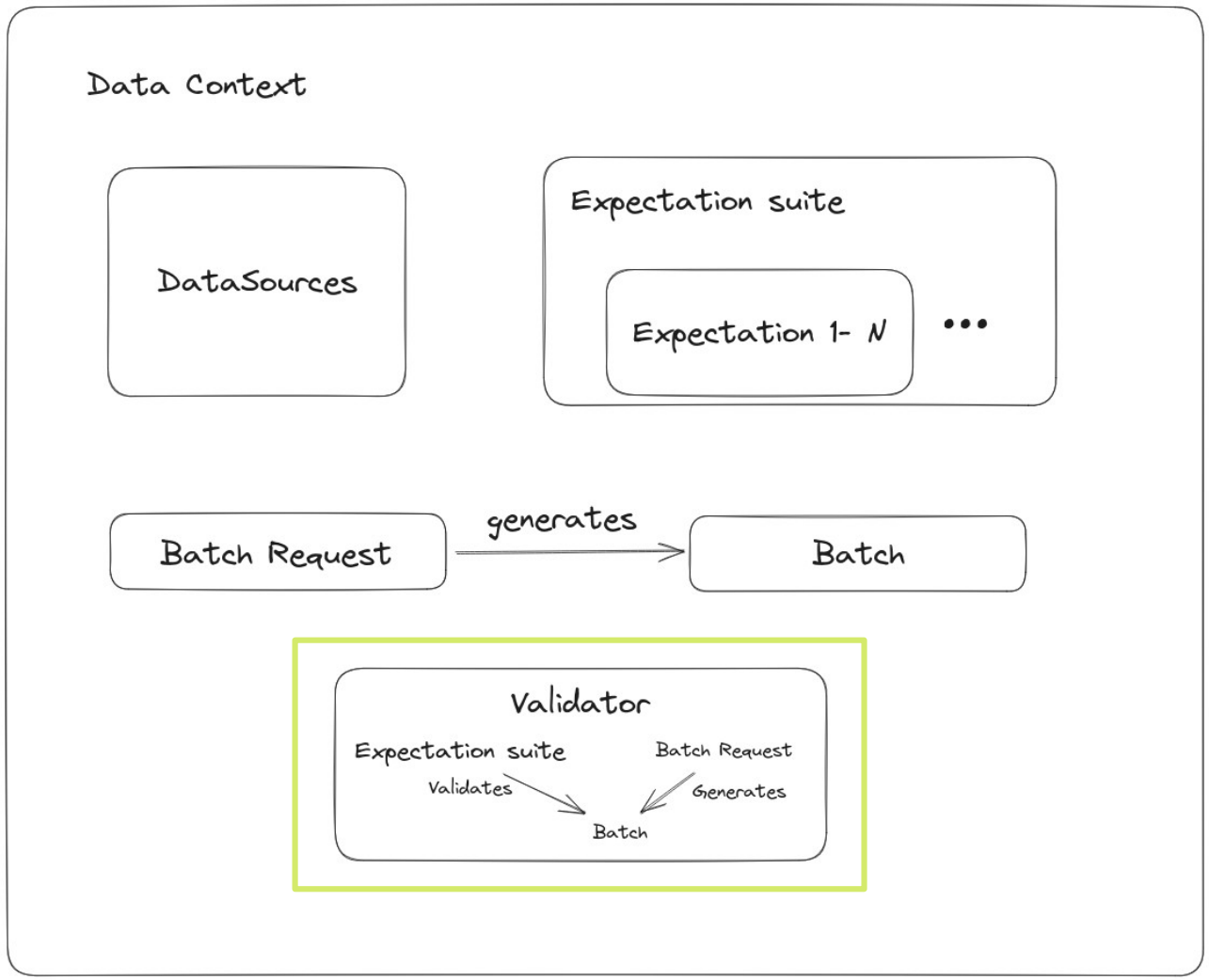
Glossary

- **Expectation Suite:** grouping a bunch of expectations



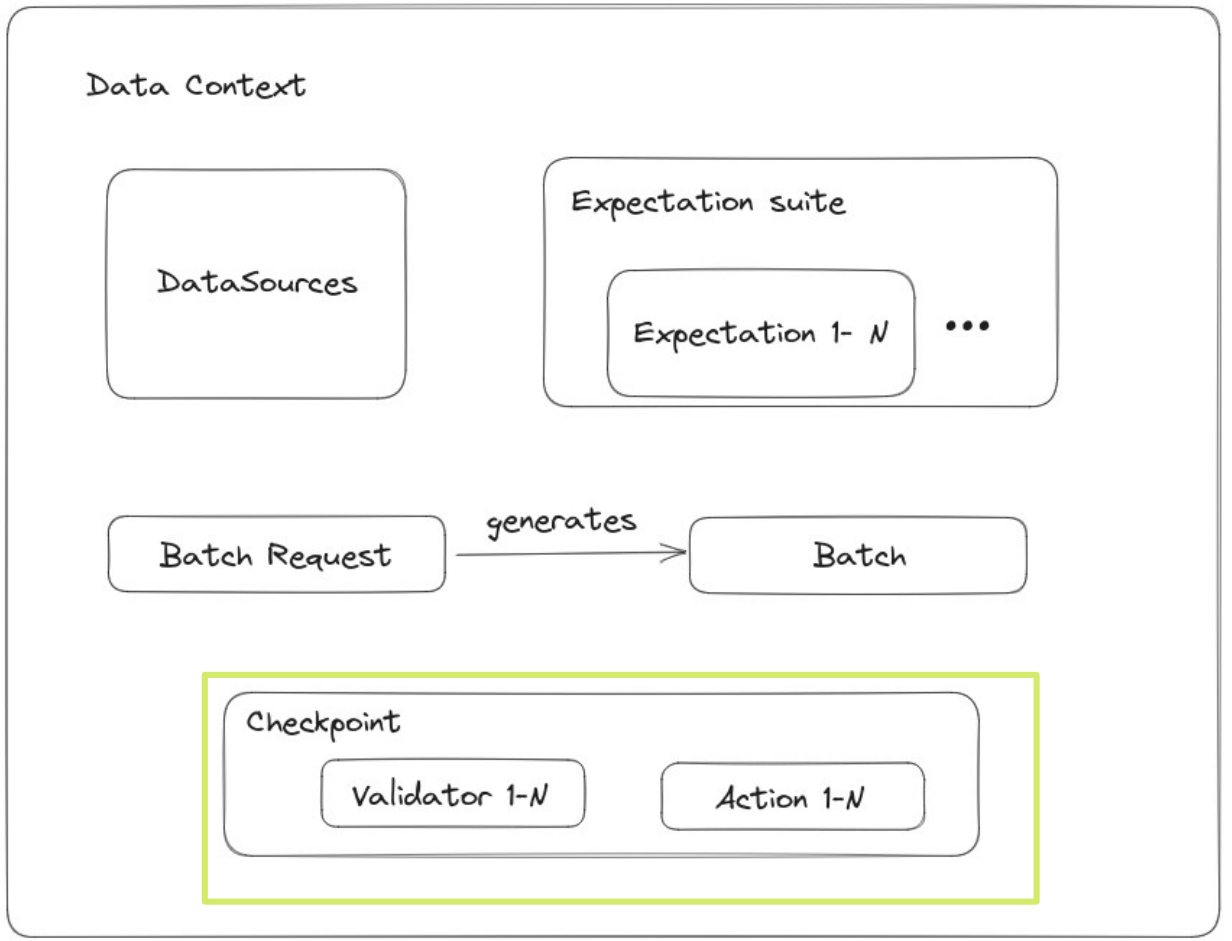
Glossary

- **Validators:** are responsible for running an Expectation Suite against a Batch Request.



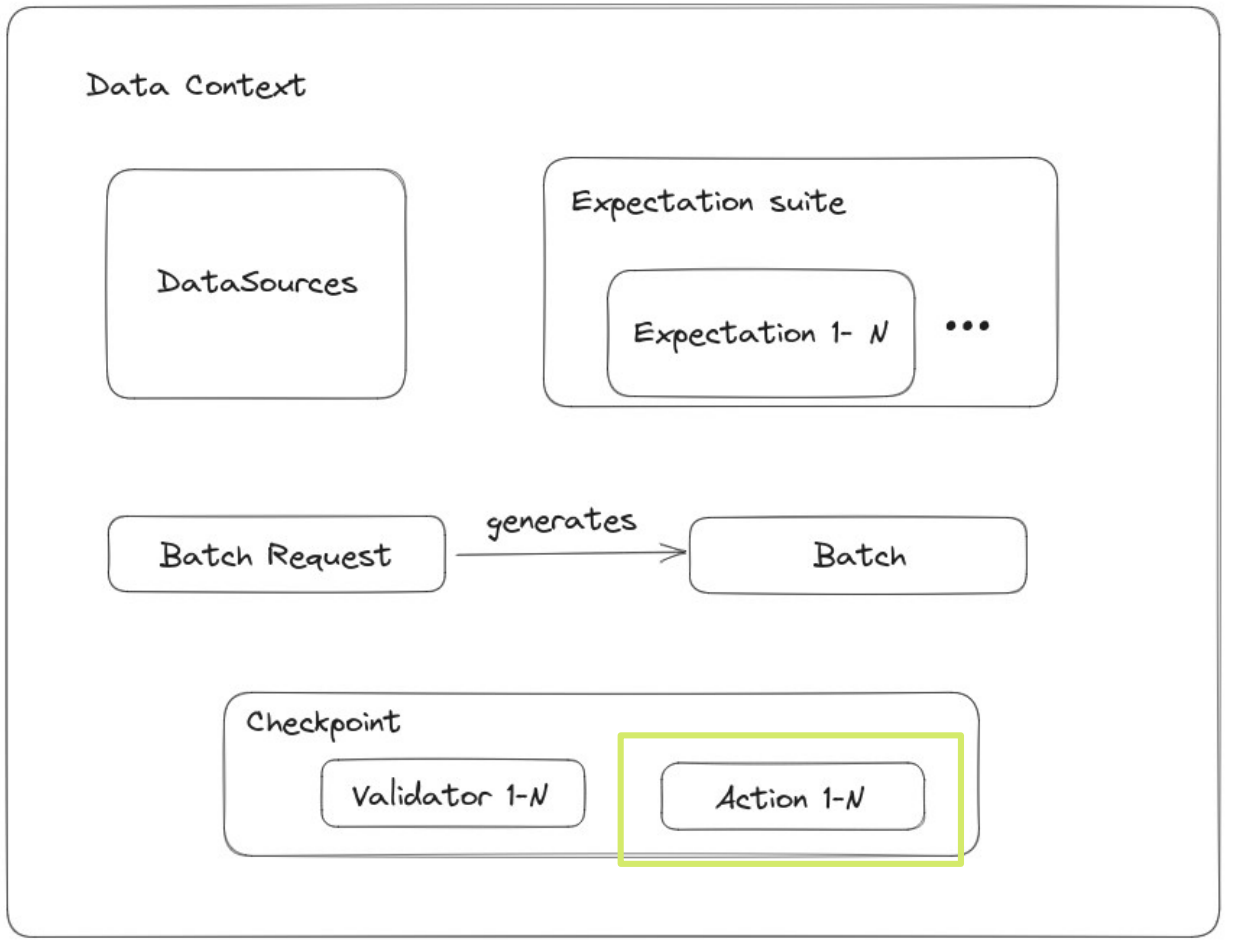
Glossary

- **Checkpoint:** A convenient abstraction for bundling the Validation of a Batch (or Batches) of data against an Expectation Suite (or several), as well as the Actions that should be taken after the validation..



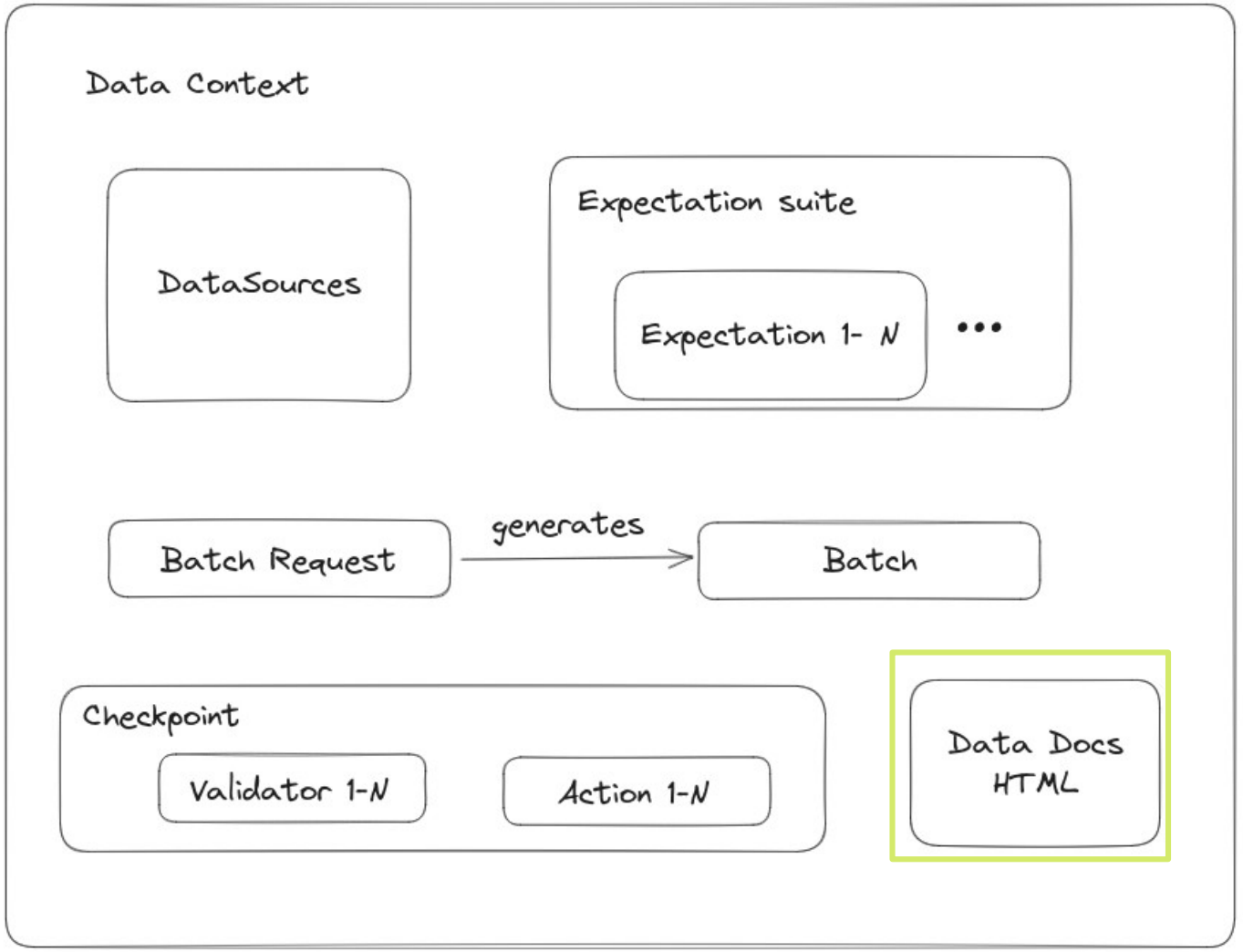
Glossary

- **Actions:**
  - Store Data Docs
  - Store Metrics
  - Store Validation Results
  - Email notification
  - Slack notification
  - MS Teams notification
  - Ops Genie Notification..



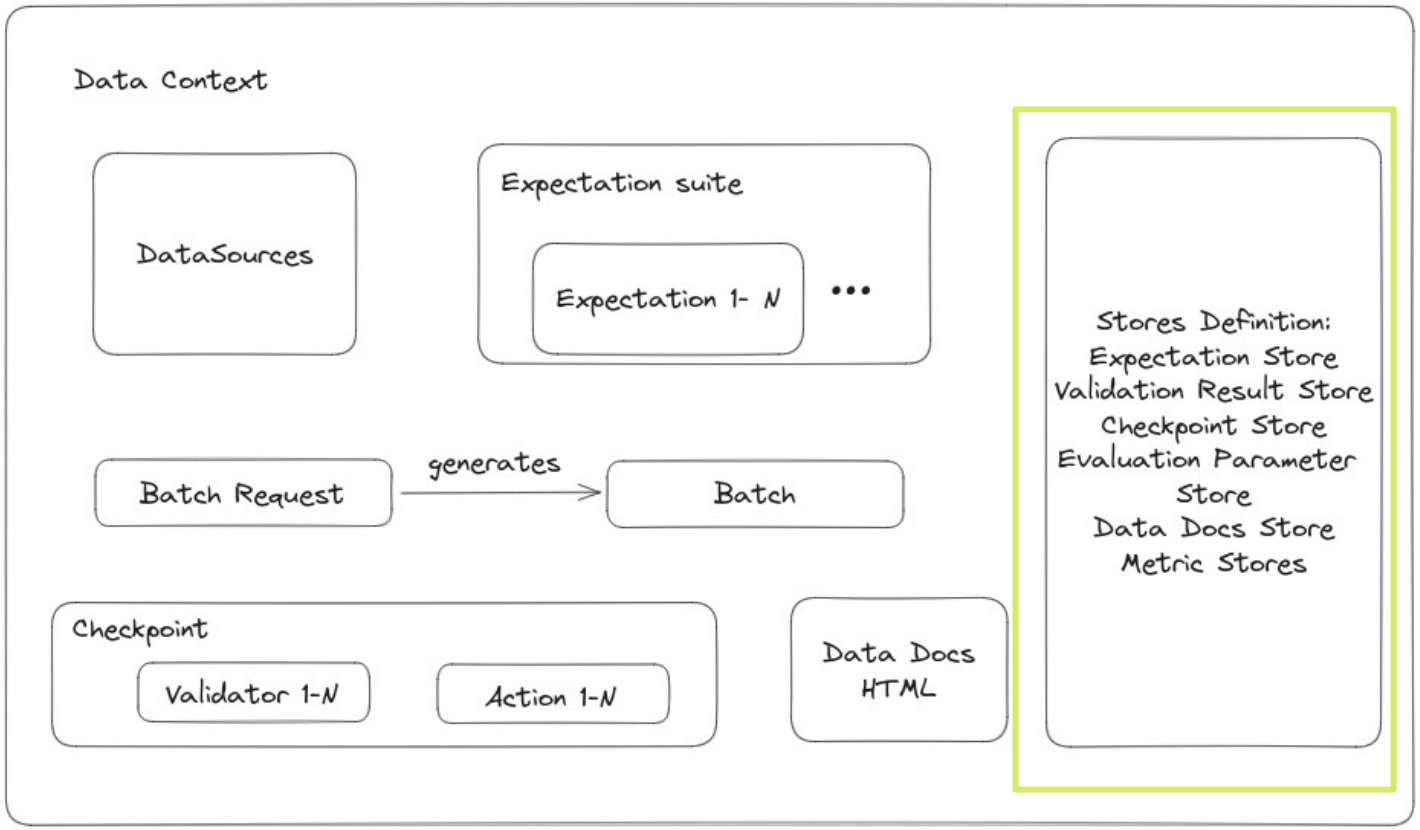
Glossary

- **Data Docs:**
  - Azure
  - GCS
  - S3
  - File System

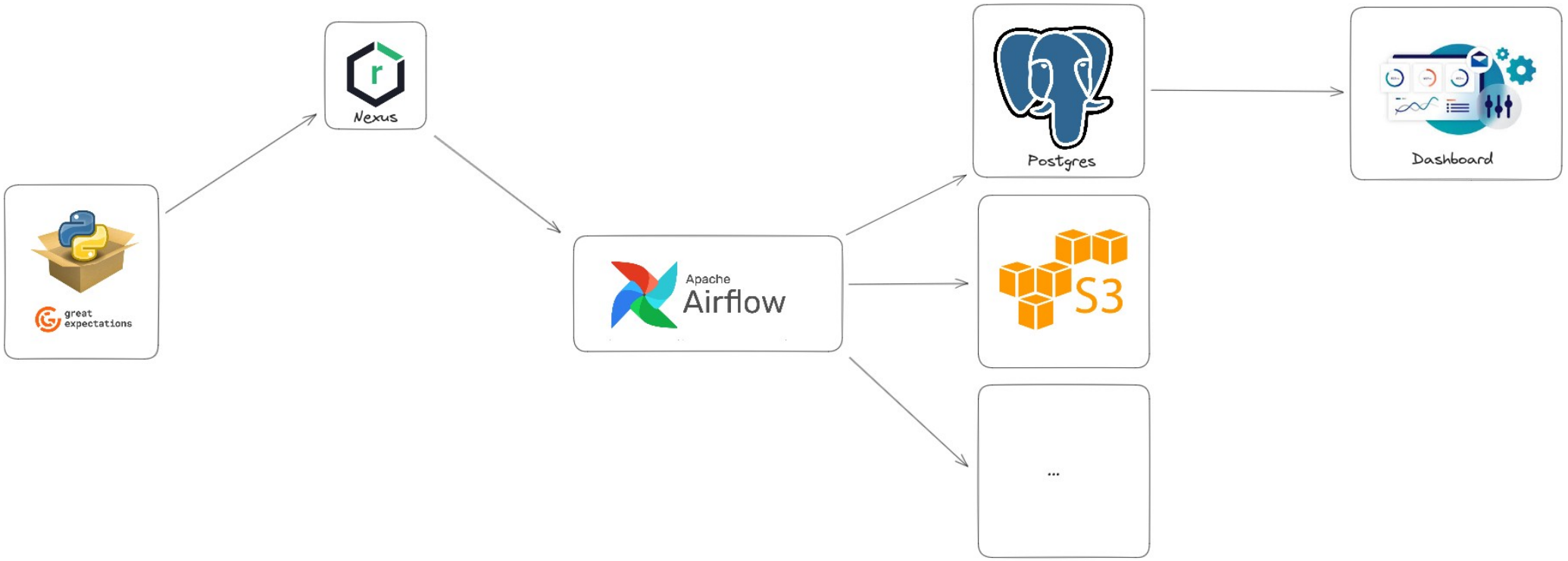


Glossary

- **Stores:**
  - Expectation Store
  - Validation Result Store
  - Checkpoint Store
  - Evaluation Parameter Store
  - Data Docs Store
  - Metric Stores

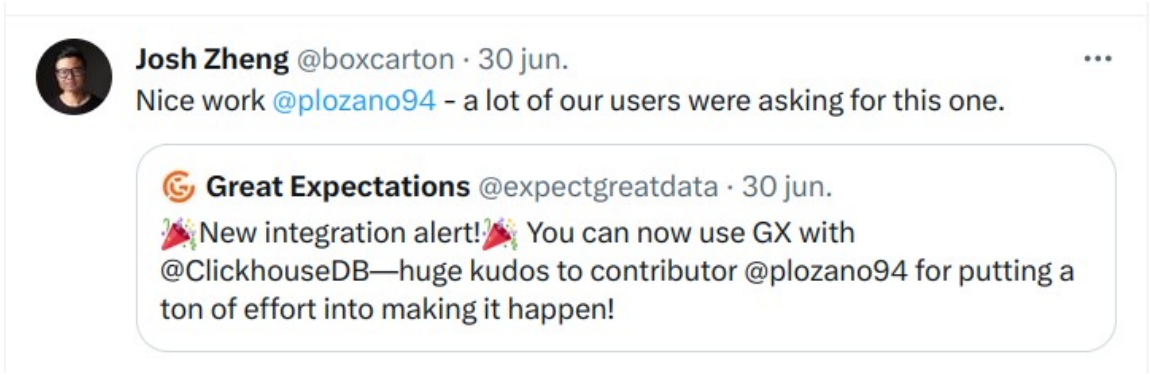


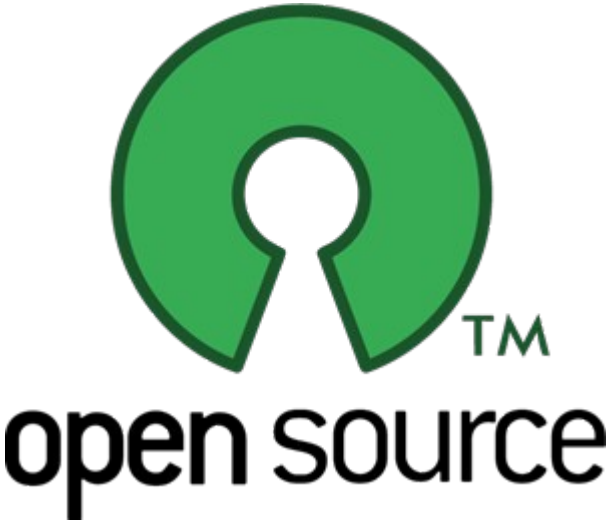
Production time



Production time

Project contribution





idealista

Gracias!

Ruegos y Preguntas