

idealista

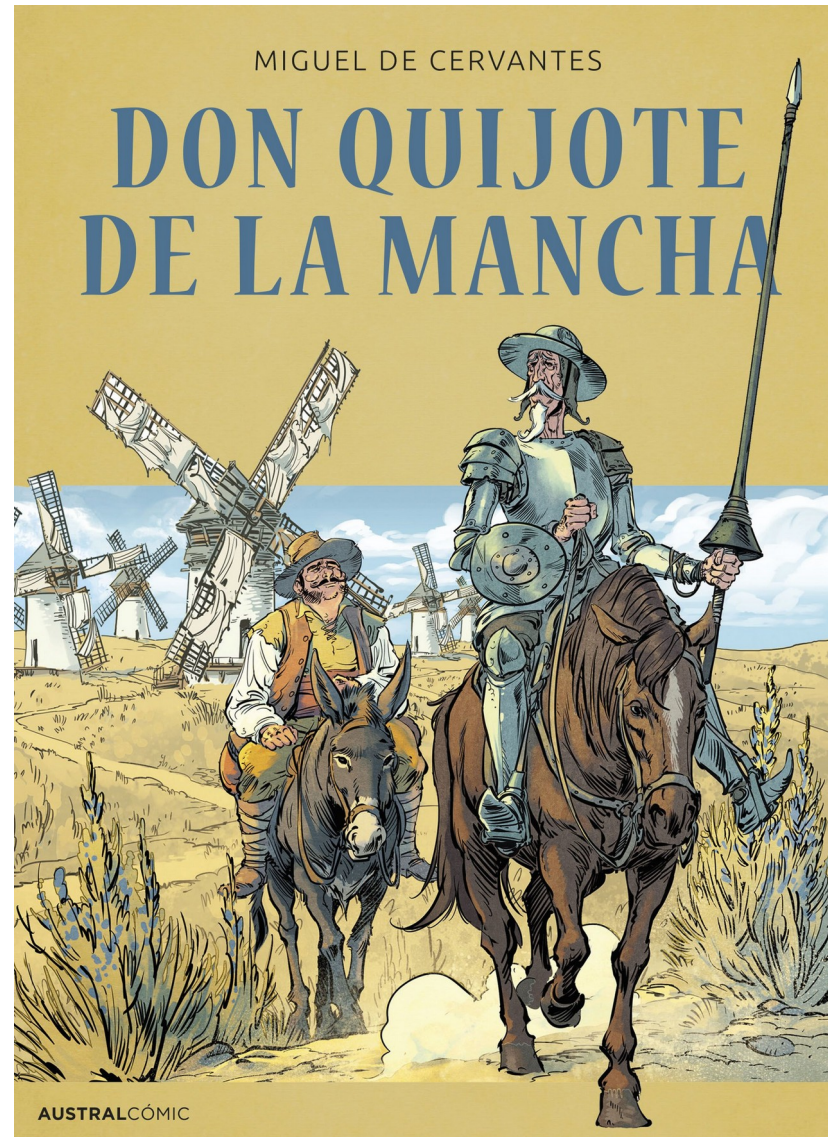
Great Expectations: A ring to rule them all



Pablo Lozano Santiuste



Royal Society - Benders



idealista





Comprar

Alquilar

Compartir

Viviendas



Q Escribe dónde buscas

Buscar

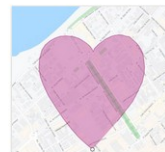
A map of Pamplona, Spain, showing the study area. A red pin marks the 'Clínica Universidad de Navarra'. A blue rectangle highlights the 'Cerro Comercial Plenilunio' area. Other landmarks include 'Centro comercial Arturo Soria Plaza', 'Parque de la Quinta de los Molinos', 'ILUNION Alcalá Norte', 'SIMANCAS', 'Estadio Civitas Metropolitano', 'ALAMEDA DE OSUNA', 'AEROPUERTO', and 'Leroy Merlin Madrid Barajas'. The map is credited to Google and includes a copyright notice for 2023.

hasta 220.000€

2 filtros más



✓ Búsqueda guardada



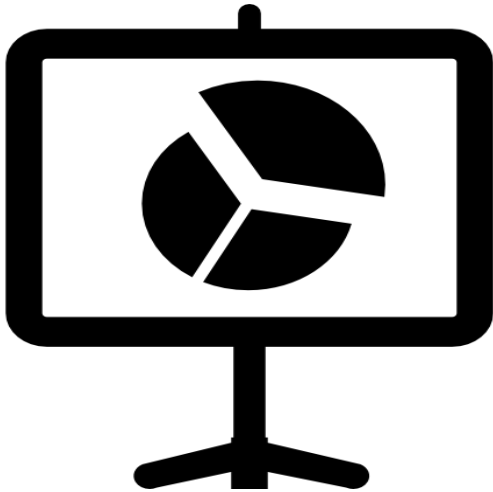
Elegir exactamente la zona que estás buscando sobre un mapa.

A close-up photograph of a person's finger tapping the 'Piso' (Apartment) option on a mobile application. The screen displays a menu titled 'Poner anuncio' (Post ad) with the instruction 'Elige el tipo de inmueble' (Choose the type of property). The menu items include 'Piso', 'Comercios' (Commercial), 'Venta' (Sale), 'Alquiler' (Rental), 'Ubicación del inmueble' (Location of the property), and 'Usar mi posición para ubicarlo' (Use my position to locate it). The 'Piso' option is highlighted with a green bar.

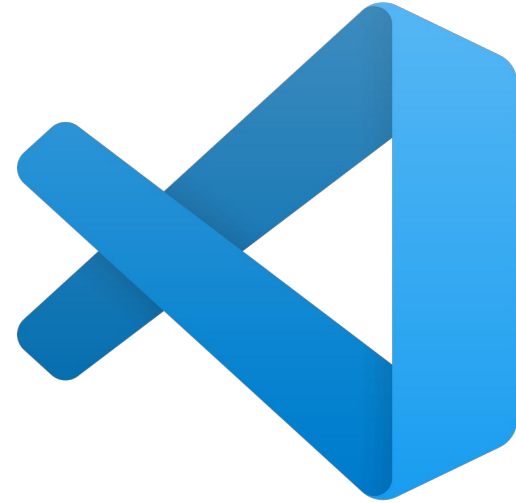
Tus 2 primeros anuncios son gratis. Casas, habitaciones, oficinas... ¡Todo cabe!

Poner tu anuncio gratis

Talk Structure



+



https://github.com/idealista/great_expectations_talk

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 it's 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 :)**

Great Expectations

Documentation

Welcome

- Get started with GX >
- Configure your GX environment >
- Connect to source data >
- Create Expectations ✓
 - Expectation creation workflow
- Manage Expectations and Expectation Suites >
- Profilers and Data Assistants >
- Create Custom Expectations >
- Add Features to Custom Expectations >
- Use a Custom Expectation >
- Validate Data >
- Integrations >
- Concepts >
- Reference >
- Changelog
- Migration guide
- Contribute

Github

 [great-expectations / great_expectations](#) Public

 Fork 1.4k

 Star 8.7k

Slack

Great Expectations ▾



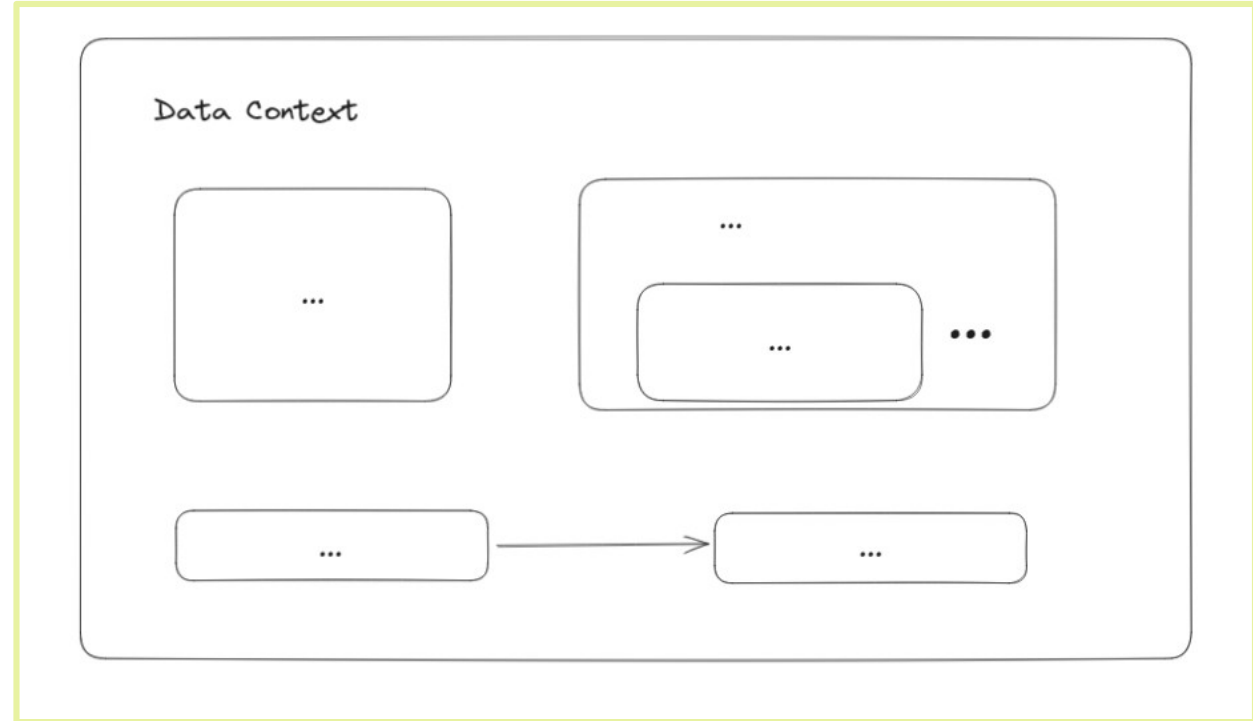
- ▼ Canales
- # announcements
- # chit-chat
- # contributing
- # data-quality-discussions
- # events
- # gx-community-support
- # gx-feedback
- # gx-releases
- # gx-share
- # introductions
- # job-board
- # tools-and-services
- + Añadir canales

¿Como instalamos la herramienta?

```
pyproject.toml
1  [tool.poetry]
2  name = "great-expectations-talk"
3  version = "0.1.0"
4  description = ""
5  authors = ["Pablo Lozano Santiuste <plozano94@gmail.com>"]
6  readme = "README.md"
7
8  [tool.poetry.scripts]
9  accidentes_checkpoint = "great_expectations_talk.checkpoint.accidentes_checkpoint:run"
10
11 [tool.poetry.dependencies]
12 python = "^3.9"
13 great-expectations = "^0.17.14"
14 sqlalchemy = "^1"
15 clickhouse-sqlalchemy = "^0.2.4"
16 psycpg2 = "^2.9.7"
17 click = "^8.1.7"
18
19
20 [[tool.poetry.source]]
21 name = "id"
22 url = "http://nexus.int.sys.idealista/repository/pypi-idealista/"
23 priority = "default"
24
25 [build-system]
26 requires = ["poetry-core"]
27 build-backend = "poetry.core.masonry.api"
```

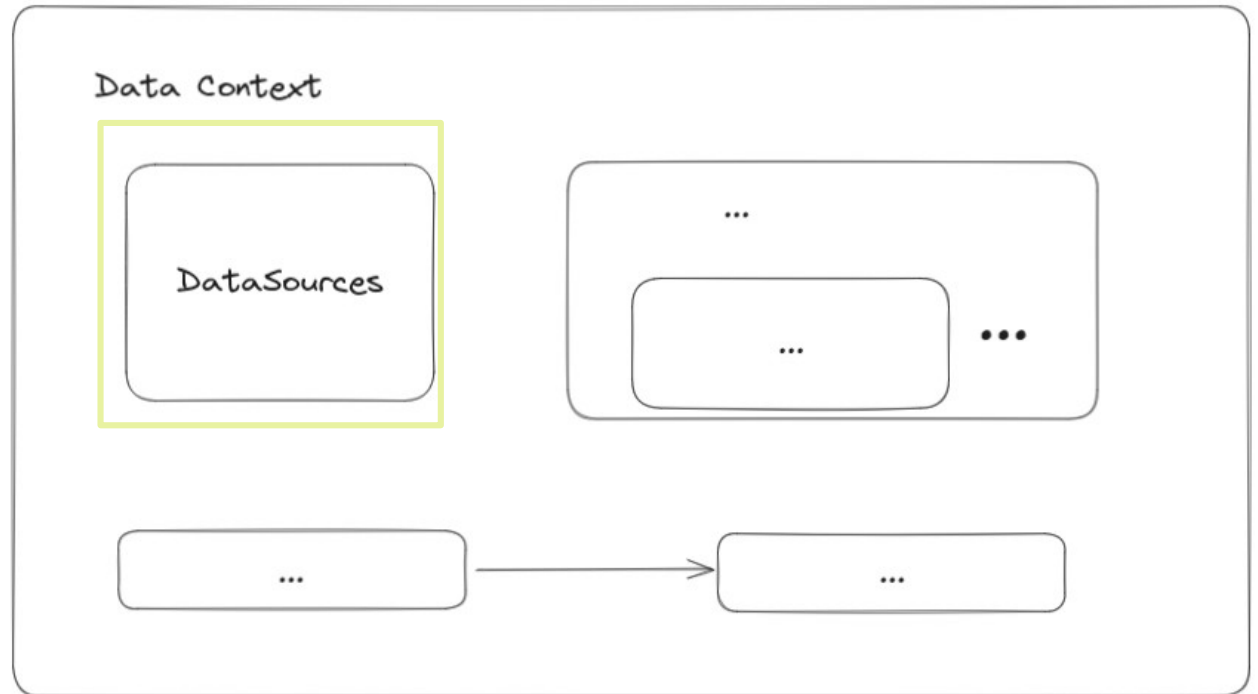
Glossary

- **Data Context**
 - EphemeralDataContext
 - **FileDataContext**
 - CloudDataContext



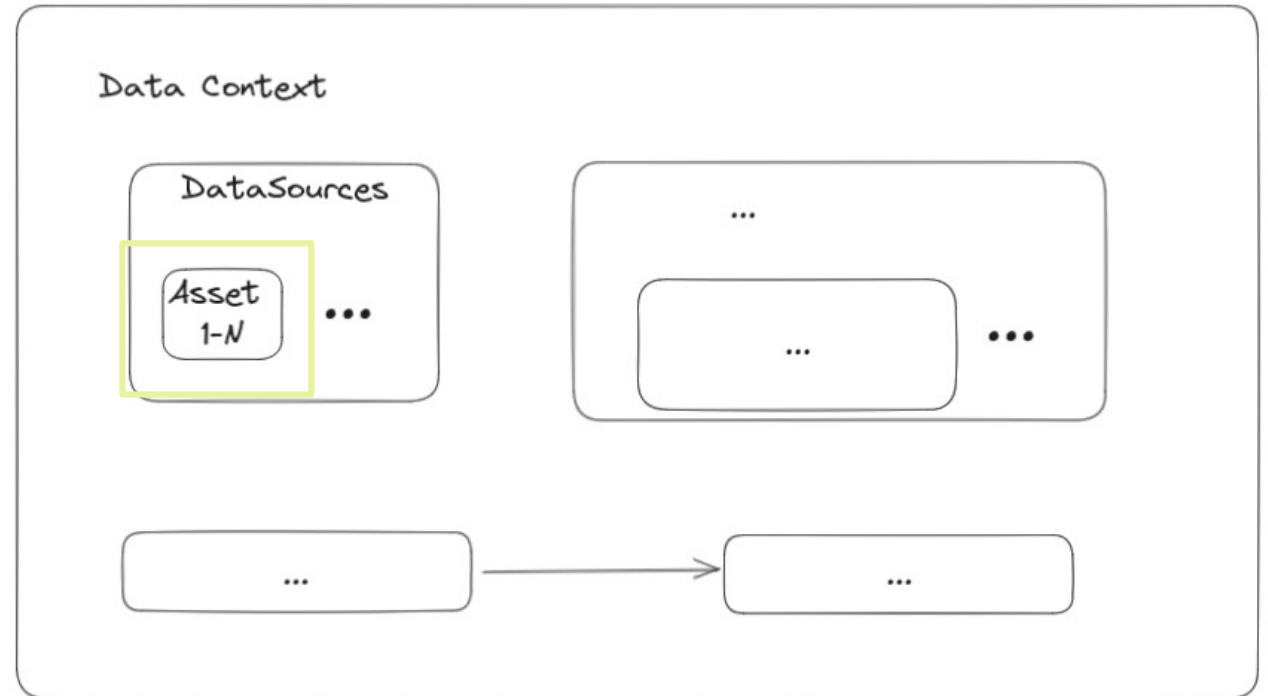
Glossary

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



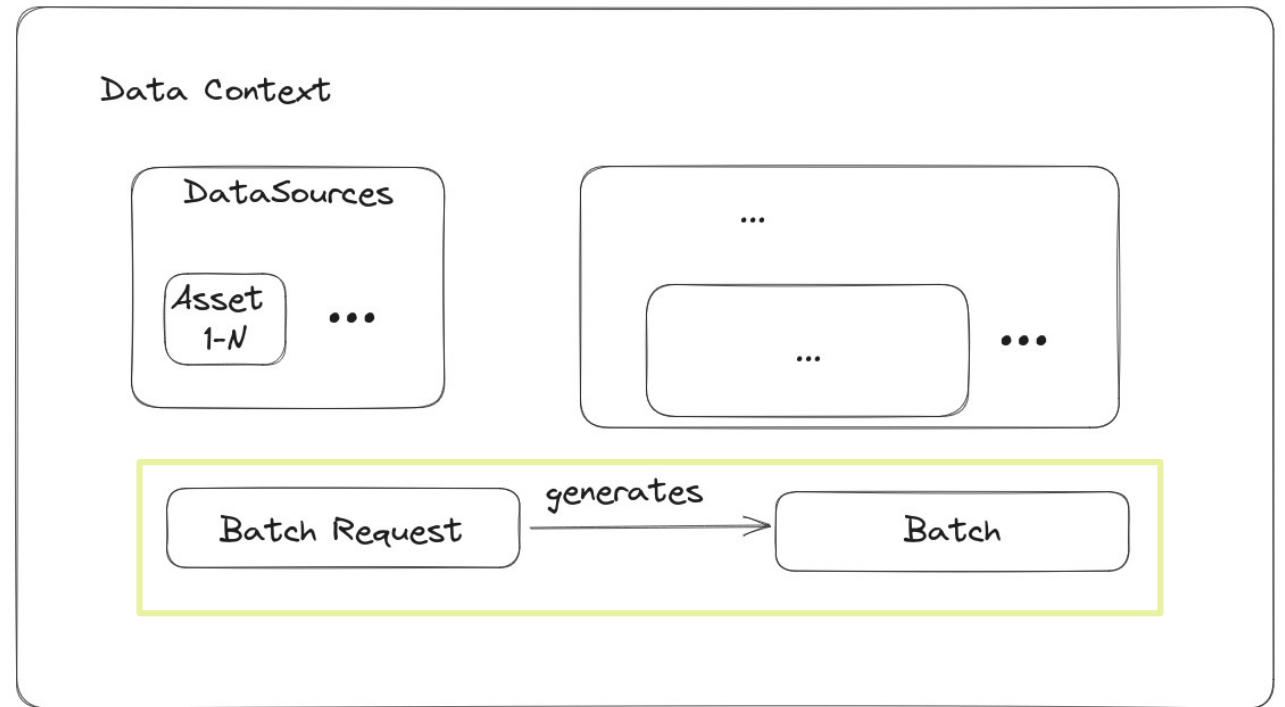
Glossary

- **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

```
{
  "expectation_type": "expect_column_distinct_values_to_equal_set",
  "kwargs": {
    "column": "tipo_vehiculo",
    "value_set": [
      "Bicicleta EPAC (pedaleo asistido)",
      "Bicicleta"
    ]
  },
  "meta": {}
}
```

Glossary

Expectations Gallery

Expectations: 324 Total

Search here...

Coverage

Summary

Completeness

Datasource

Clear filters

Filter by: Backend support Select Items

expect_column_kl_divergence_to_be_less_than (Core ColumnAggregateExpectation)

Expect the Kulback-Leibler (KL) divergence (relative entropy) of the specified column with respect to the partition object to be lower than the provided threshold.

Tags: core expectation column aggregate ex... distributional expect...

Support:

Contribution status:

Experimental4 / 4

Beta3 / 3

Production2 / 2

expect_column_values_to_be_between (Core ColumnMapExpectation)

Expect the column entries to be between a minimum value and a maximum value (inclusive).

Tags: core expectation column map expecta...

Support:

Contribution status:

Experimental4 / 4

Beta2 / 3

Production2 / 2

expect_column_distinct_values_to_contain_set (Core ColumnAggregateExpectation)

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

Tags: core expectation column aggregate ex...

Support:

Contribution status:

Experimental4 / 4

Beta3 / 3

Production2 / 2

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.

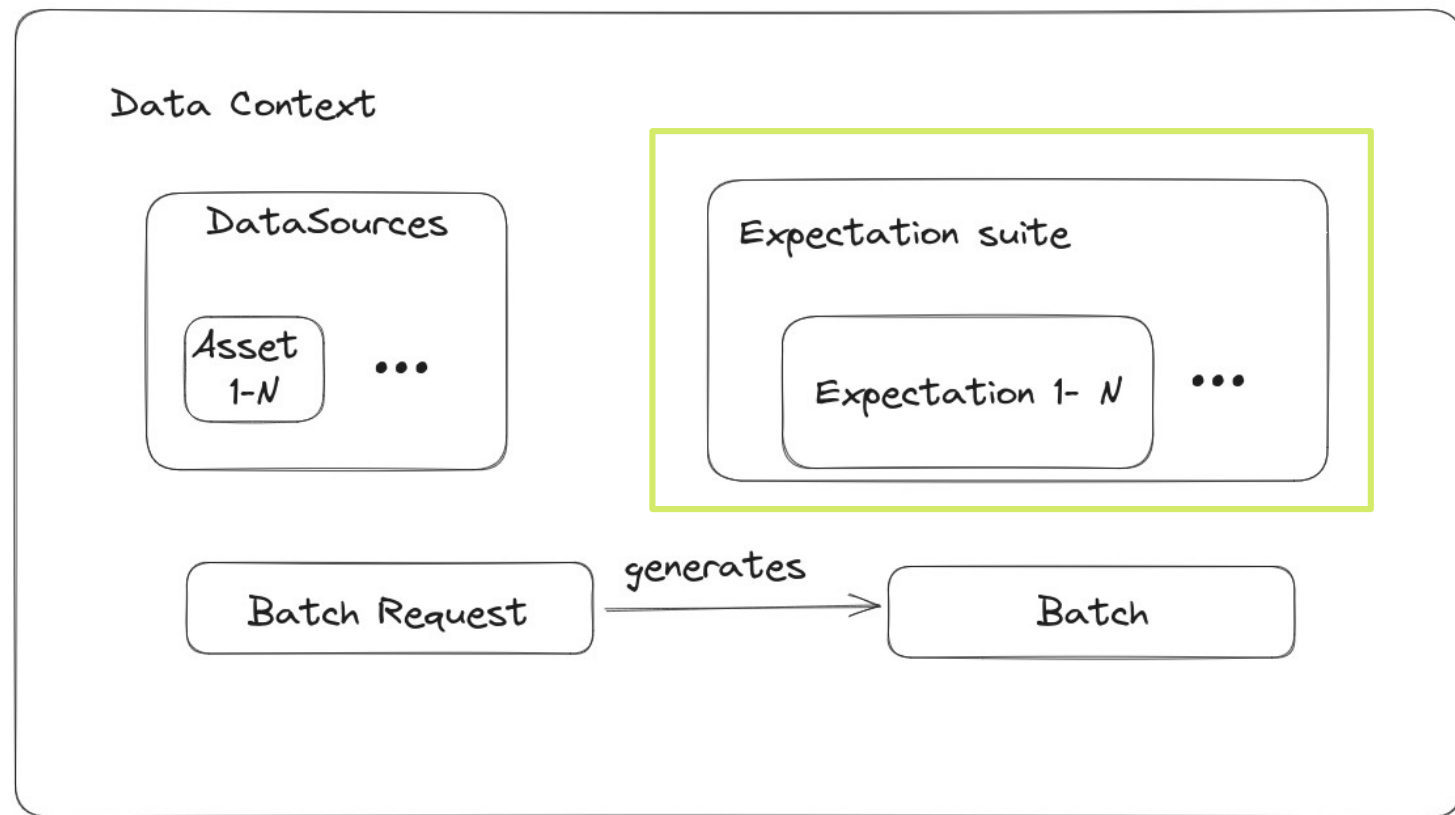
Glossary

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`

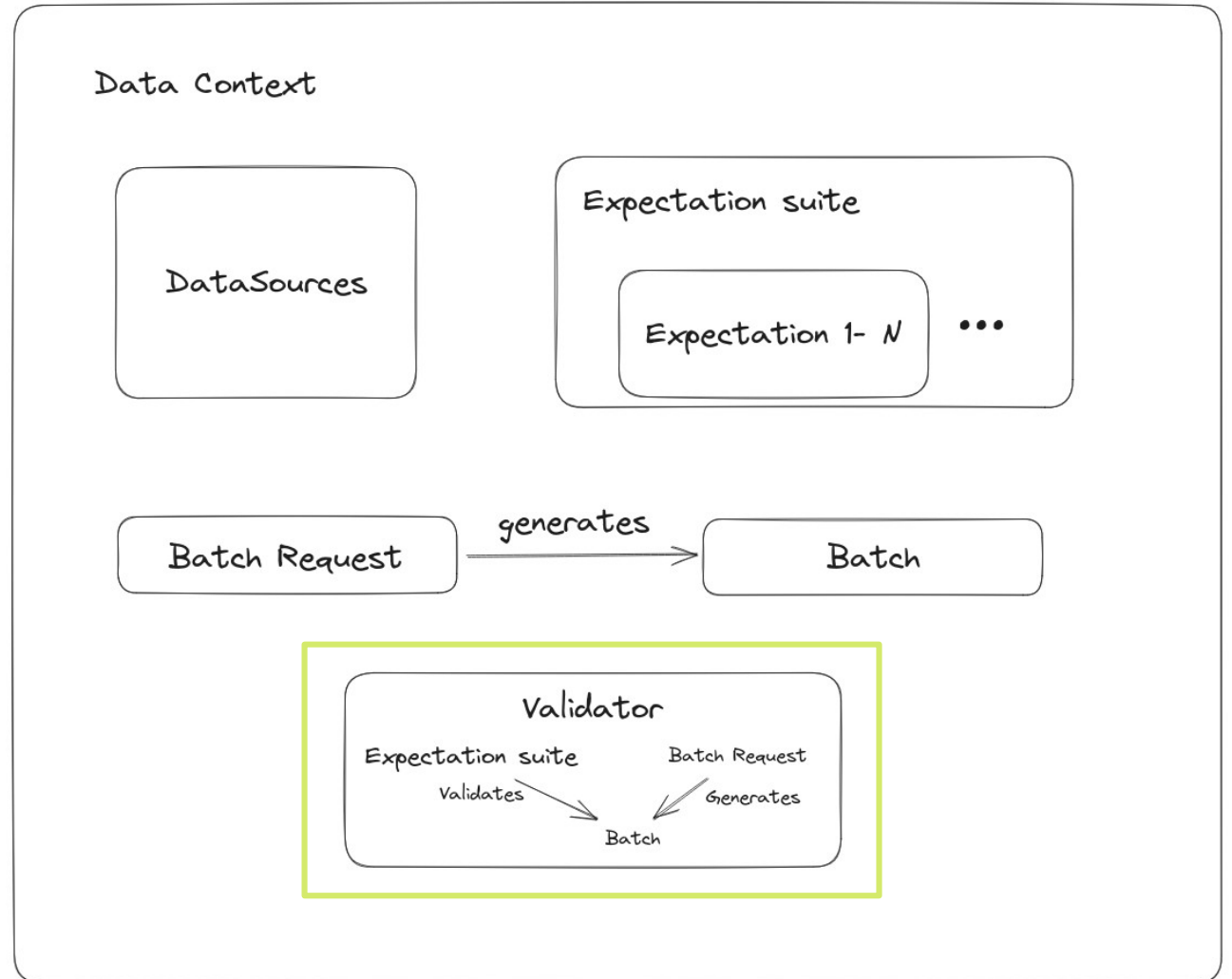
Glossary

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



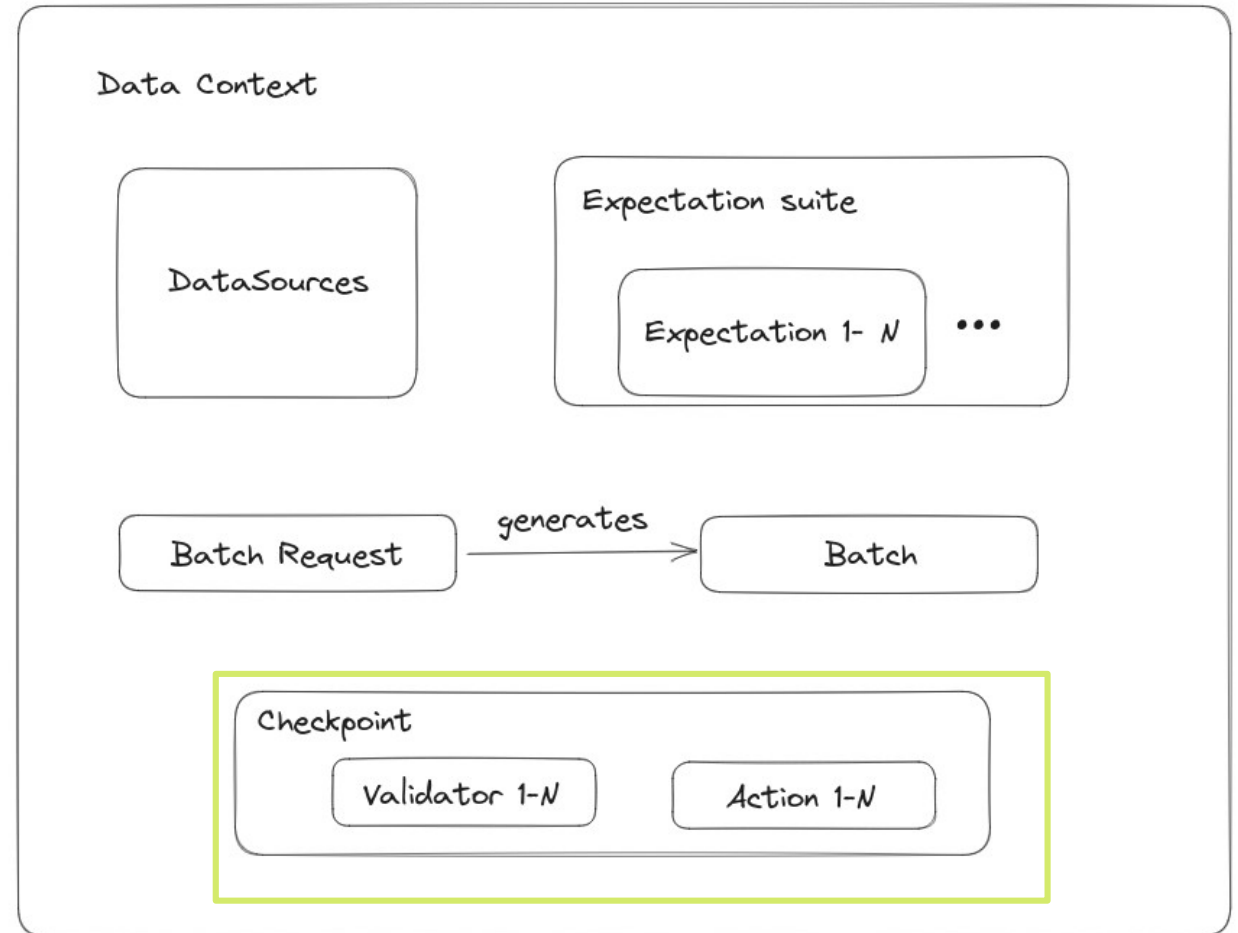
Glossary

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



Glossary

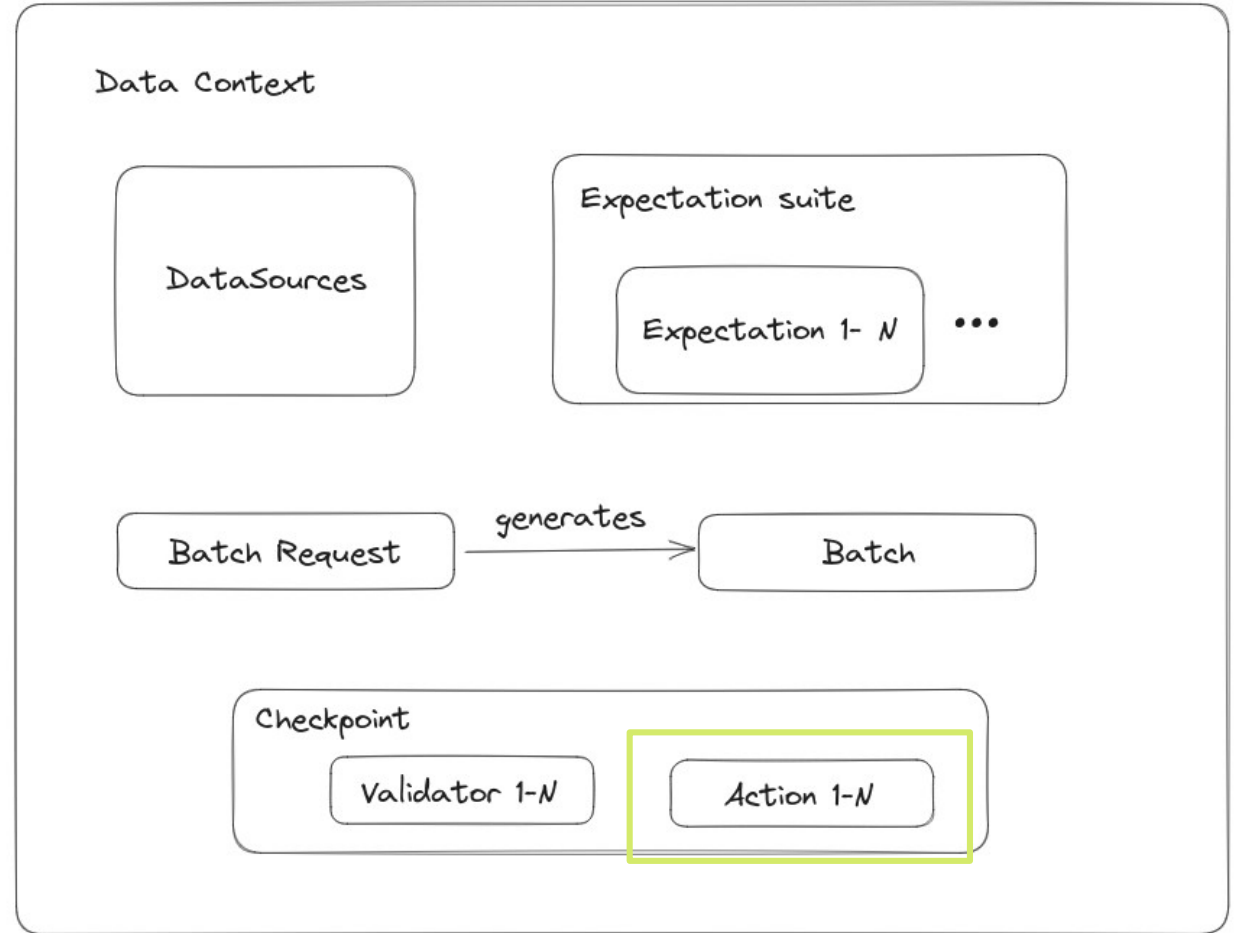
- **Checkpoint:** A convenient abstraction for bundling the Validation of a Batch (or Batches) of data against an Expectation Suite (or several), as well as the Actions that should be taken after the validation..



Glossary

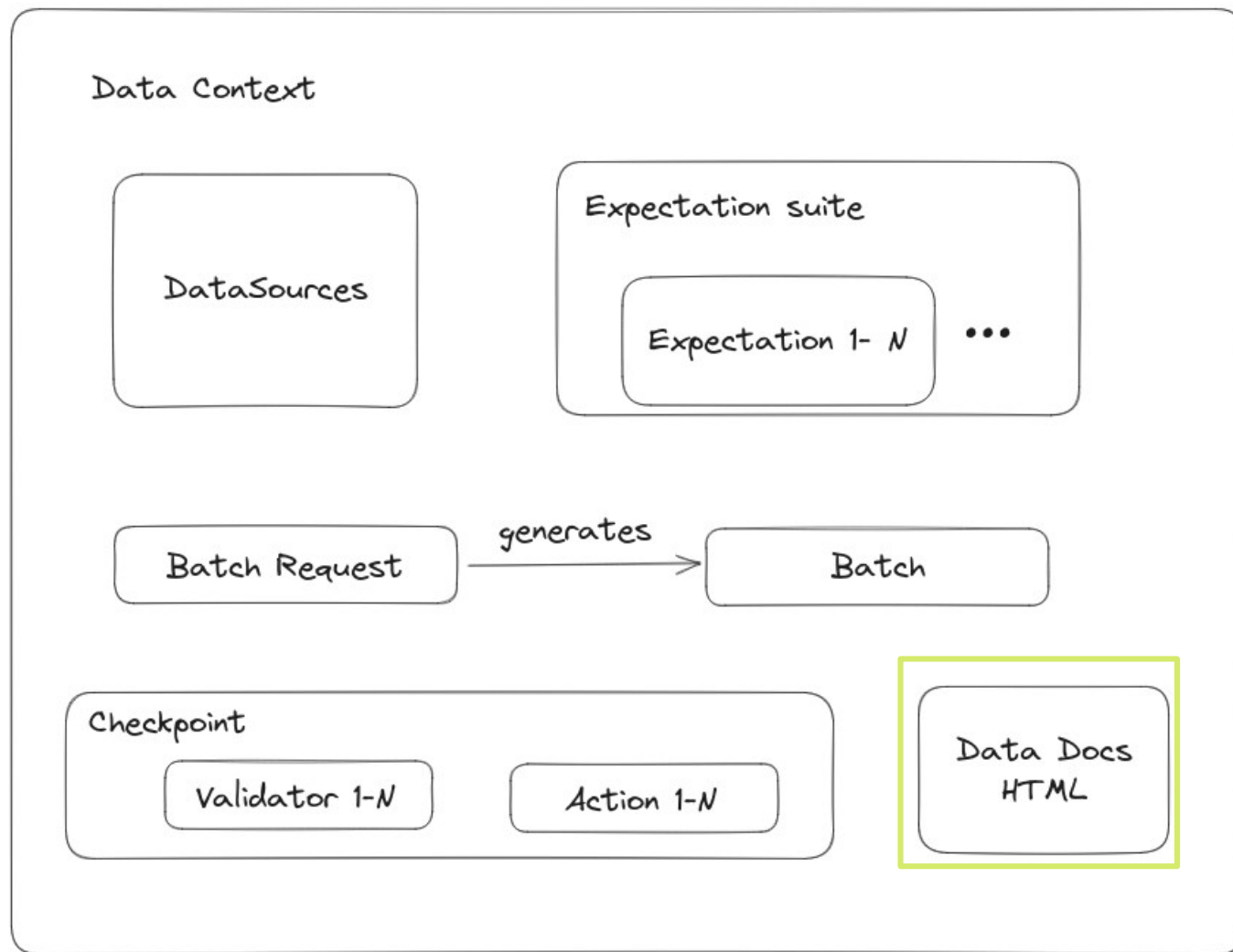
- **Actions:**

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



Glossary

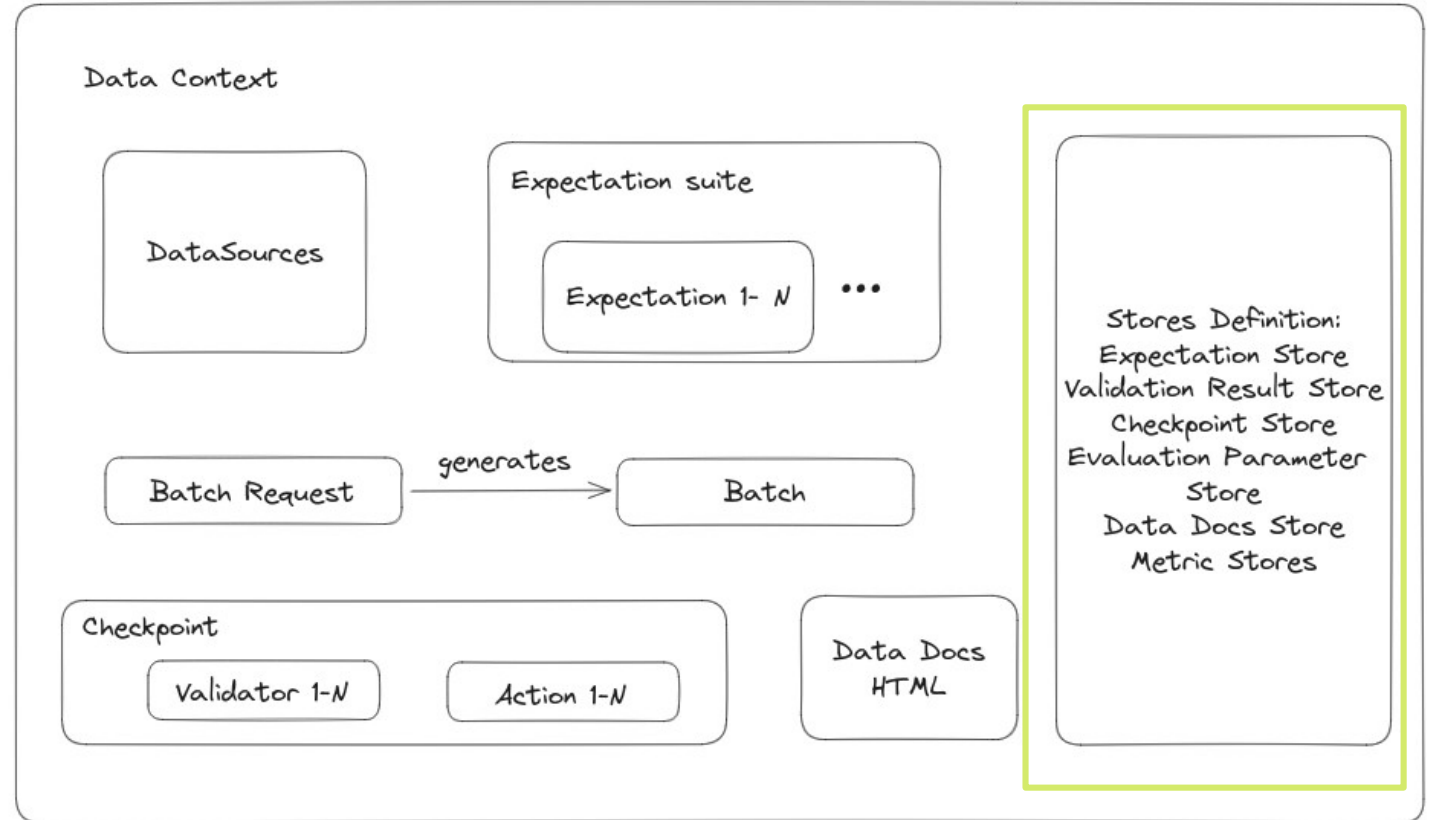
- **Data Docs:**
 - Azure
 - GCS
 - S3
 - File System



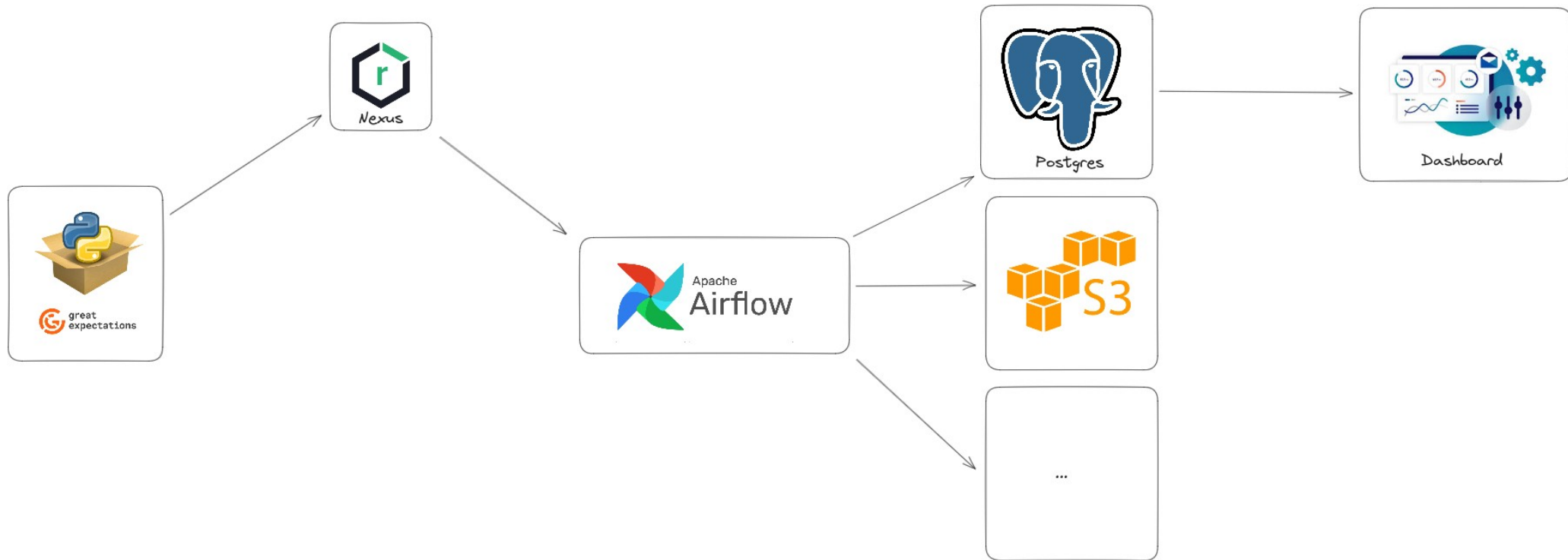
Glossary

• Stores:

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



Production time



Production time

Project contribution



Josh Zheng @boxcarton · 30 jun. ...

Nice work @plozano94 - a lot of our users were asking for this one.



Great Expectations @expectgreatdata · 30 jun.

🎉 New integration alert! 🎉 You can now use GX with @ClickhouseDB—huge kudos to contributor @plozano94 for putting a ton of effort into making it happen!



Great Expectations

2.710 seguidores

2 meses • 🌐

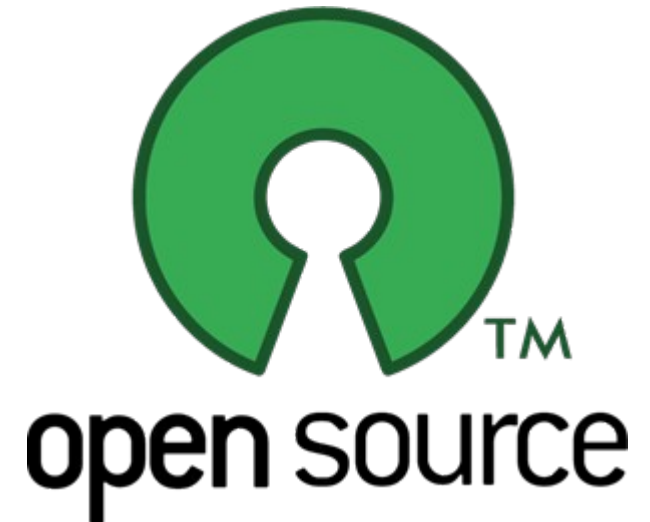
ICYMI: GX now integrates with [ClickHouse](#), courtesy of contributor [Pablo Lozano Santiuste](#)!

[#data](#) [#dataquality](#) [#clickhouse](#) [#integrations](#)



Great Expectations @expectgreatdata · 5 jul. ...

ICYMI: GX now integrates with [@ClickhouseDB](#), thanks to contributor [@plozano94](#)!



idealista

Gracias!

Ruegos y Preguntas

