



**being in control of data quality**

# Thank you, sponsors





# Dave Ruijter

Solution Architect Data & Analytics  
Blue Rocket IT



dave@blue-rocket.it



@DaveRuijter



linkedin.com/in/DaveRuijter

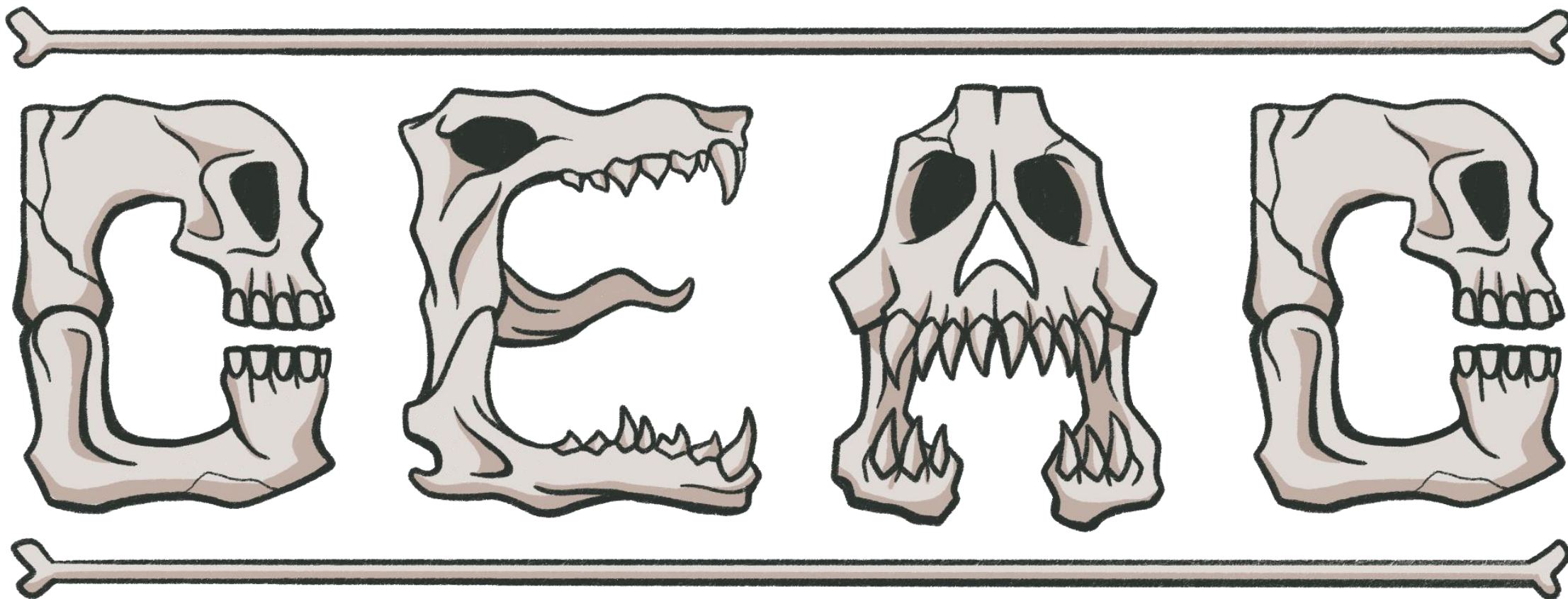


ModernData.ai

**THERE ARE NO DATA ISSUES**

**IF YOU DON'T CHECK DATA QUALITY**





Dumbledilps Equipment Adventuring INC



01

# Great Expectations

---



# How to work with Great Expectations

---

- data testing
- documentation
- profiling



# What are Expectations?

---

- `expect_column_to_exist`
- `expect_table_row_count_to_be_between`
- `expect_column_values_to_be_unique`
- `expect_column_values_to_not_be_null`
- `expect_column_values_to_be_between`
- `expect_column_values_to_match_regex`
- `expect_column_mean_to_be_between`
- `expect_column_kl_divergence_to_be_less_than`





# Glossary of expectations

## Dataset

Dataset objects model tabular data and include expectations with row and column semantics. Many Dataset expectations are implemented using `column_map_expectation` and `column_aggregate_expectation` decorators.

Not all expectations are currently available for each backend. A table describing available implementations per-backend is available here: [Table of Expectation Implementations By Backend](#).

## Table shape

- `expect_column_to_exist`
- `expect_table_columns_to_match_ordered_list`
- `expect_table_row_count_to_be_between`
- `expect_table_row_count_to_equal`

## Missing values, unique values, and types

- `expect_column_values_to_be_unique`
- `expect_column_values_to_not_be_null`
- `expect_column_values_to_be_null`
- `expect_column_values_to_be_of_type`
- `expect_column_values_to_be_in_type_list`



# Tests are docs and docs are tests

---

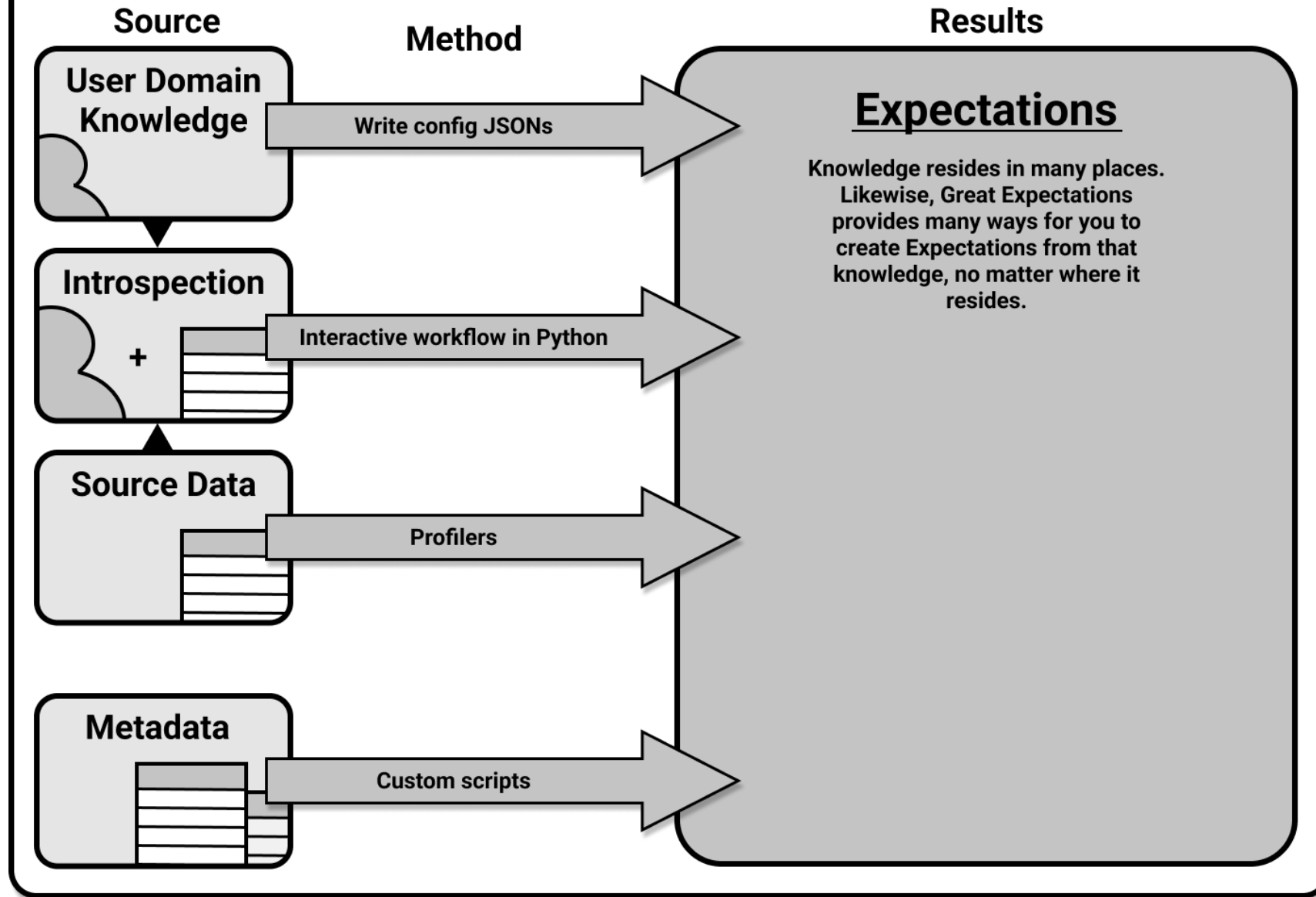
```
expect_column_values_to be  
between (  
    column="room_temp",  
    min_value=60,  
    max_value=75,  
    mostly=.95  
)
```



"Values in this column should be between 60 and 75, at least 95% of the time."

"Warning: more than 5% of values fell outside the specified range of 60 to 75."

# Where do Expectations come from?



# Validation produces a validation result object

---

```
{
  "success": false,
  "result": {
    "element_count": 253405,
    "unexpected_count": 7602,
    "unexpected_percent": 2.999
  },
  "expectation_config": {
    "expectation_type": "expect_column_values_to_not_be_null",
    "kwargs": {
      "column": "user_id"
    }
  }
}
```



# Validation results save you time.



Status	Expectation	Observed Value
✓	values must never be null.	100% not null



values must belong to this set: **Y** **N**.

≈20.08%  
unexpected

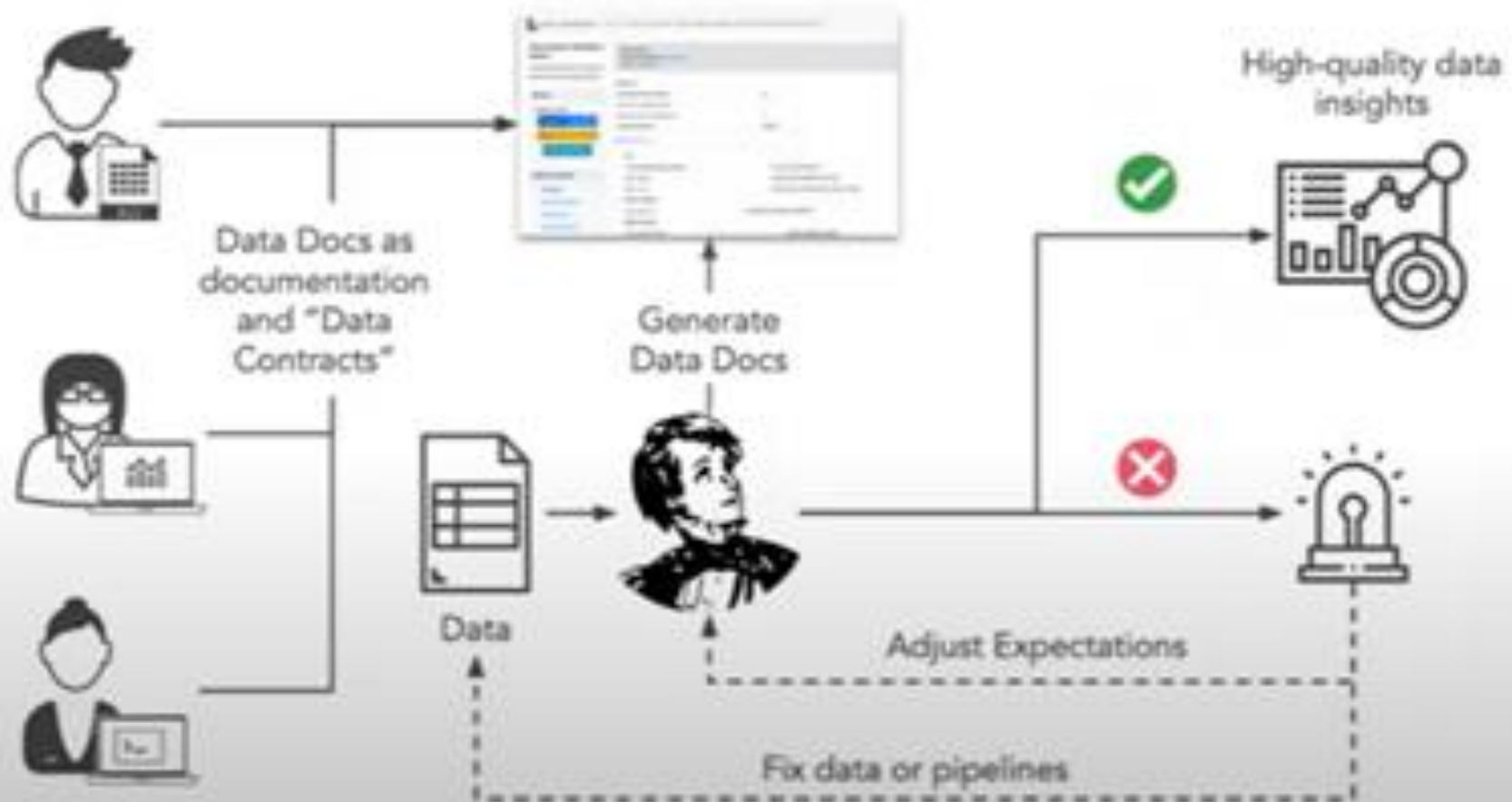
**218612 unexpected values found. ≈20.08% of 1088696 total rows.**

Unexpected Value	Count
yes	10
no	4
No.	2
n	2
y	2

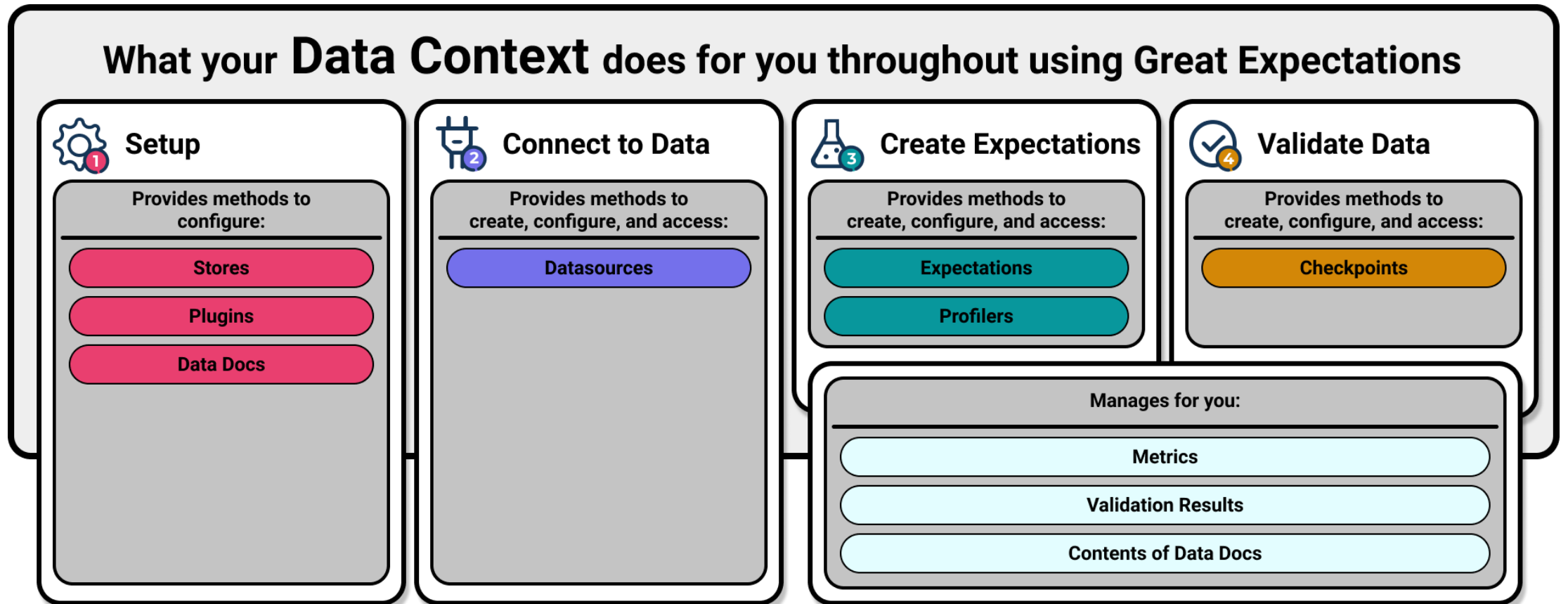




# The workflows



# The Data Context



# DataDocs on Azure

---

<https://dlswedataplatformlab.z6.web.core.windows.net/index.html>



02

# Azure Databricks

---



# Libraries in cluster

<input type="checkbox"/>	Name	Type	Status
<input type="checkbox"/>	great-expectations	PyPI	🟢 Installed
<input type="checkbox"/>	azure.identity	PyPI	🟢 Installed
<input type="checkbox"/>	azure-storage-blob	PyPI	🟢 Installed



04

# Power BI

---



# Testing the data model, not the reports

---

- On datasets in the Power BI Service (not Desktop)
- Does not require Premium workspace

*<https://docs.microsoft.com/en-us/rest/api/power-bi/datasets/execute-queries>*




# Validations

---

- *Exacte waarde van 1 order*
- *Totale waarde van 2020, 2021*
- *Complex measure*
  - *Percentage van iets*
  - *Time intelligence*
  - *Period over period*
- *Verify if all dimension values (e.g. product brands) are accounted for*





```
{
  "queries": [
    {
      "query": "EVALUATE
SUMMARIZECOLUMNS(Products[Brand
Class])"
    }
  ],
  "serializerSettings": {
    "includeNulls": true
  },
  "impersonatedUserName":
"someuser@mycompany.com"
}
```

```
{
  "results": [
    {
      "tables": [
        {
          "rows": [
            {
              "Products[Brand Class]": "NON-MAGIC PRODUCTS"
            },
            {
              "Products[Brand Class]": "MAGIC PRODUCTS"
            },
            {
              "Products[Brand Class]": "MAGIC BRANDS"
            },
            {
              "Products[Brand Class]": "NON-MAGIC BRANDS"
            }
          ]
        }
      ]
    }
  ]
}
```







# Setup

 Refresh |  Got feedback?

## Configured permissions

Applications are authorized to call APIs when they are granted permissions by users/admins as part of the consent process. The list of configured permissions should include all the permissions the application needs. [Learn more about permissions and consent](#)

 Add a permission  Grant admin consent for Dave Ruijter

API / Permissions name	Type	Description	Admin consent requ...	Status
Microsoft Graph (1) ...				
User.Read	Delegated	Sign in and read user profile	No	 Granted for Dave Ruijter ...
Power BI Service (1) ...				
Dataset.ReadWrite.All	Delegated	Read and write all datasets	No	 Granted for Dave Ruijter ...





# Setup

## Integration settings

- ▶ Allow XMLA endpoints and Analyze in Excel with on-premises datasets

*Enabled for the entire organization*

- ▲ Dataset Execute Queries REST API

*Enabled for the entire organization*

Users in the organization can query datasets by using Data Analysis Expressions (DAX) through Power BI REST APIs.

☒ Enabled

Apply to:

- ☒ The entire organization
- ☐ Specific security groups
- ☐ Except specific security groups

Apply

Cancel

## Admin API settings

- ▲ Allow service principals to use read-only admin APIs

*Enabled for a subset of the organization*

Web apps registered in Azure Active Directory (Azure AD) will use an assigned service principal to access read-only admin APIs without a signed in user. To allow an app to use service principal authentication, its service principal must be included in an allowed security group. By including the service principal in the allowed security group, you're giving the service principal read-only access to all the information available through admin APIs (current and future). For example, user names and emails, dataset and report detailed metadata. [Learn more](#)

☒ Enabled

Apply to:

- ☐ The entire organization
- ☒ Specific security groups

PBI\_Service\_API\_Permissions X Enter security groups

Apply

Cancel



# Impersonation

---

- Testing row-level security (RLS)



# Limitations

---

- Datasets that are hosted in Azure Analysis Services or that have a live connection to an on-premises Azure Analysis Services model aren't supported.
- The tenant setting Dataset Execute Queries REST API, found under Integration settings, must be enabled.
- One query per API call.
- One table request per query.
- Maximum of 100,000 rows or 1,000,000 values per query (whichever is hit first). For example if you query for 5 columns, you can get back max 100,000 rows. If you query for 20 columns, you can get back max 50,000 rows (1 million divided by 20).
- Maximum of 15MB of data per query. Once 15MB is exceeded, the current row will be completed but no additional rows will be written.
- Maximum of 120 requests per user per minute. Target dataset does not impact this rate limit.
- Service Principals aren't supported for datasets with RLS per RLS limitations or with SSO enabled. To use Service Principals, make sure the admin tenant setting Allow service principals to use Power BI APIs under Developer settings is enabled.



# Other notes

---

- [Announcing general availability of the ExecuteQueries REST API | Microsoft Power BI Blog | Microsoft Power BI](#)
- [Monitoring Power BI using REST APIs from Python — DATA GOBLINS \(data-goblins.com\)](#)



03

# Azure Synapse Analytics

---





# Setup

---

- Adding packages (e.g. `great_expectations`) is more complex if you have Data exfiltration protection enabled!



# Take aways on great\_expectations

---

- Impressive
- Configuration nightmare
- Timesaver



# Session Feedback



dave@blue-rocket.it



@DaveRuijter



linkedin.com/in/DaveRuijter



ModernData.ai

[https://bit.ly/dMC2022\\_SessionFeedback](https://bit.ly/dMC2022_SessionFeedback)