# Exploratory Data Analysis

March 19, 2025

## 1 Personal Information

Name: **Sijn Hering**

StudentID: **12377422**

Email: **stijn.hering@student.uva.nl**

Submitted on: **23-03-2025**

GitHub: **https://github.com/melchiorhering/IS-master-thesis**

## 2 Data Context

This dataset is sourced from an evaluation examples framework designed to test various data applications. It focuses on tasks across 21 professional applications including BigQuery, Snowflake, MySQL, and others. We will be using the Spider2-V data tasks, which represent structured examples with associated configurations, evaluations, and account requirements. Each example contains metadata like unique IDs, instructions, action counts, and evaluation methods, primarily aimed at supporting retrieval augmented generation for data processing workflows.

## 3 Data Description: Evaluation Examples Framework

This document provides an overview of the evaluation examples framework used for testing and benchmarking retrieval-augmented generation tasks across various professional applications.

### 3.1 Overview

The framework consists of structured task examples designed to evaluate performance across 21 professional applications, including data warehousing tools, analysis platforms, and development environments. Each task includes detailed configuration information, evaluation metrics, and documentation references.

### 3.2 Framework Components

#### 3.2.1 1. Document Warehouse (`documents/`)

- Contains crawled and pre-processed HTML from official documentation
- Supports retrieval augmented generation framework
- Organized by application type

#### 3.2.2 2. Task Format Examples (`examples/`)

Each task example is defined by a JSON configuration file with the following key fields:

| Field | Description |
|---|---|
| `id` | Globally unique UUID |
| `snapshot` | The focused professional application |
| `instruction` | Task goal or user intent |
| `source` | List of URLs used to construct the example |
| `action_number` | Number of steps required to complete the task |
| `related_apps` | List of professional applications used |
| `tags` | Categorization labels |
| `config` | Environment setup functions |
| `evaluator` | Metrics and comparison methods |
| `counterpart` | Reference to related example |

#### 3.2.3 3. Real Accounts (`settings/`)

- Credentials for applications requiring authentication
- JSON templates for account information
- Supports applications like Snowflake, BigQuery, ServiceNow, etc.

### 3.3 Application Coverage

The framework spans 21 professional applications:

- **Data Warehousing**: Snowflake, BigQuery, DuckDB
- **Data Processing**: Airflow, Dagster, DBT
- **Databases**: MySQL, PostgreSQL, SQLite3
- **Development**: VSCode, Docker, Terminal
- **Visualization**: Metabase, Superset, Jupyter
- **Other**: ServiceNow, Airbyte, Chromium, Excel, etc.

### 3.4 Task Categorization

Tasks are categorized by various tags including:

- Interface type: CLI, GUI, or both
- Instruction detail: Verbose vs. abstract
- Authentication requirements
- Functional category (7 types):
  - Data ingestion and integration
  - Data warehousing
  - Data orchestration
  - Data analysis and visualization
  - Traditional data processing
  - IT service management
  - Data transformation

### 3.5 Evaluation Methodology

Tasks are evaluated using predefined metrics that compare predicted outputs against golden results. The framework supports various evaluation methods including:

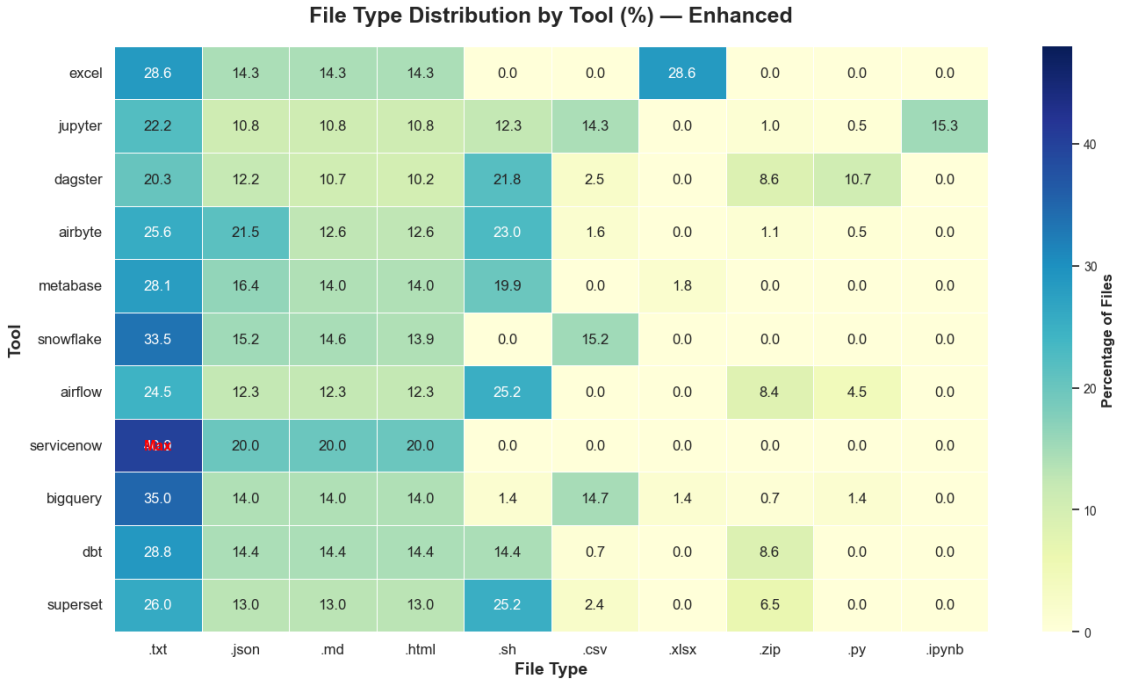
- File comparison (e.g., CSV comparison)
- VM state extraction
- Result validation against golden outputs

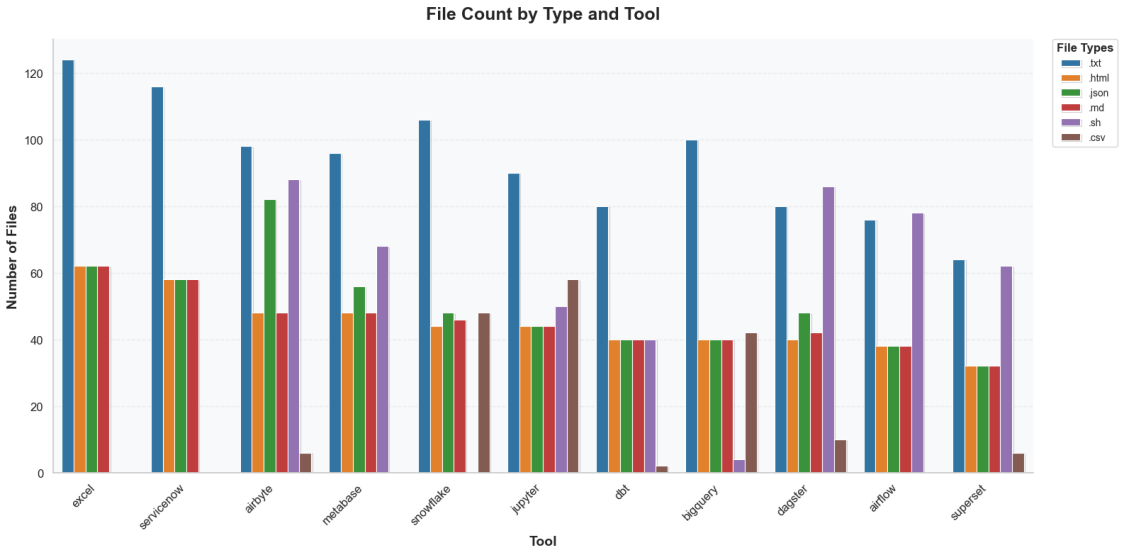
[1]:
```
# Imports
import os
import numpy as np
import pandas as pd

from anaylze import BenchmarkAnalyzer
```

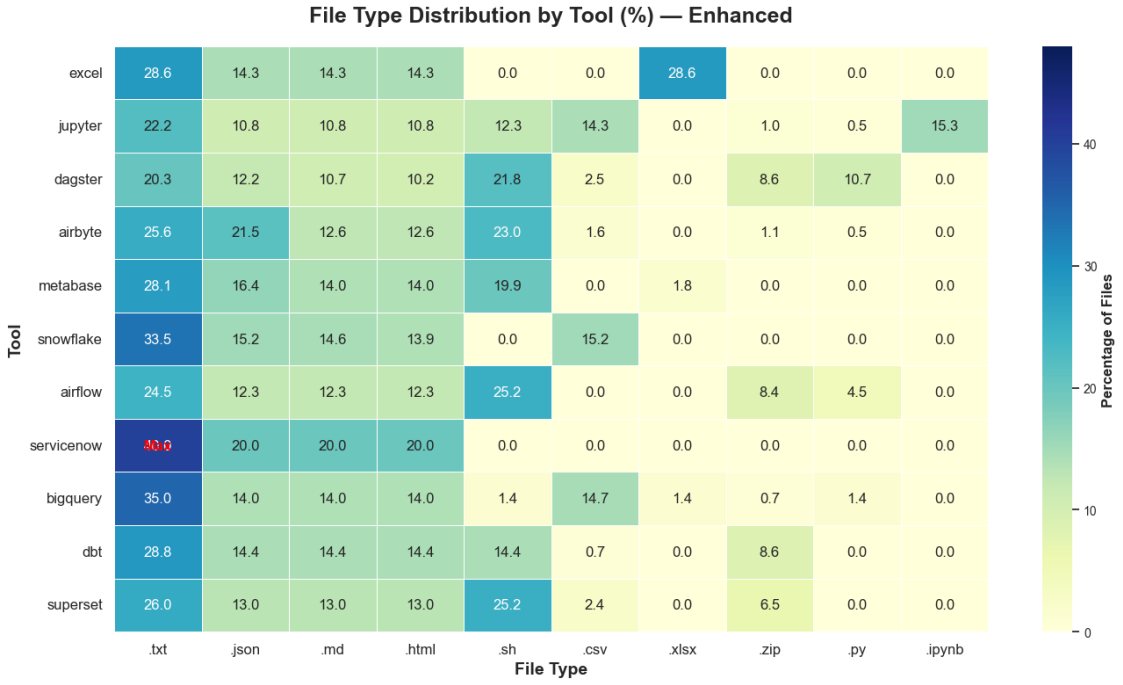
#### 3.5.1 Data Loading

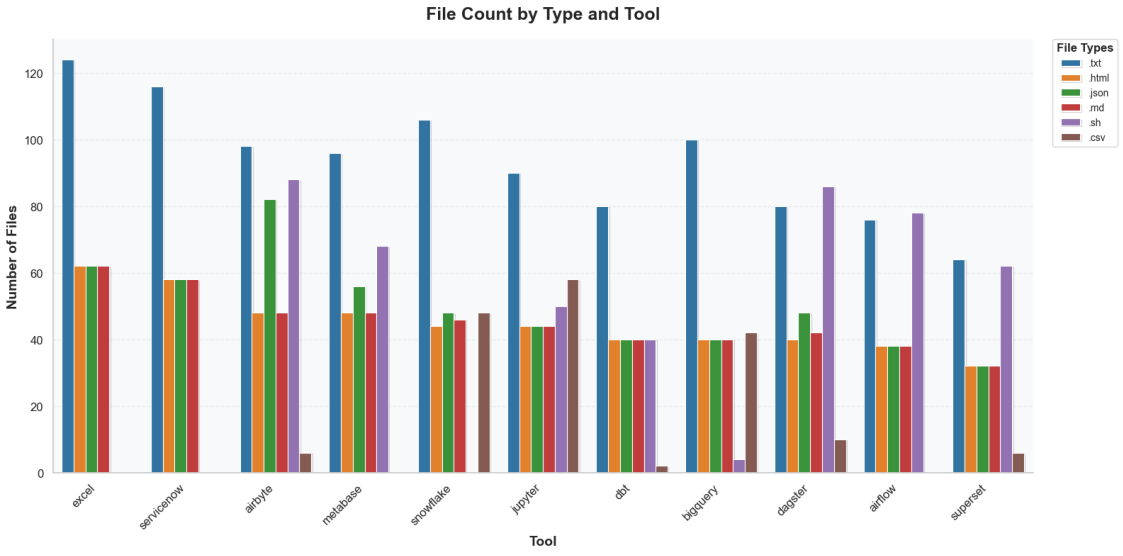
[2]:
```
# Load your data here
analyzer = BenchmarkAnalyzer(root_dir="../evaluation_examples/examples")


# Get comprehensive file type insights
analyzer.display_file_type_report()

# Visualize file type distribution
analyzer.plot_file_type_distribution()

# Visualize benchmark structure
analyzer.visualize_benchmark_structure()

# Fixed connection visualization (handles the Airbyte-only case)
analyzer.plot_connection_analysis()

# Visualize the benchmarking tool comparison
analyzer.plot_tool_comparison()


analyzer.get_file_signature_by_tool()

# Visualize the benchmarking tool comparison
# analyzer.plot_json_complexity()
```

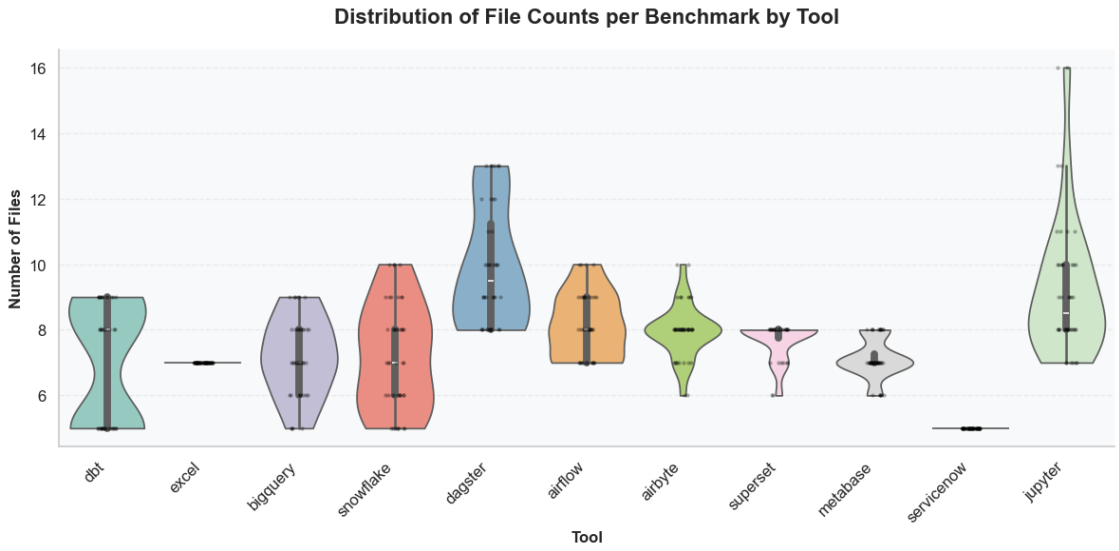
```
<IPython.core.display.HTML object>
```

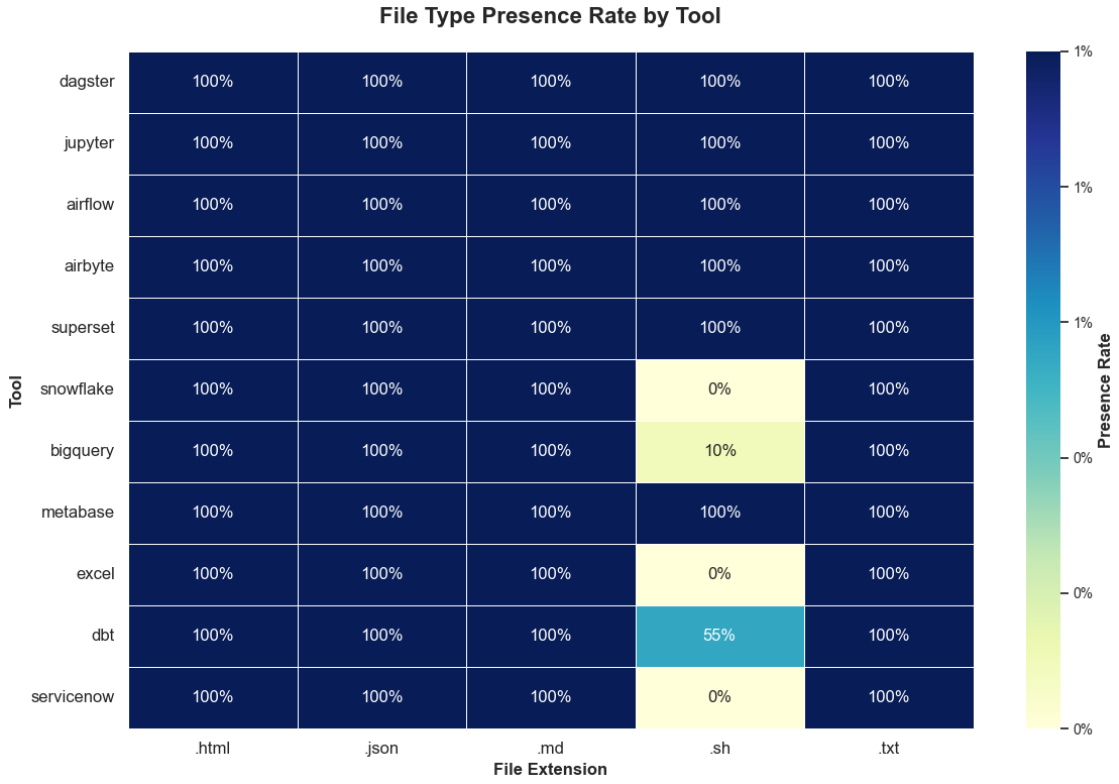

File Type Distribution by Tool (%) — Enhanced
excel 28.6 14.3 14.3 14.3 0.0 0.0 28.6 0.0 0.0 0.0
jupyter 22.2 10.8 10.8 10.8 12.3 14.3 0.0 1.0 0.5 15.3
dagster 20.3 12.2 10.7 10.2 21.8 2.5 0.0 8.6 10.7 0.0
airbyte 25.6 21.5 12.6 12.6 23.0 1.6 0.0 1.1 0.5 0.0
metabase 28.1 16.4 14.0 14.0 19.9 0.0 1.8 0.0 0.0 0.0
snowflake 33.5 15.2 14.6 13.9 0.0 15.2 0.0 0.0 0.0 0.0
airflow 24.5 12.3 12.3 12.3 25.2 0.0 0.0 8.4 4.5 0.0
servicenow 40.0 20.0 20.0 20.0 0.0 0.0 0.0 0.0 0.0 0.0
bigquery 35.0 14.0 14.0 14.0 1.4 14.7 1.4 0.7 1.4 0.0
dbt 28.8 14.4 14.4 14.4 14.4 0.7 0.0 8.6 0.0 0.0
superset 26.0 13.0 13.0 13.0 25.2 2.4 0.0 6.5 0.0 0.0
Tool
.txt .json .md .html .sh .csv .xlsx .zip .py .ipynb
File Type
Percentage of Files
0 10 20 30 40

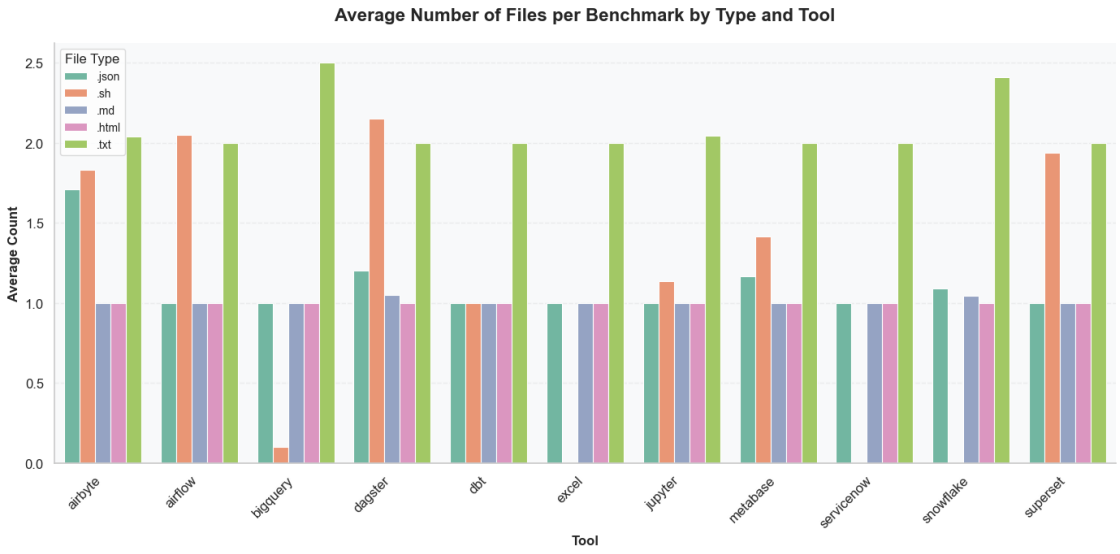

File Count by Type and Tool
Number of Files
0 20 40 60 80 100 120
excel servicenow airbyte metabase snowflake jupyter dbt bigquery dagster airflow superset
Tool
File Types
.txt
.html
.json
.md
.sh
.csv

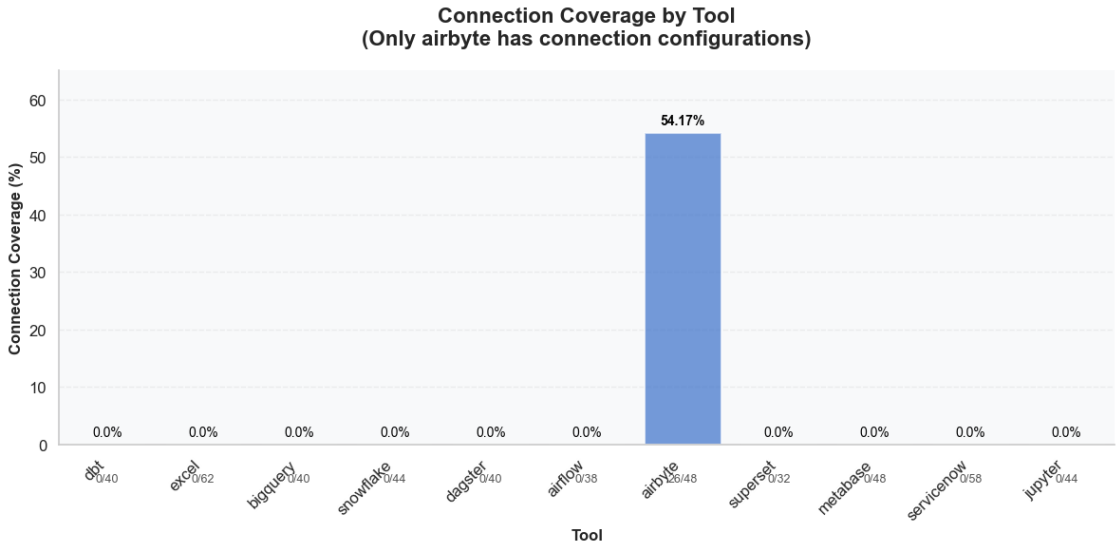

File Type Distribution by Tool (%) — Enhanced
excel 28.6 14.3 14.3 14.3 0.0 0.0 28.6 0.0 0.0 0.0
jupyter 22.2 10.8 10.8 10.8 12.3 14.3 0.0 1.0 0.5 15.3
dagster 20.3 12.2 10.7 10.2 21.8 2.5 0.0 8.6 10.7 0.0
airbyte 25.6 21.5 12.6 12.6 23.0 1.6 0.0 1.1 0.5 0.0
metabase 28.1 16.4 14.0 14.0 19.9 0.0 1.8 0.0 0.0 0.0
snowflake 33.5 15.2 14.6 13.9 0.0 15.2 0.0 0.0 0.0 0.0
airflow 24.5 12.3 12.3 12.3 25.2 0.0 0.0 8.4 4.5 0.0
servicenow 40.0 20.0 20.0 20.0 0.0 0.0 0.0 0.0 0.0 0.0
bigquery 35.0 14.0 14.0 14.0 1.4 14.7 1.4 0.7 1.4 0.0
dbt 28.8 14.4 14.4 14.4 14.4 0.7 0.0 8.6 0.0 0.0
superset 26.0 13.0 13.0 13.0 25.2 2.4 0.0 6.5 0.0 0.0
Tool
.txt .json .md .html .sh .csv .xlsx .zip .py .ipynb
File Type
Percentage of Files
0 10 20 30 40


File Count by Type and Tool
Number of Files
0 20 40 60 80 100 120
excel servicenow airbyte metabase snowflake jupyter dbt bigquery dagster airflow superset
Tool
File Types
.txt
.html
.json
.md
.sh
.csv


Distribution of File Counts per Benchmark by Tool
Number of Files
16
14
12
10
8
6
dbt
excel
bigquery
snowflake
dagster
airflow
airbyte
superset
metabase
servicenow
jupyter
Tool

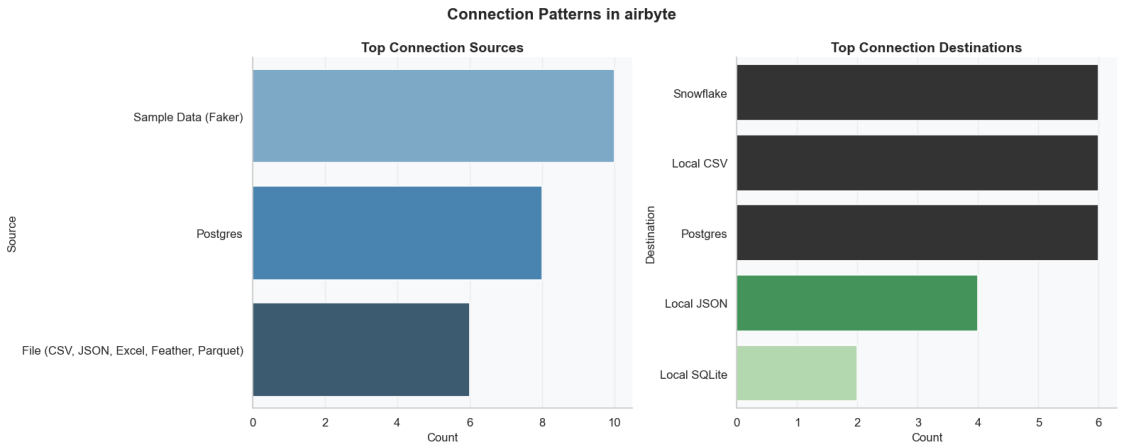

File Type Presence Rate by Tool
Tool
dagster 100% 100% 100% 100% 100%
jupyter 100% 100% 100% 100% 100%
airflow 100% 100% 100% 100% 100%
airbyte 100% 100% 100% 100% 100%
superset 100% 100% 100% 100% 100%
snowflake 100% 100% 100% 0% 100%
bigquery 100% 100% 100% 10% 100%
metabase 100% 100% 100% 100% 100%
excel 100% 100% 100% 0% 100%
dbt 100% 100% 100% 55% 100%
servicenow 100% 100% 100% 0% 100%
.html
.json
.md
.sh
.txt
File Extension
Presence Rate
1%
1%
1%
0%
0%
0%



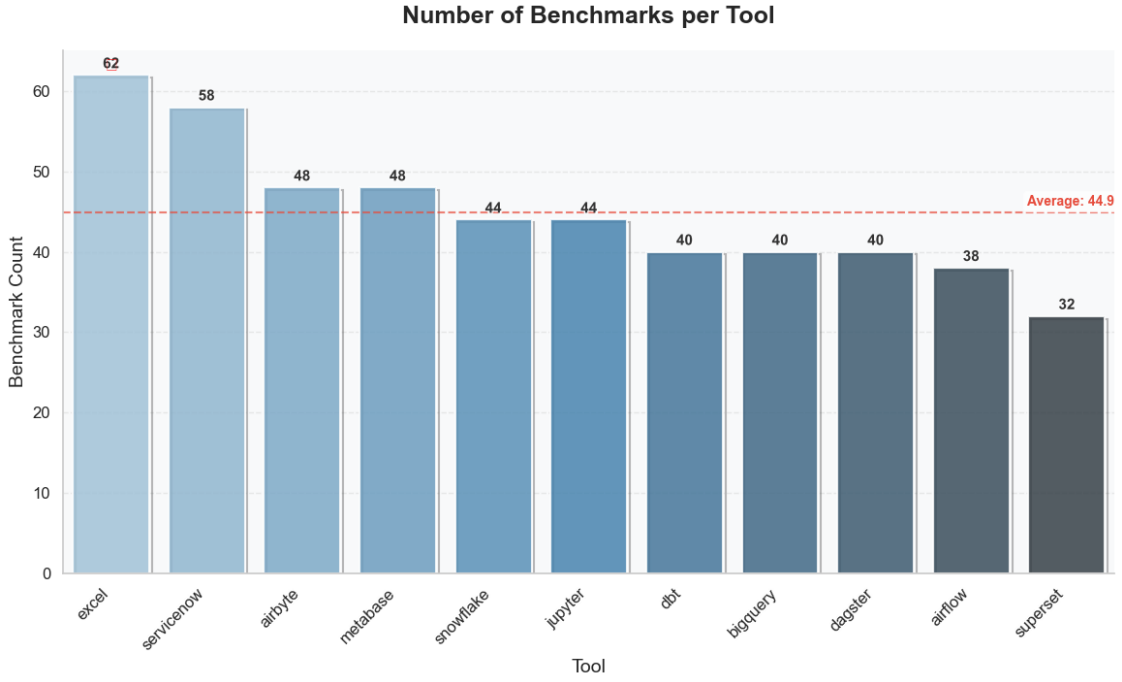




Note: Only airbyte contains connection configurations in this dataset.
Found 26 connections in 48 benchmarks.

<Figure size 1400x600 with 0 Axes>




```
/Users/stijn.hering/Desktop/Master-Thesis/.venv/lib/python3.13/site-
packages/IPython/core/pylabtools.py:170: UserWarning: Glyph 9733 (\N{BLACK
STAR}) missing from font(s) Arial.
  fig.canvas.print_figure(bytes_io, **kw)
```




[2]:

| | tool | benchmark_count | avg_files_per_benchmark | \ |
|---|---|---|---|---|
| 0 | dbt | 40 | 6.95 | |
| 1 | excel | 62 | 7.00 | |
| 2 | bigquery | 40 | 7.15 | |

```
3    snowflake              44                    7.18
4      dagster              40                    9.90
5      airflow              38                    8.16
6      airbyte              48                    7.96
7     superset              32                    7.69
8     metabase              48                    7.12
9   servicenow              58                    5.00
10     jupyter              44                    9.23

                                          common_files semi_common_files  \
0   retrieved_chunk_size_512_chunk_overlap_20_topk…  eval.sh, init.sh
1   retrieved_chunk_size_512_chunk_overlap_20_topk…
2   retrieved_chunk_size_512_chunk_overlap_20_topk…
3   retrieved_chunk_size_512_chunk_overlap_20_topk…
4   eval.sh, init.sh, retrieved_chunk_size_512_chu…
5   eval.sh, init.sh, retrieved_chunk_size_512_chu…
6   eval.sh, init.sh, retrieved_chunk_size_512_chu…   connection.json
7   eval.sh, init.sh, retrieved_chunk_size_512_chu…
8   init.sh, retrieved_chunk_size_512_chunk_overla…           eval.sh
9   retrieved_chunk_size_512_chunk_overlap_20_topk…
10  init.sh, retrieved_chunk_size_512_chunk_overla…

        top_extensions  unique_file_count
0     .txt, .sh, .json                 57
1   .xlsx, .txt, .html                111
2    .txt, .csv, .html                 72
3    .txt, .json, .csv                 84
4     .sh, .txt, .json                 91
5     .sh, .txt, .json                 60
6     .txt, .sh, .json                 66
7     .txt, .sh, .html                 44
8     .txt, .sh, .json                 71
9   .txt, .html, .json                 62
10  .txt, .ipynb, .csv                103
```

[ ]: