AIRFLOW SUMMIT

Simplifying Data Lineage: How OpenLineage Empowers Airflow and Beyond

3.0

Julien Le Dem & Harel Shein, Datadog

Who we are

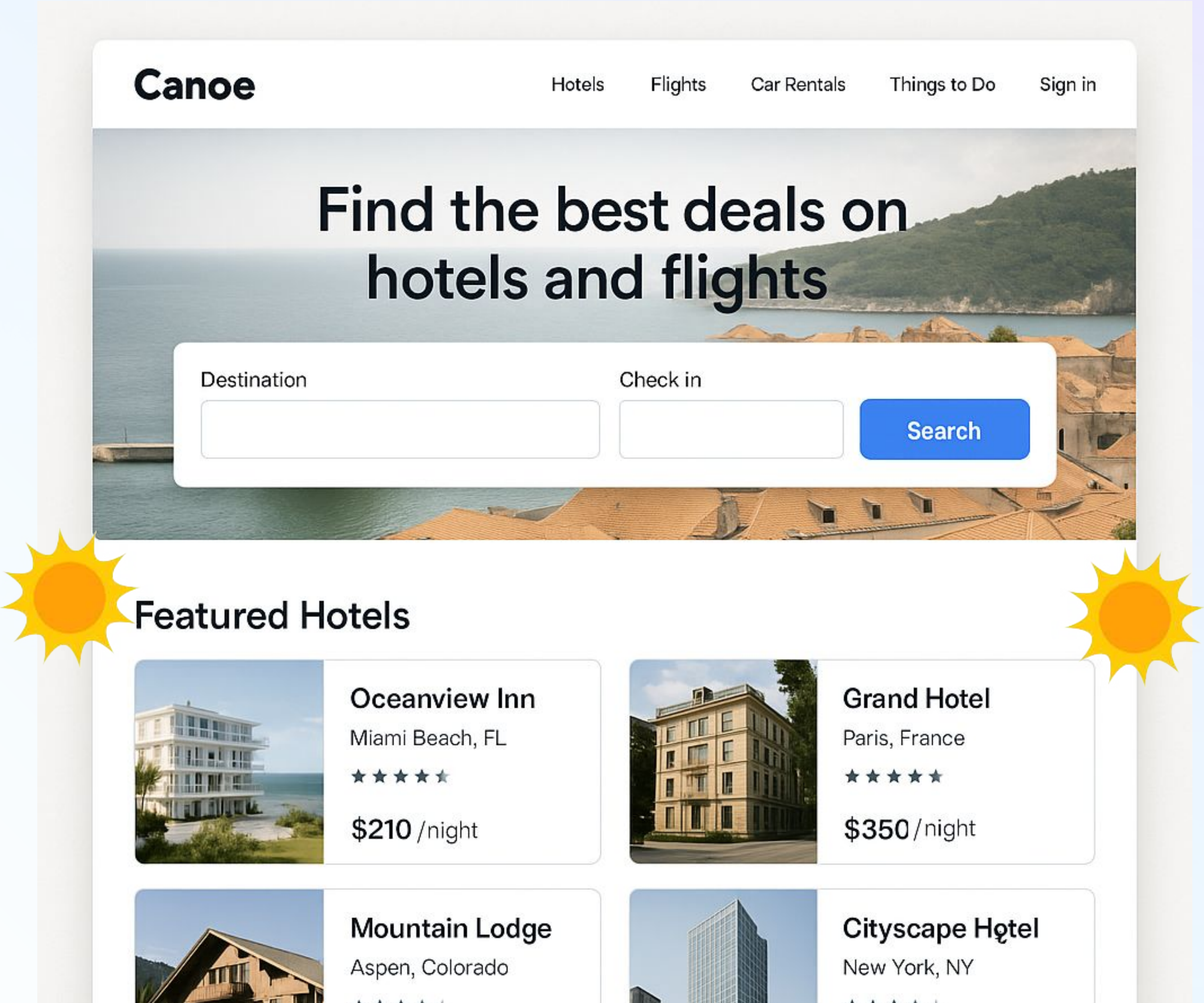
Datadog is a world-class data platform ingesting more than a 100 trillion events a day, providing real-time insights.

Harel Shein
Senior Engineering Manager
OpenLineage TSC

Julien Le Dem
Principal Engineer
co-creator of Parquet, Arrow
and OpenLineage

DATADOG

Imagine we work at Canoe



Canoe

Find the best deals on hotels and flights

Destination

Check in

Featured Hotels

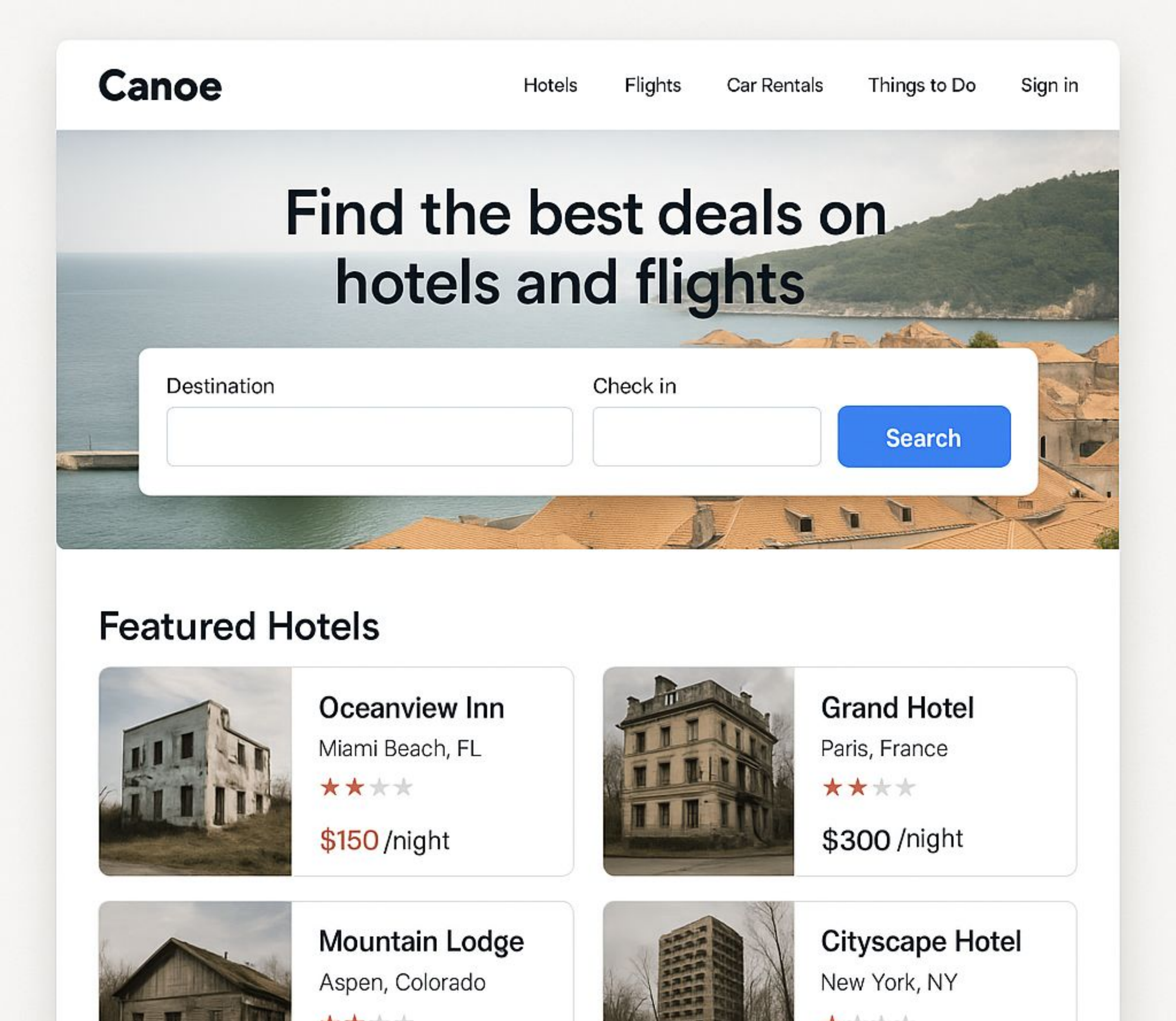
Oceanview Inn
Miami Beach, FL
★★★★½
$210 /night

Grand Hotel
Paris, France
★★★★★
$350 /night

Mountain Lodge
Aspen, Colorado

Cityscape Hǫtel
New York, NY

But something is wrong...



Find the best deals on hotels and flights

Featured Hotels

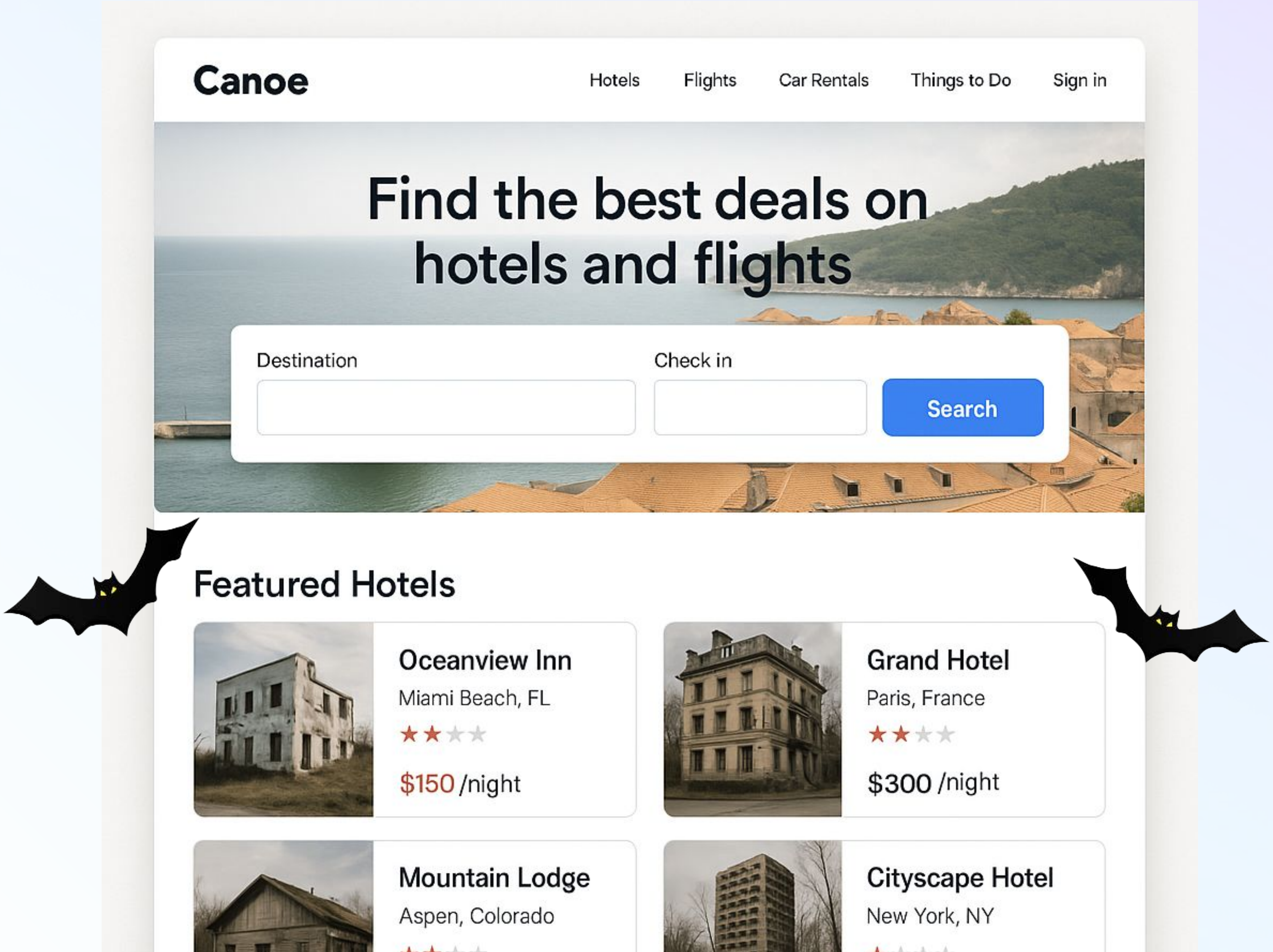
Oceanview Inn
Miami Beach, FL
★★☆☆
$150 /night

Grand Hotel
Paris, France
★★☆☆
$300 /night

Mountain Lodge
Aspen, Colorado

Cityscape Hotel
New York, NY

But something is wrong...



Canoe

Hotels Flights Car Rentals Things to Do Sign in

Find the best deals on hotels and flights

Destination

Check in

Search

Featured Hotels

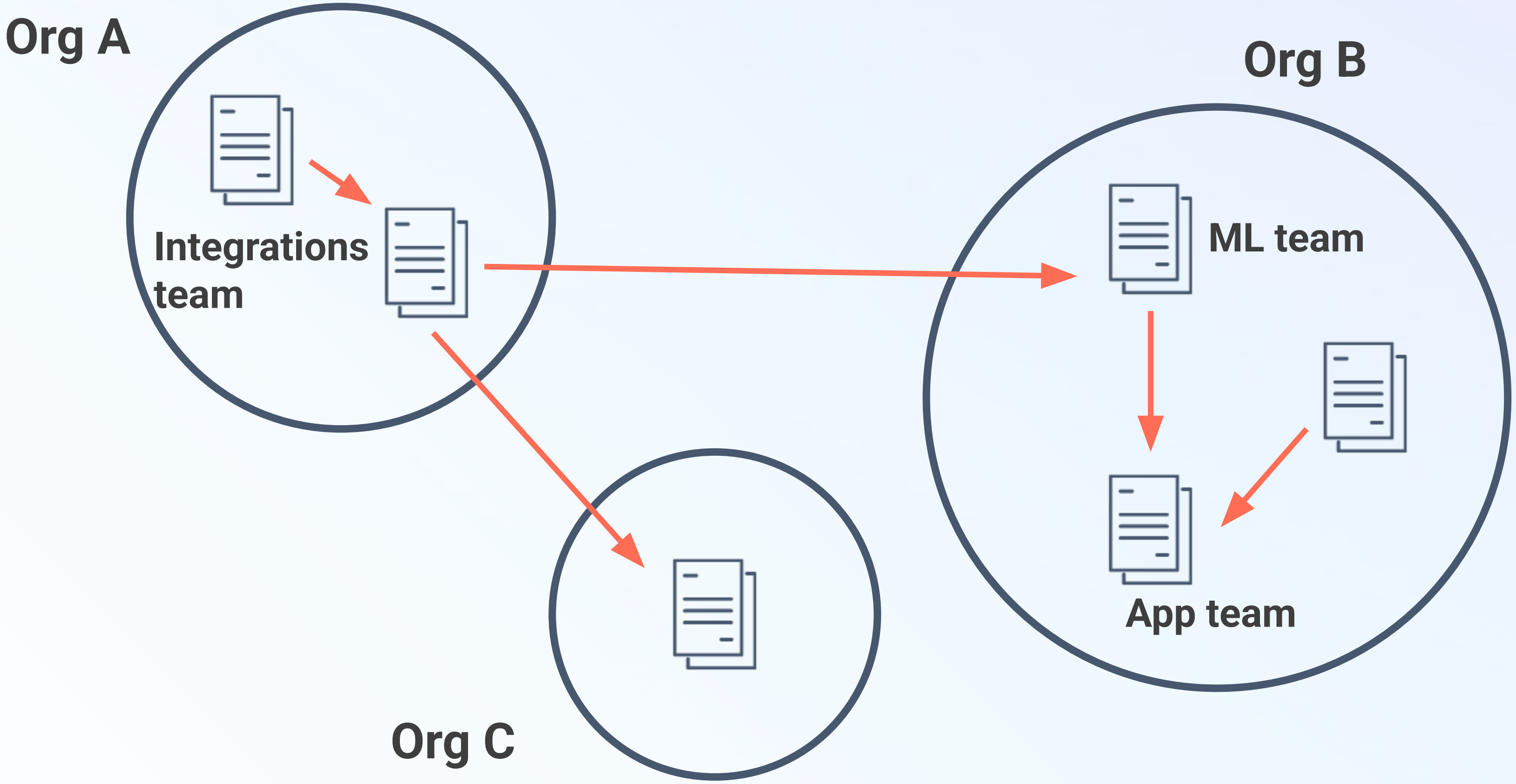
Oceanview Inn
Miami Beach, FL
★★☆☆
$150 /night

Grand Hotel
Paris, France
★★☆☆
$300 /night

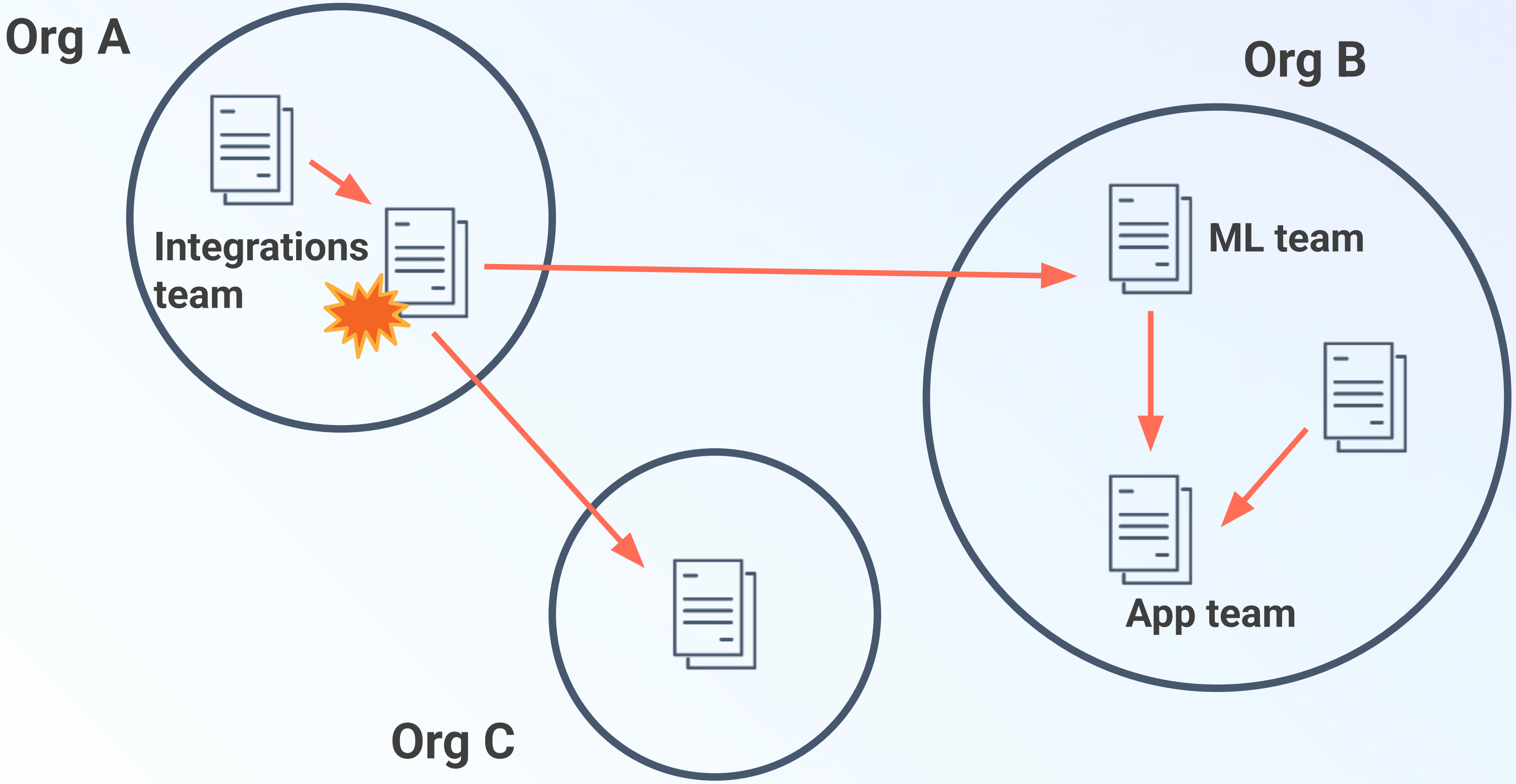
Mountain Lodge
Aspen, Colorado

Cityscape Hotel
New York, NY

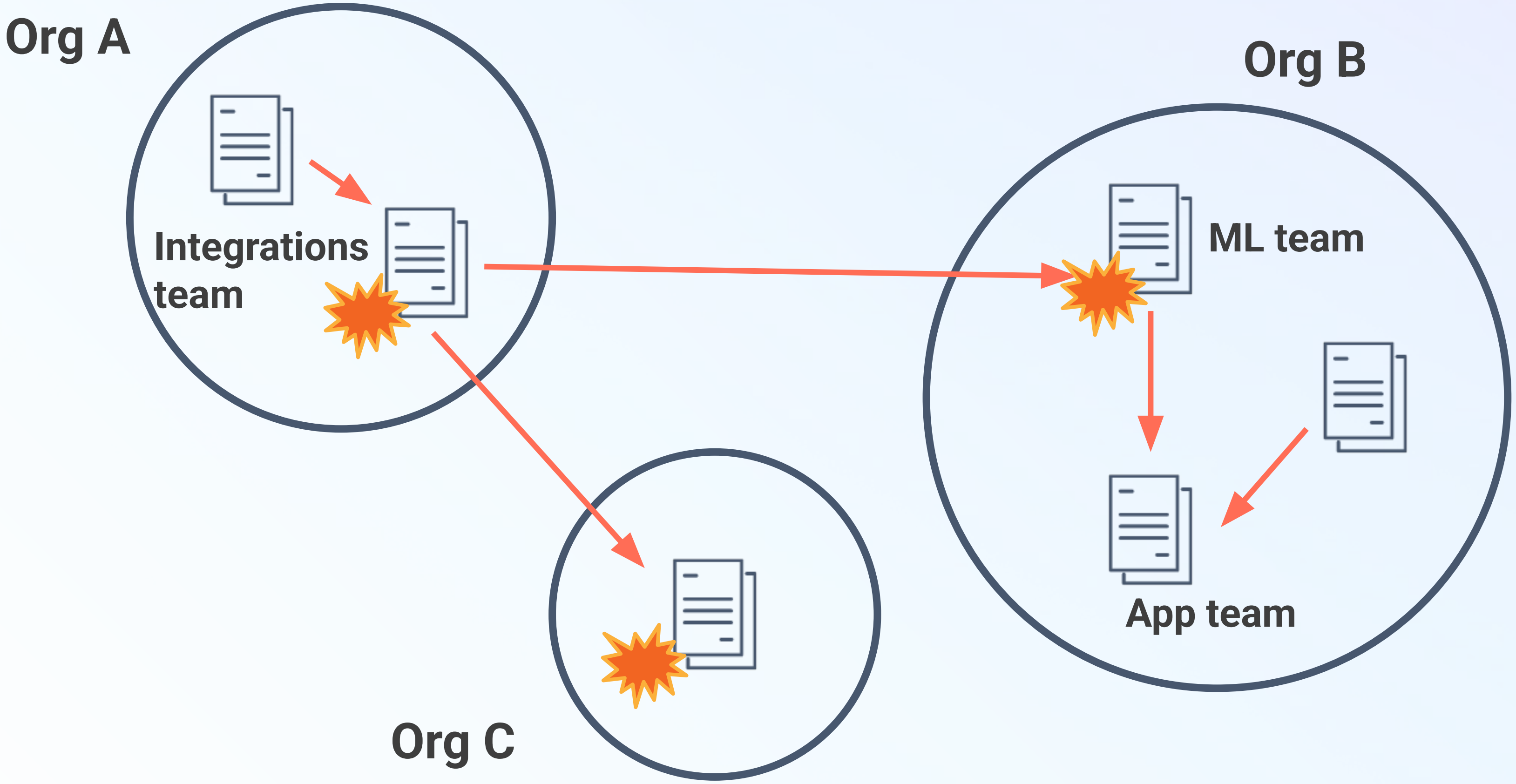
Data ecosystem at Canoe



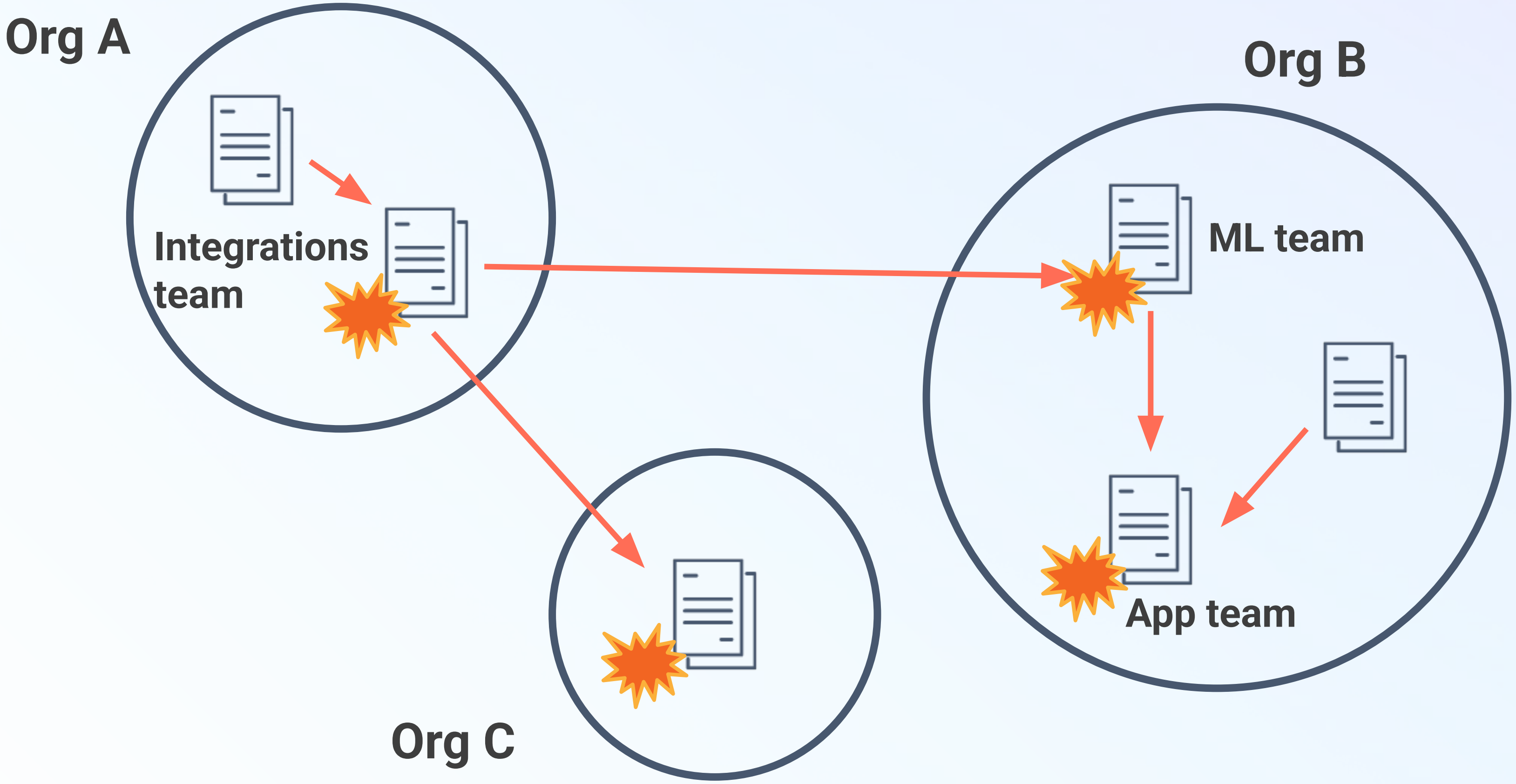
Data ecosystem at Canoe

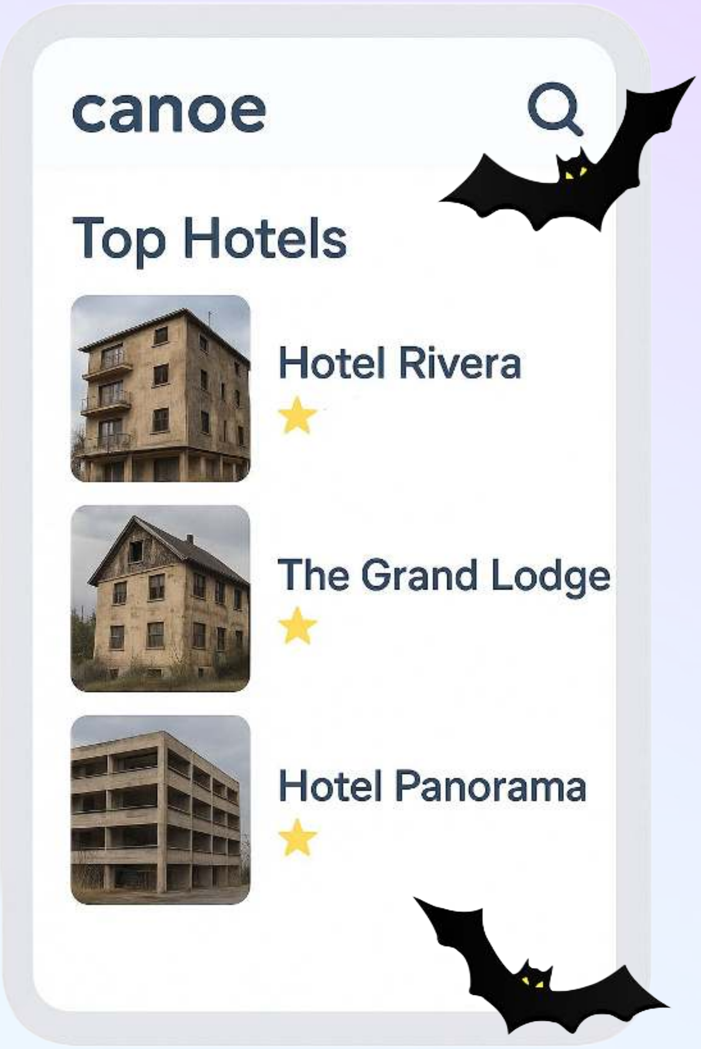


Data ecosystem at Canoe

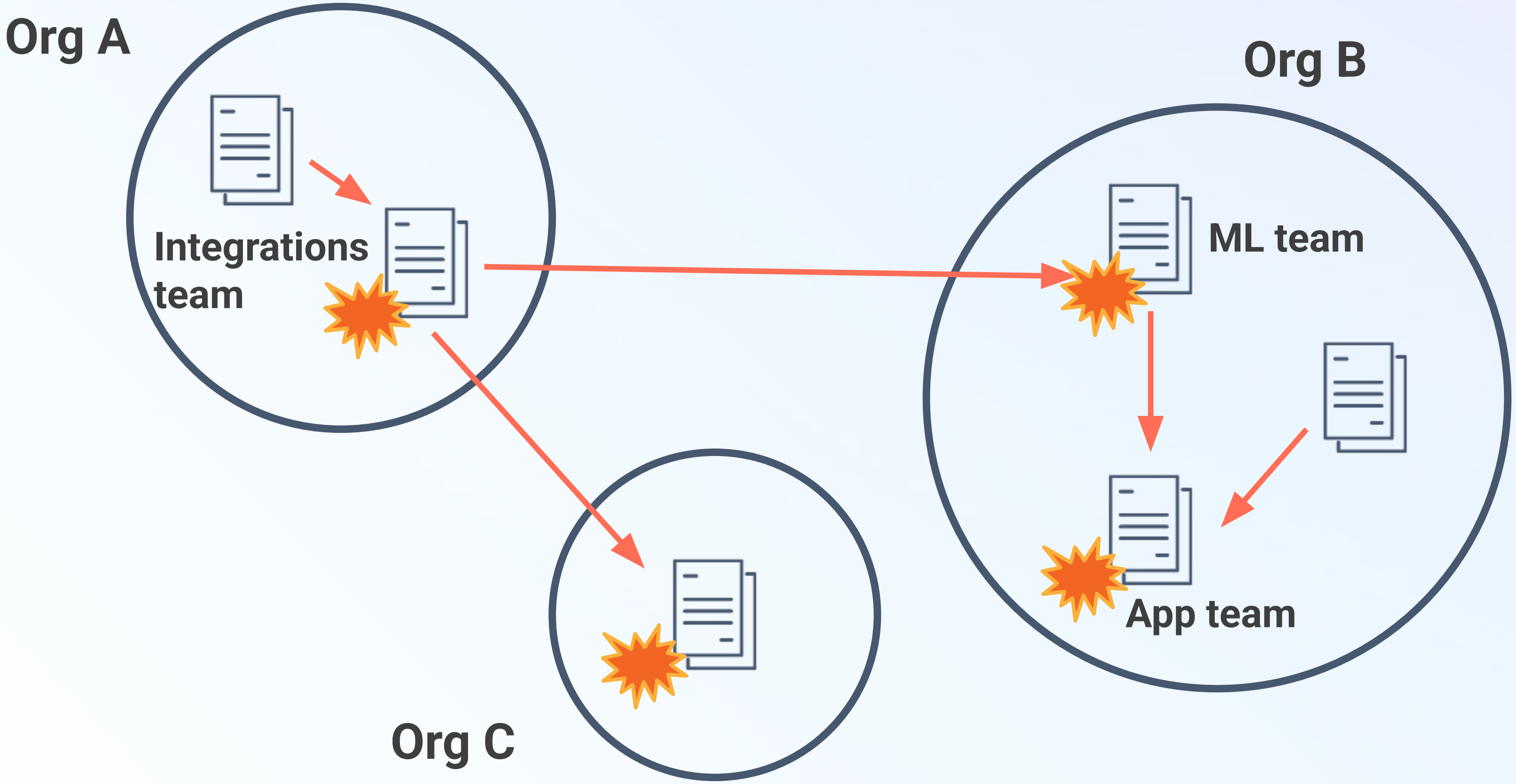


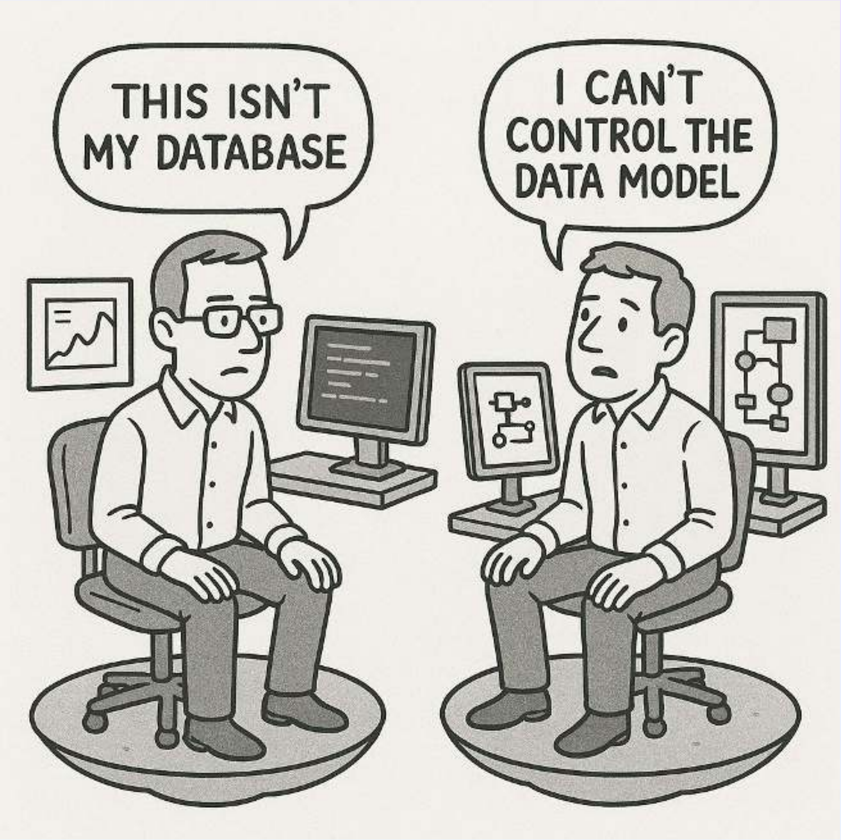
Data ecosystem at Canoe





Data ecosystem at Canoe

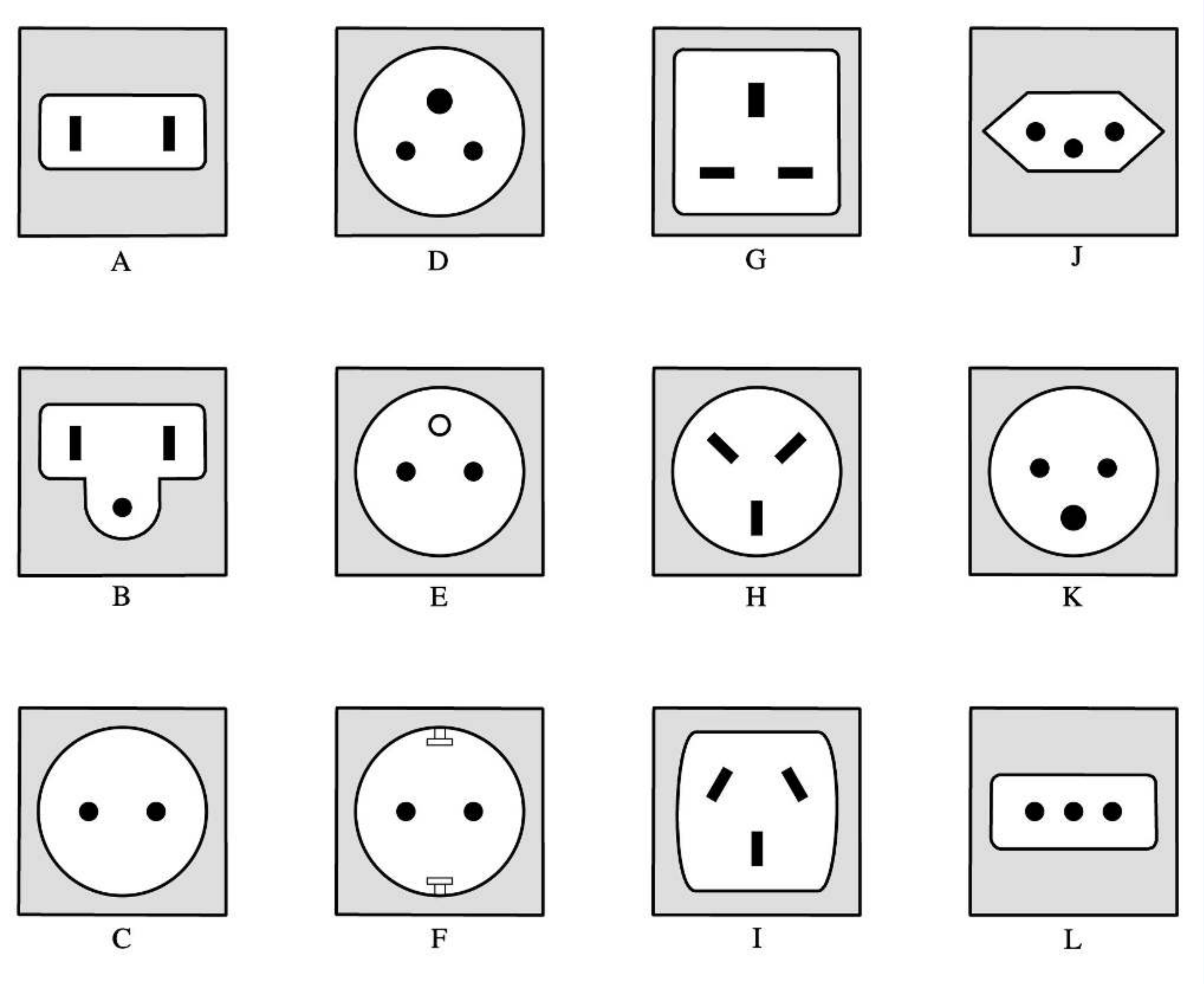




And the ecosystem is fragmented

kafka
dagster
Apache Airflow
mlflow™
RAY
HIVE
ICEBERG
dbt
trino
APACHE Spark™
Parquet

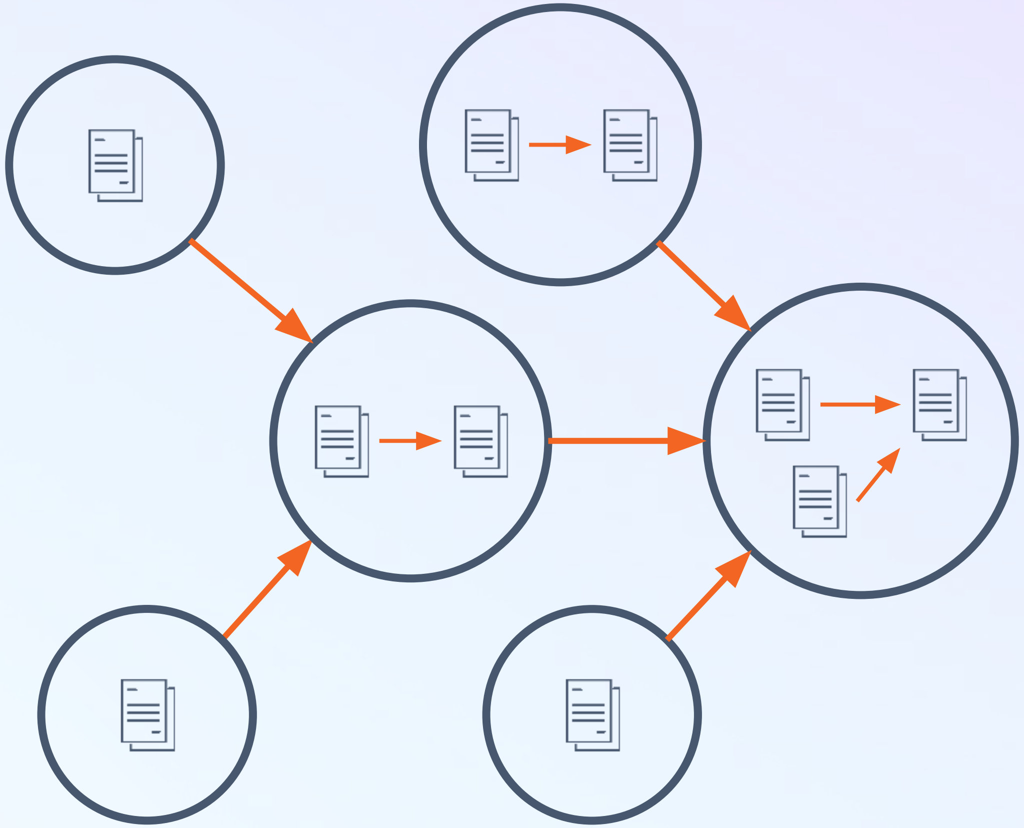
We need everyone to speak the same language

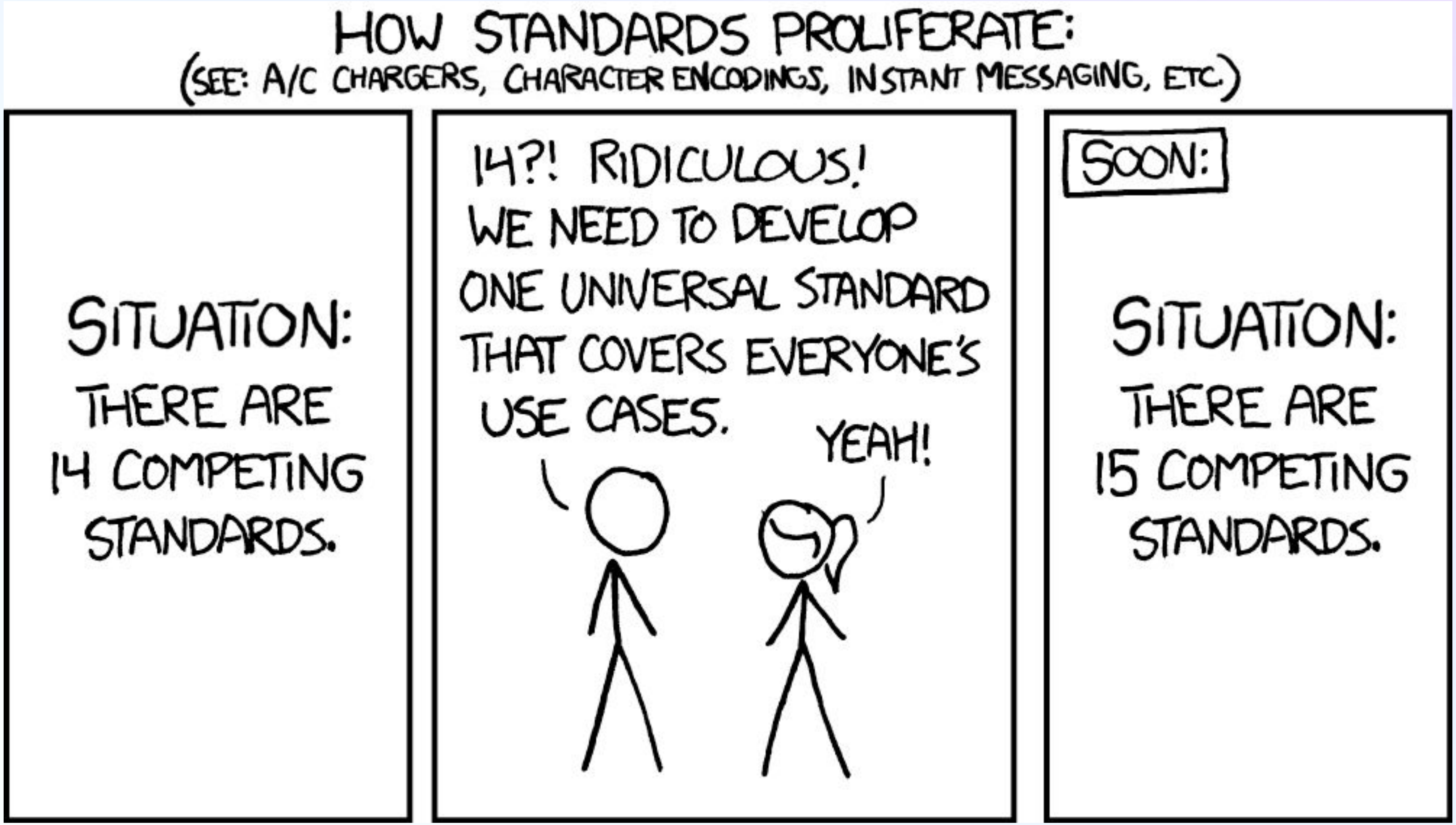


Need a common language

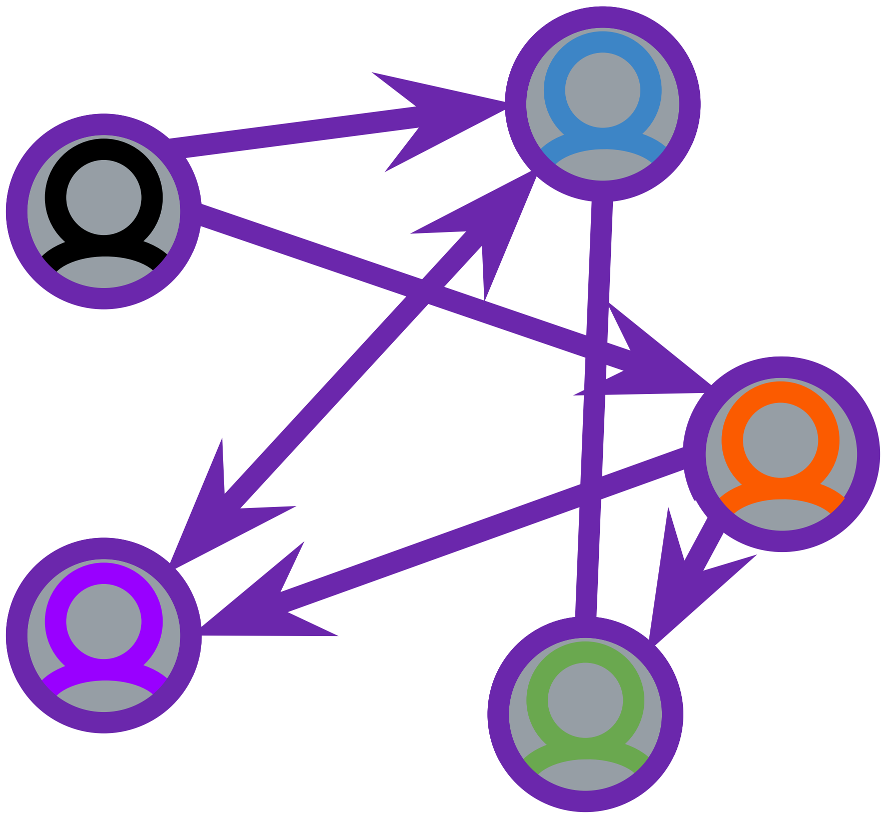
Metadata about data pipelines:

- where data is from
- how it changes
- where it goes



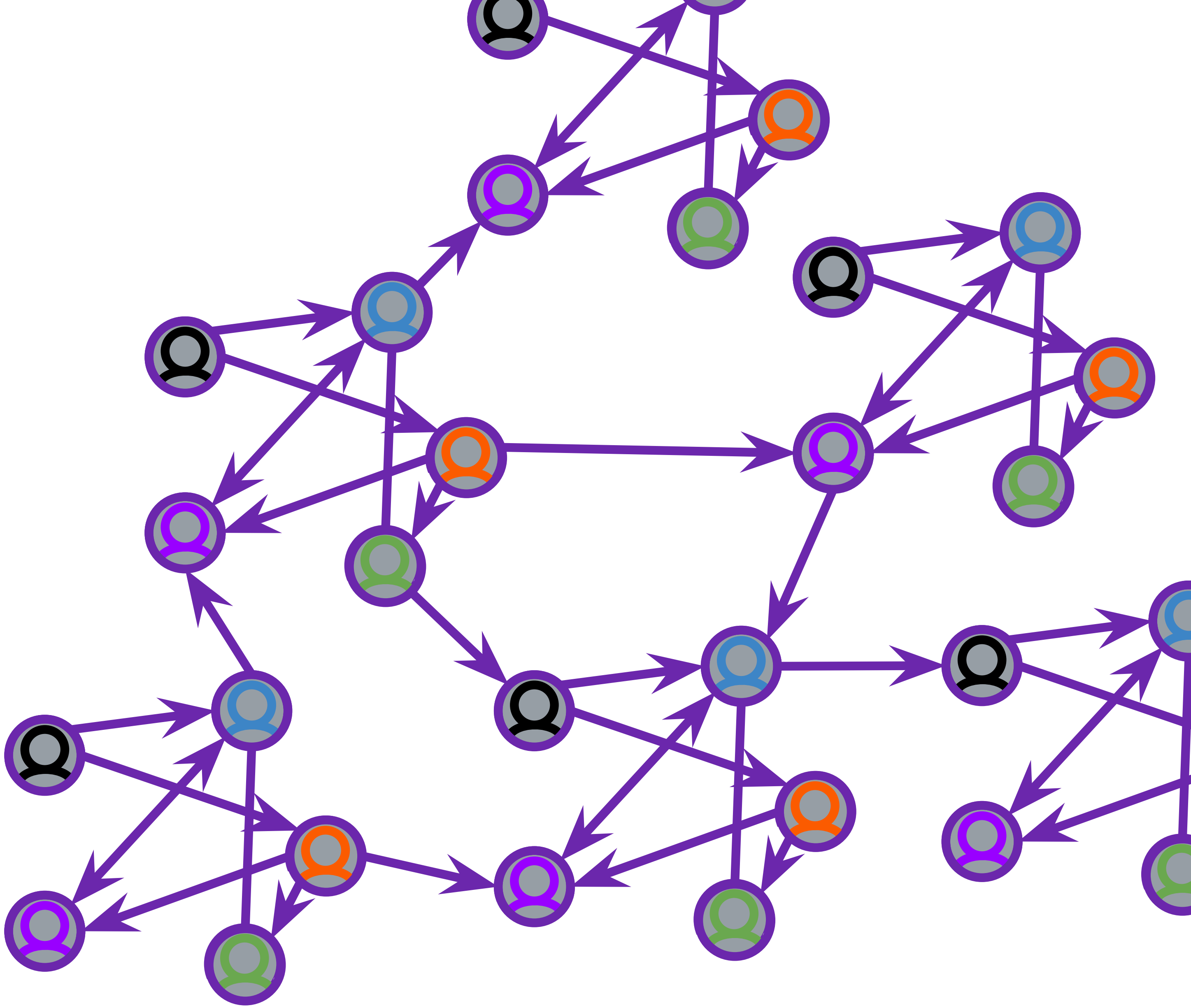

HOW STANDARDS PROLIFERATE:
(SEE: A/C CHARGERS, CHARACTER ENCODINGS, INSTANT MESSAGING, ETC.)
SITUATION: THERE ARE 14 COMPETING STANDARDS.
14?! RIDICULOUS! WE NEED TO DEVELOP ONE UNIVERSAL STANDARD THAT COVERS EVERYONE'S USE CASES.
YEAH!
SOON:
SITUATION: THERE ARE 15 COMPETING STANDARDS.

How do standards get started?



We Make Stone Soup

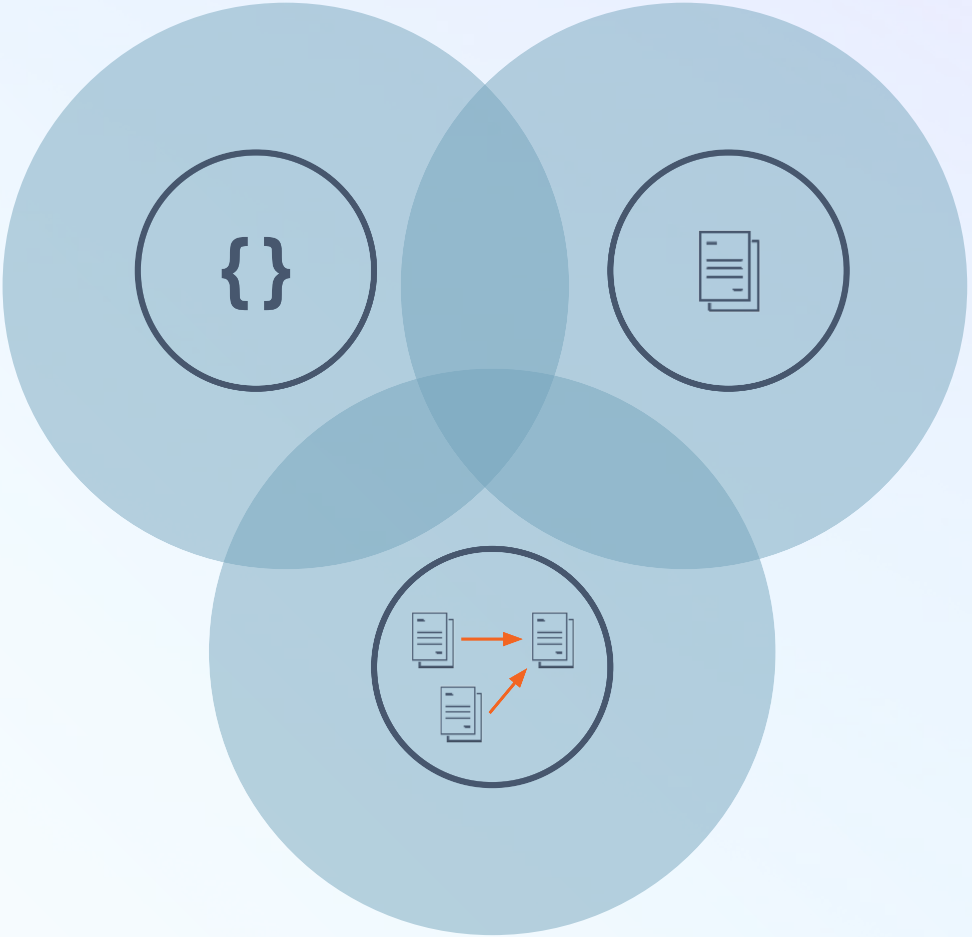
Align incentives to build a network effect



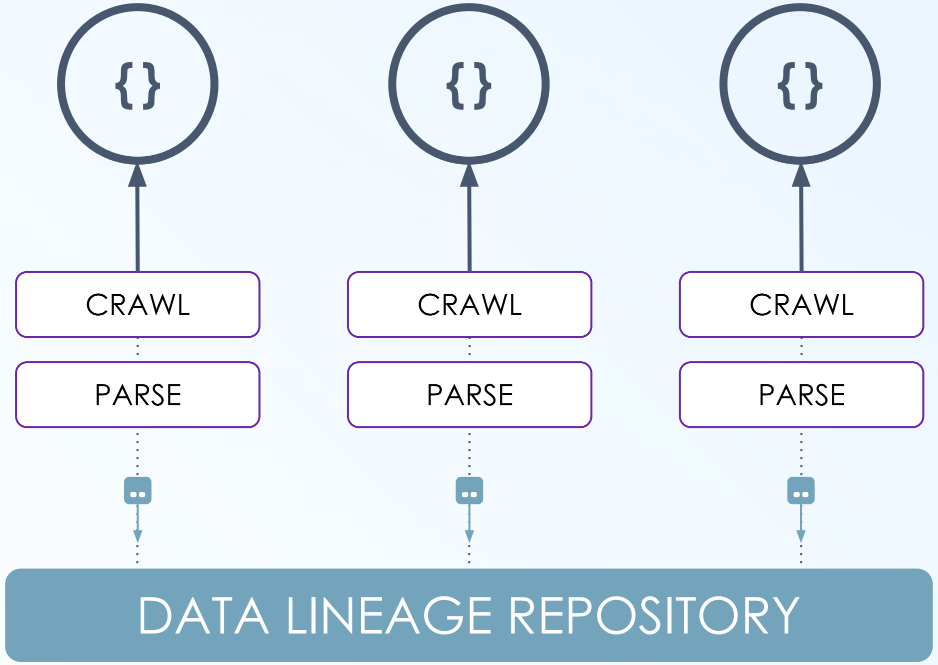
The snowball effect

so…. OpenLineage

How to observe the lineage?

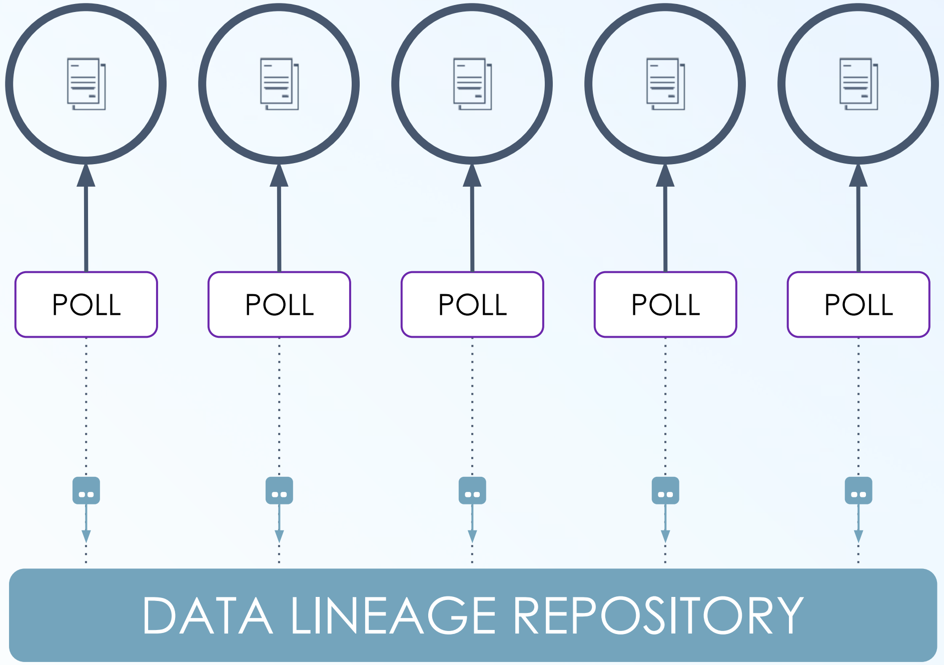


Analyze source code



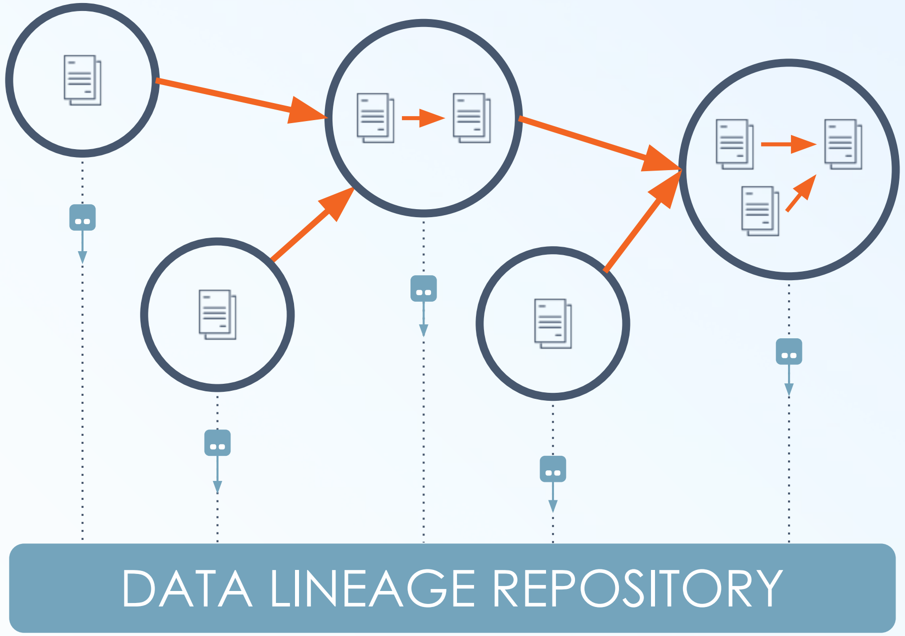
- Integrate with source code repositories
- Look for queries and parse them for lineage
- Report to a lineage metadata repository

Process query or activity logs



- Integrate with data stores and warehouses
- Regularly process query logs to trace lineage
- Report to a lineage metadata repository

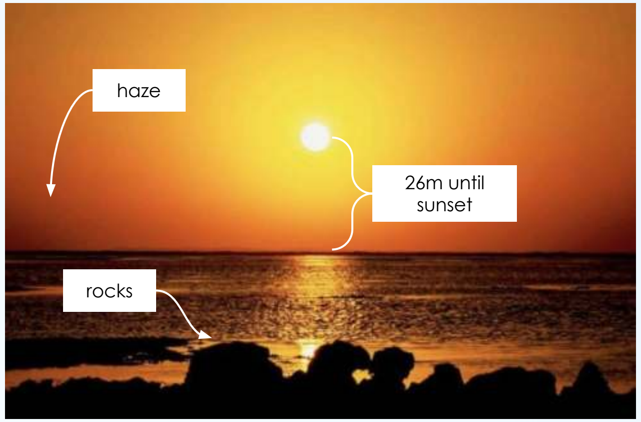
Observe the pipeline



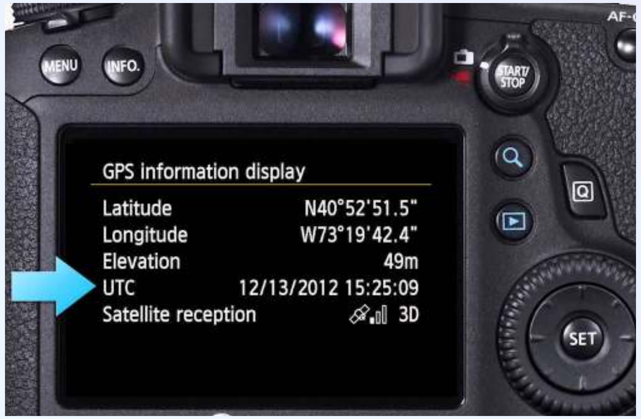
- Integrate with data orchestration systems
- As jobs run, observe the way they affect data
- Report to a lineage metadata repository

Why Collect Data at Runtime?

The best time to collect metadata



You can try to infer data after the fact…



GPS information display

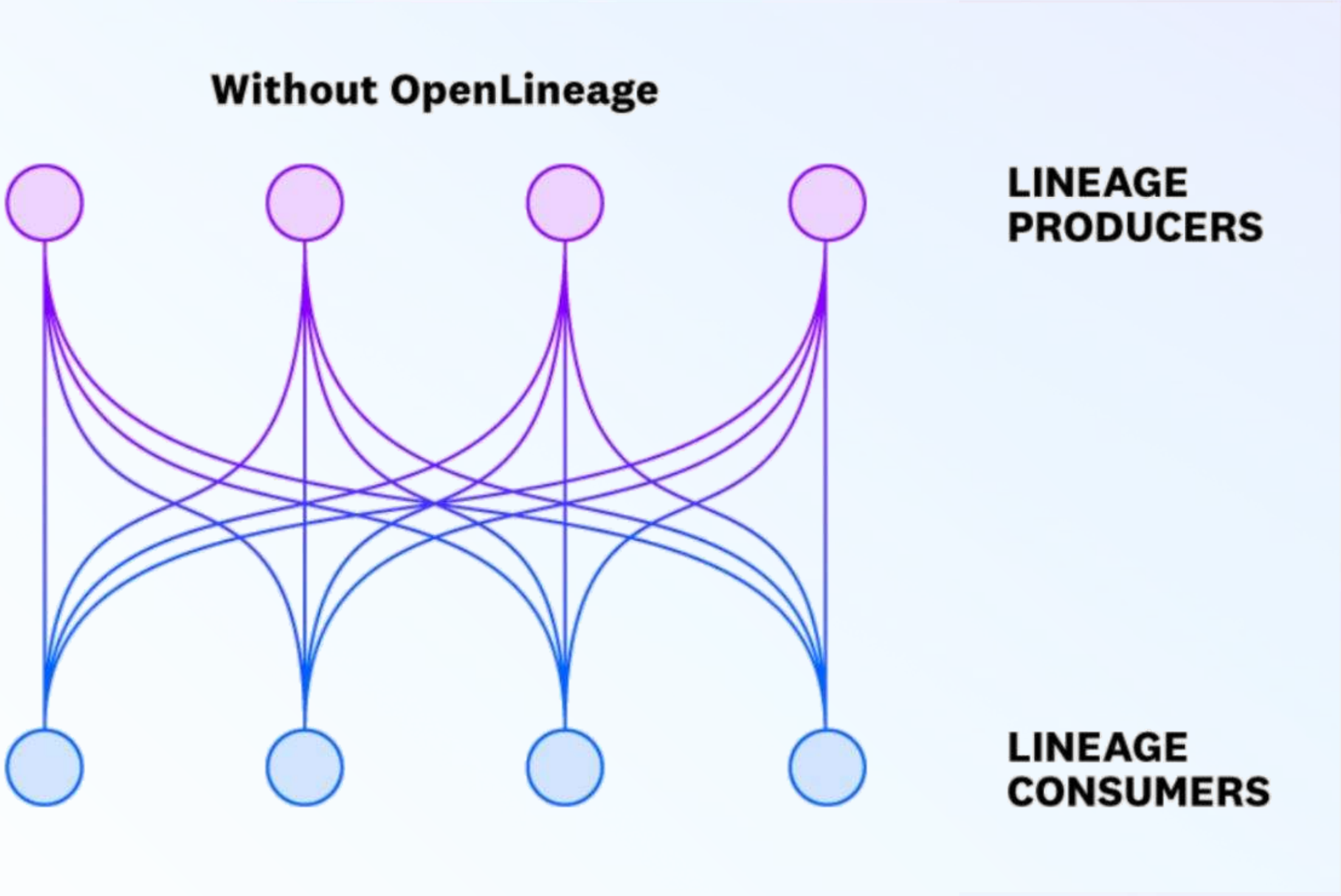
Latitude	N40°52'51.5"
Longitude	W73°19'42.4"
Elevation	49m
UTC	12/13/2012 15:25:09
Satellite reception	3D

…or you can capture it when originally created.

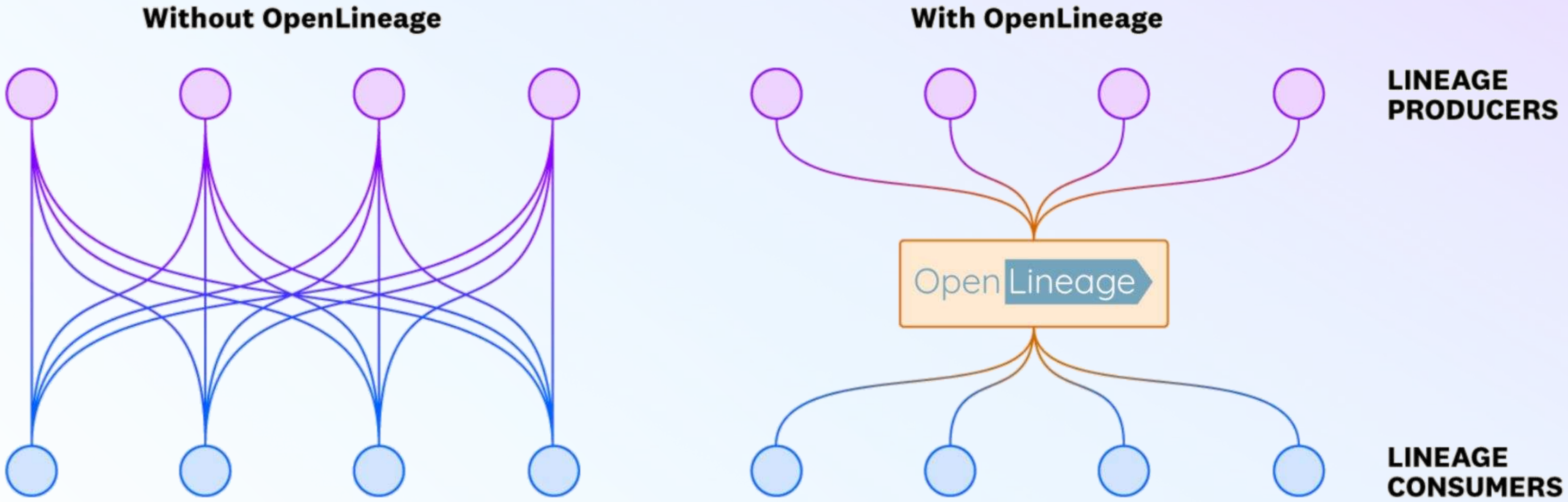
What is OpenLineage?

- Spec: open, vendor neutral
- Libraries: for common languages
- Integrations with data tools

Open Standards Matter



Simplified


Without OpenLineage
With OpenLineage
OpenLineage
LINEAGE PRODUCERS
LINEAGE CONSUMERS

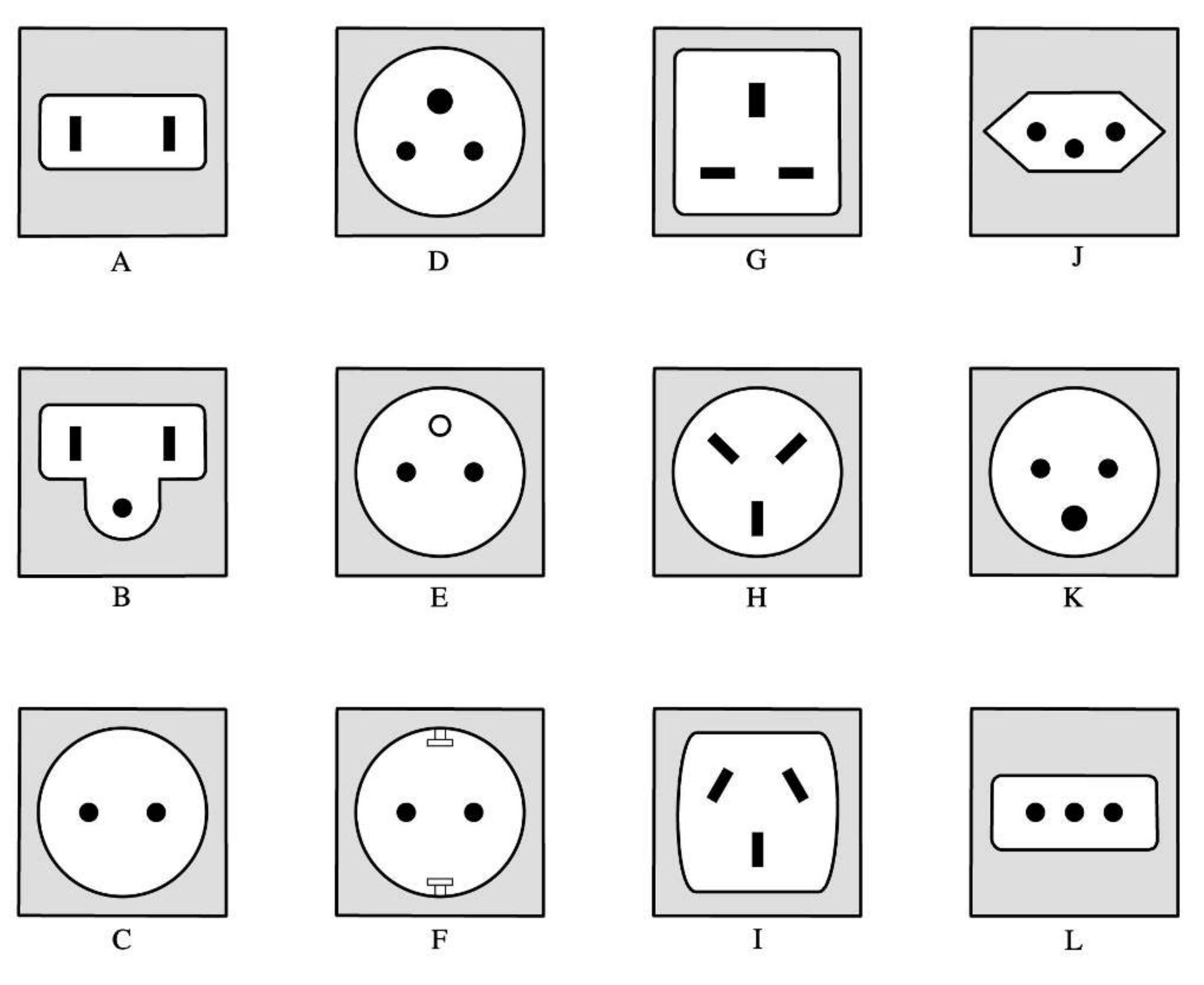
OpenLineage at Datadog

Datadog Data Observability

Data Jobs Monitoring

Data Quality Monitoring

OpenLineage benefits

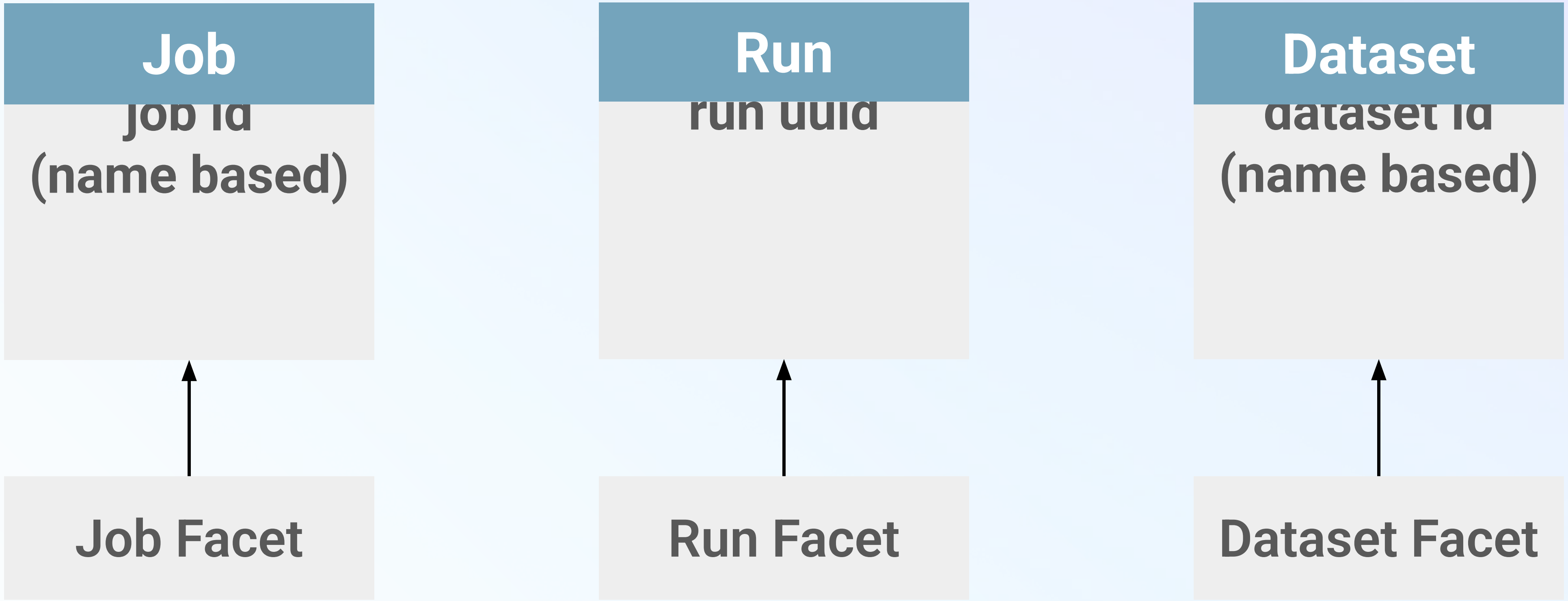


OpenLineage Spec deep dive

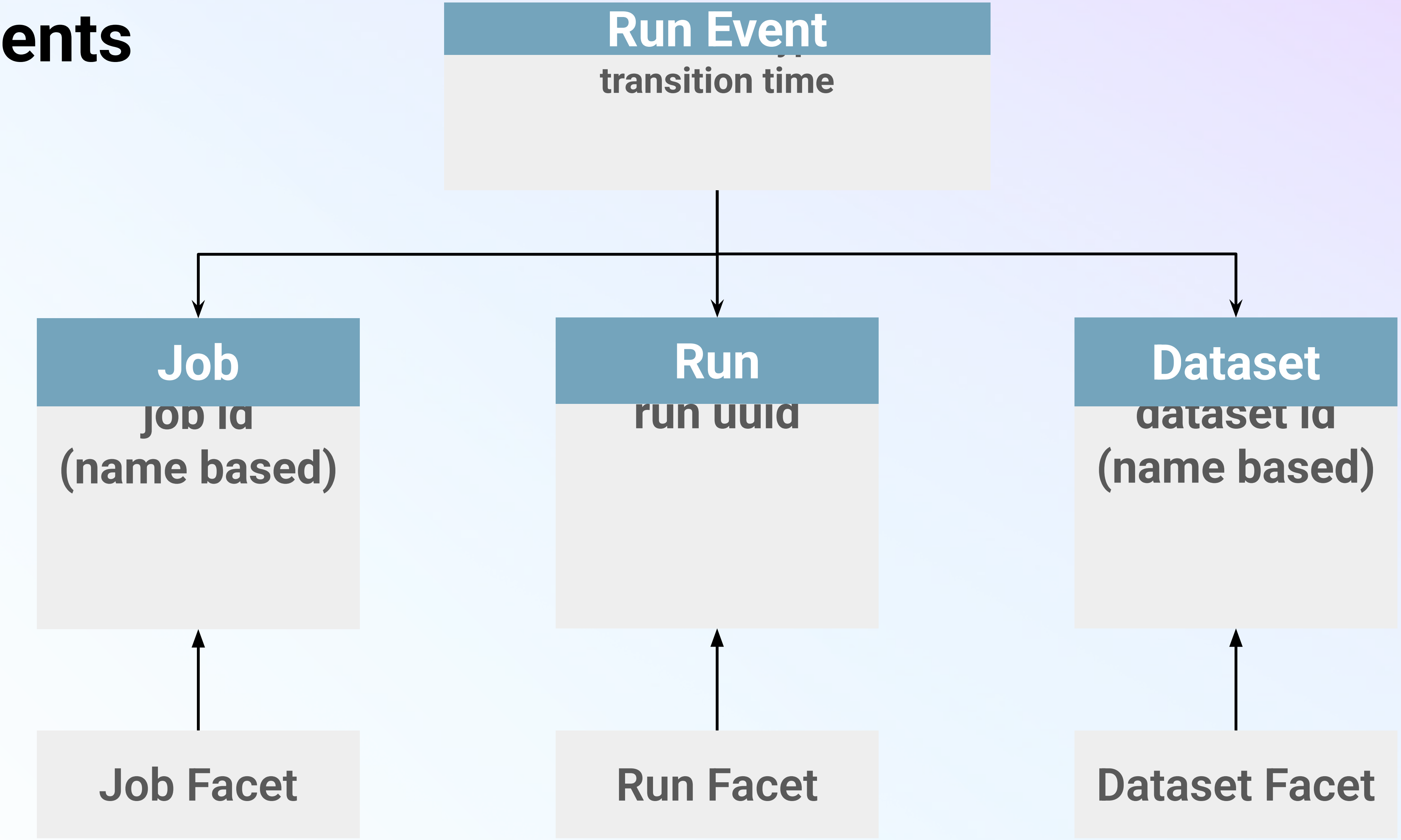
Core Concepts: Job, Run, Dataset

Job	Run	Dataset
job id (name based)	run uuid	dataset id (name based)

Core Concepts: Facets



Events



Run Event

transition time

Job

job id
(name based)

Run

run uuid

Dataset

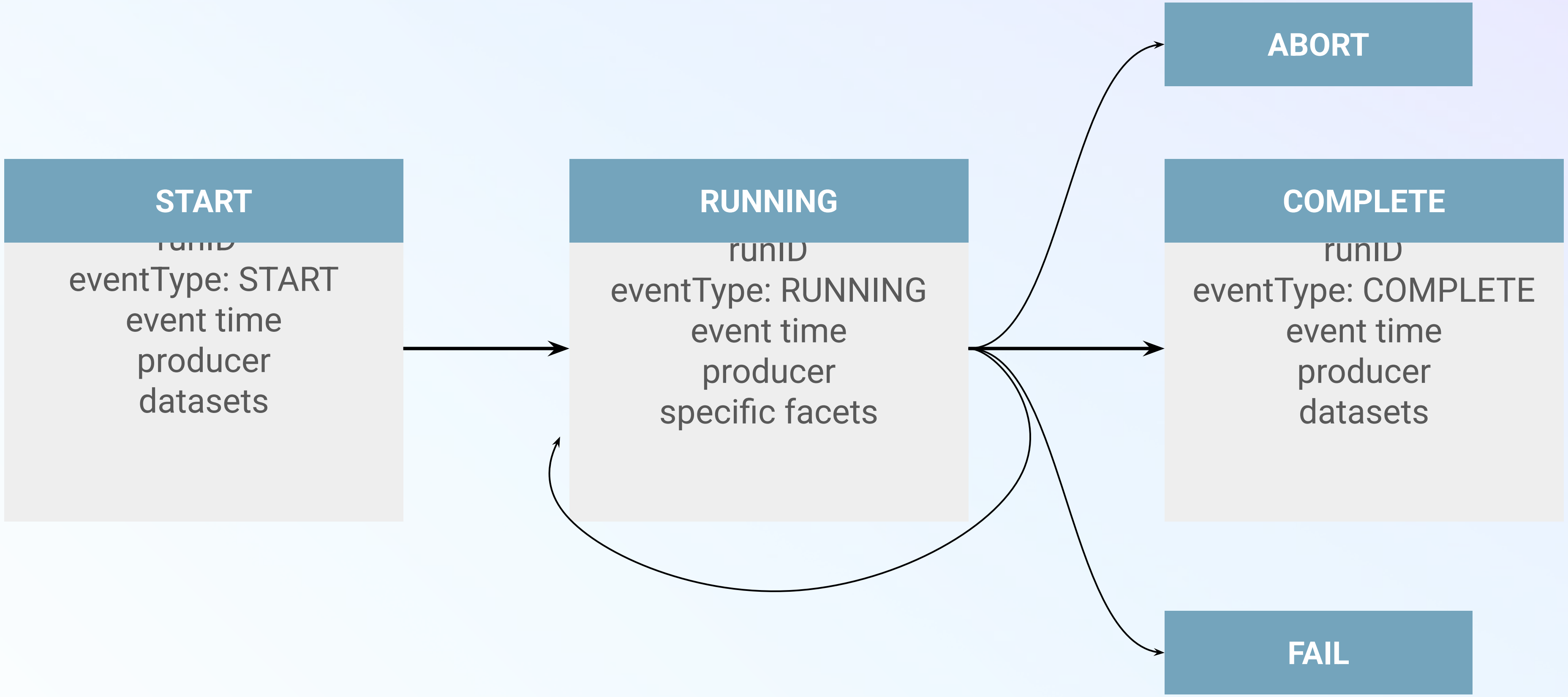
dataset id
(name based)

Job Facet

Run Facet

Dataset Facet

Run Event Lifecycle



START

runID
eventType: START
event time
producer
datasets

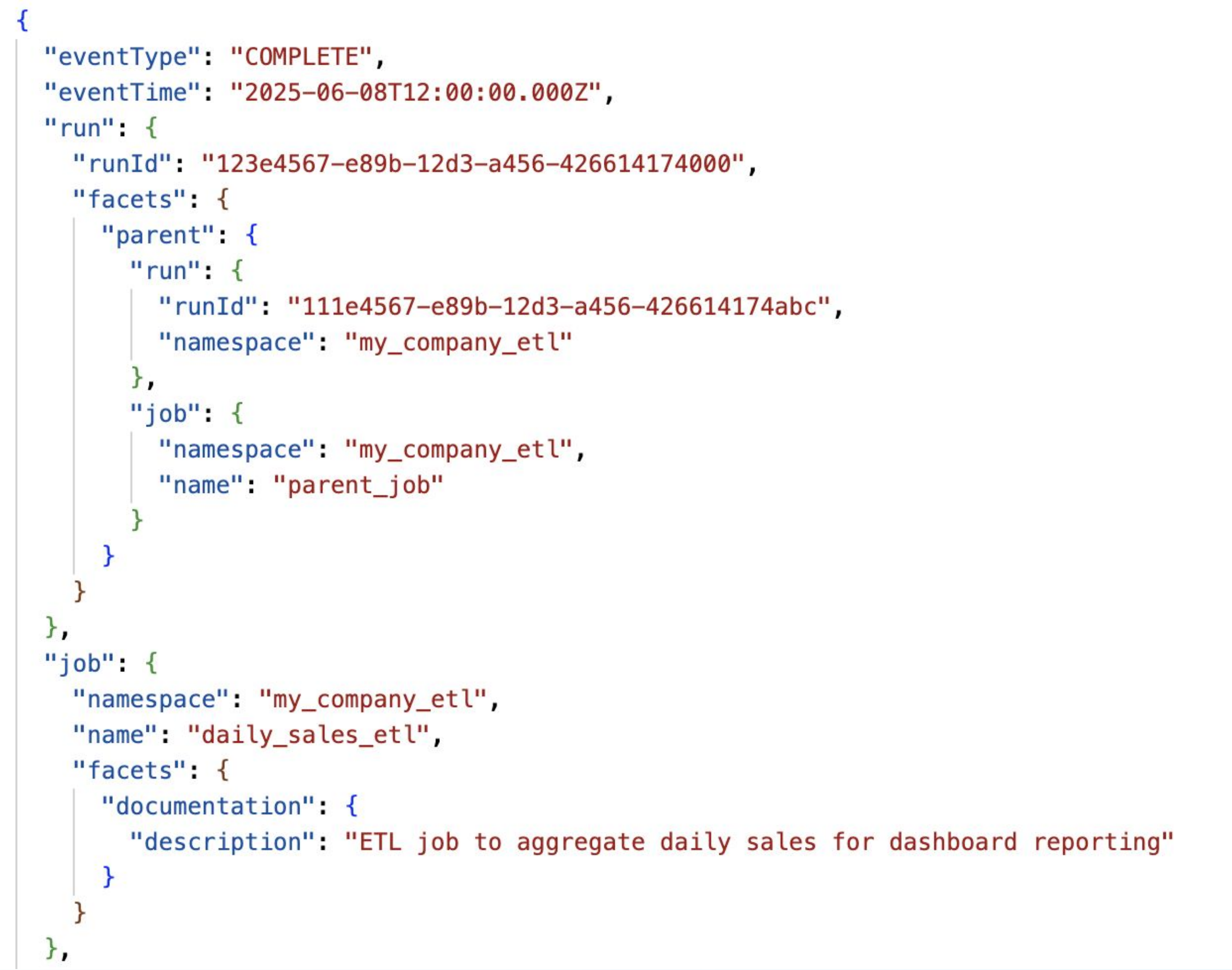
RUNNING

runID
eventType: RUNNING
event time
producer
specific facets

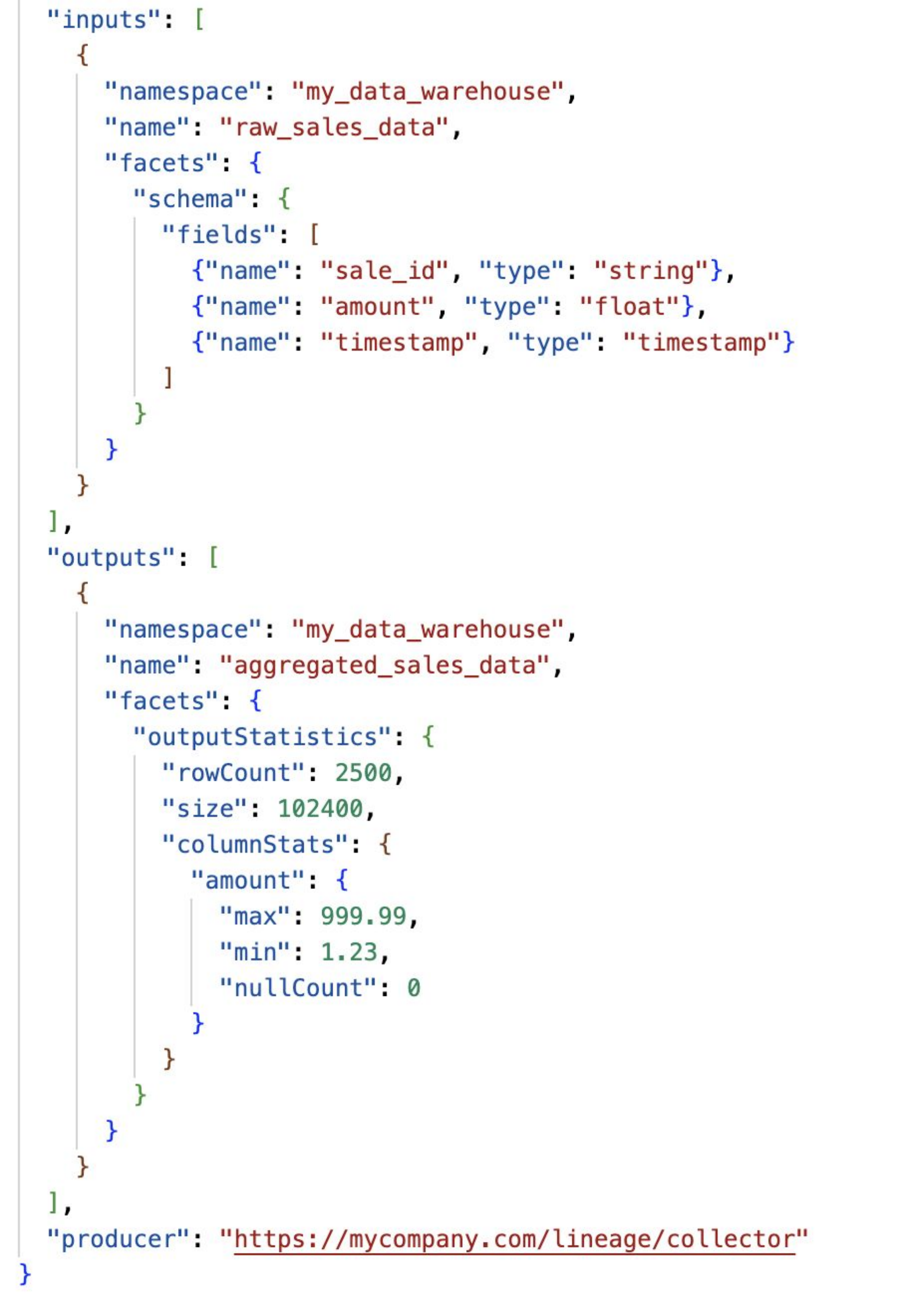
COMPLETE

runID
eventType: COMPLETE
event time
producer
datasets

Example of OpenLineage event

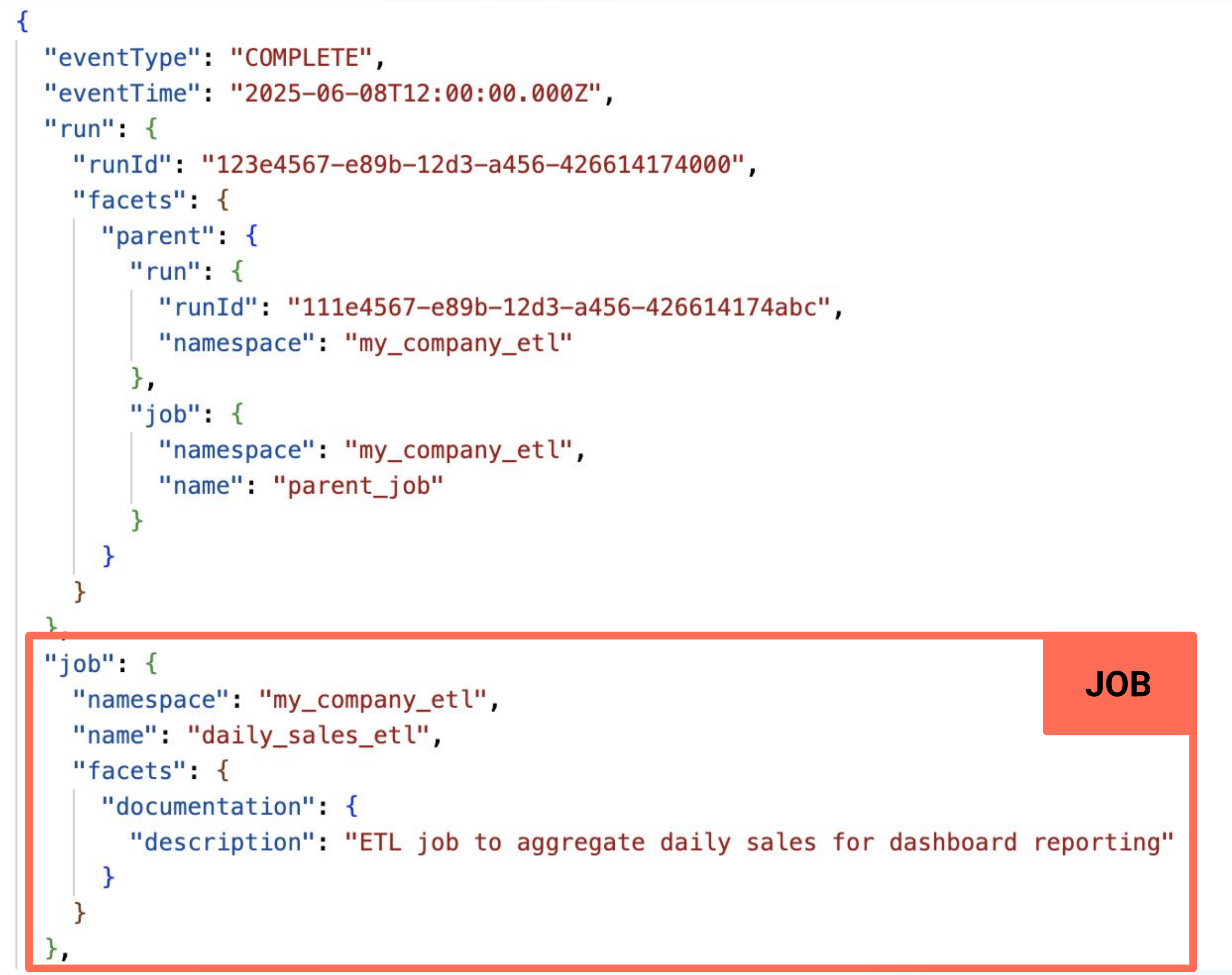


```
{
  "eventType": "COMPLETE",
  "eventTime": "2025-06-08T12:00:00.000Z",
  "run": {
    "runId": "123e4567-e89b-12d3-a456-426614174000",
    "facets": {
      "parent": {
        "run": {
          "runId": "111e4567-e89b-12d3-a456-426614174abc",
          "namespace": "my_company_etl"
        },
        "job": {
          "namespace": "my_company_etl",
          "name": "parent_job"
        }
      }
    }
  },
  "job": {
    "namespace": "my_company_etl",
    "name": "daily_sales_etl",
    "facets": {
      "documentation": {
        "description": "ETL job to aggregate daily sales for dashboard reporting"
      }
    }
  },
```



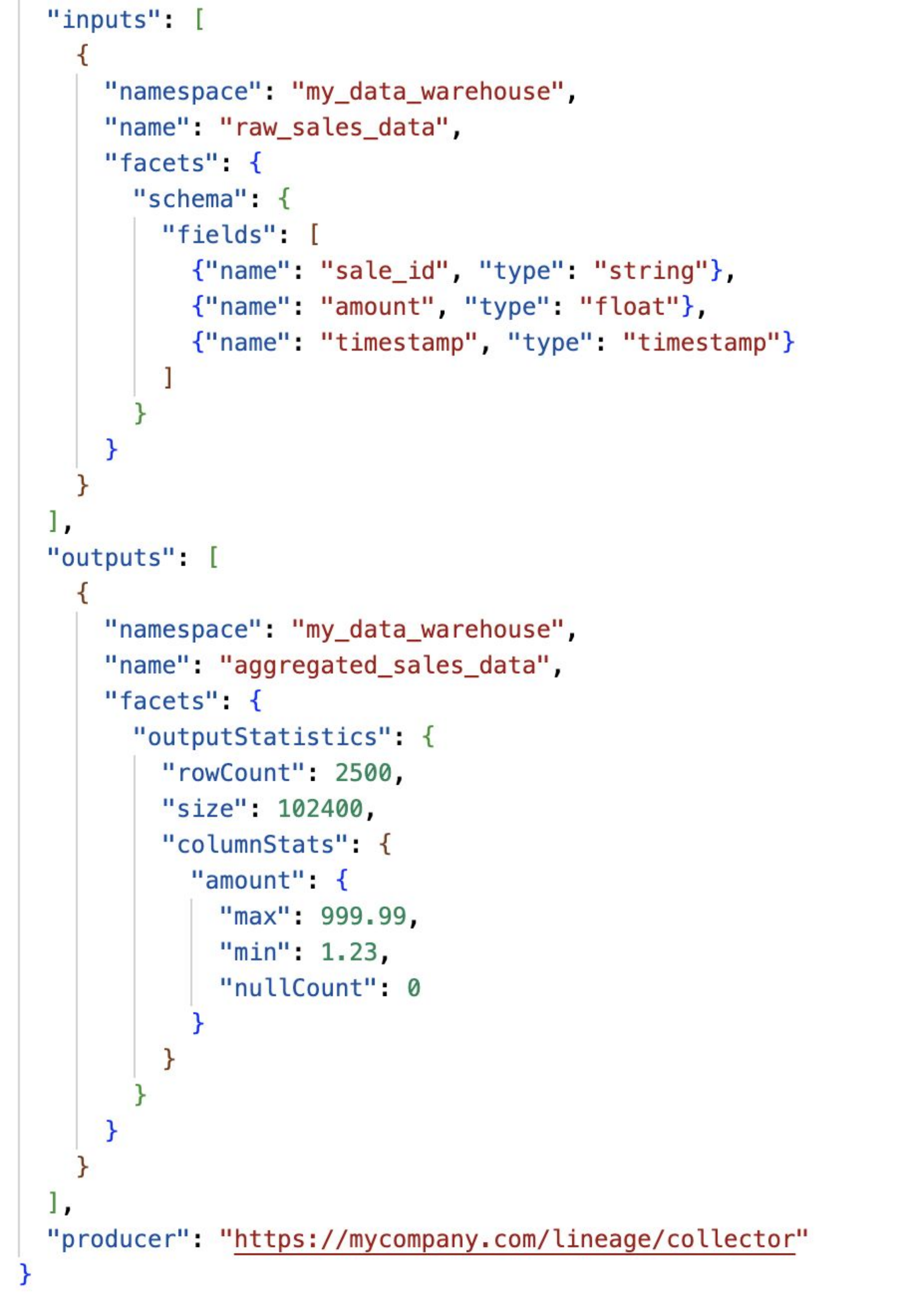
```
  "inputs": [
    {
      "namespace": "my_data_warehouse",
      "name": "raw_sales_data",
      "facets": {
        "schema": {
          "fields": [
            {"name": "sale_id", "type": "string"},
            {"name": "amount", "type": "float"},
            {"name": "timestamp", "type": "timestamp"}
          ]
        }
      }
    }
  ],
  "outputs": [
    {
      "namespace": "my_data_warehouse",
      "name": "aggregated_sales_data",
      "facets": {
        "outputStatistics": {
          "rowCount": 2500,
          "size": 102400,
          "columnStats": {
            "amount": {
              "max": 999.99,
              "min": 1.23,
              "nullCount": 0
            }
          }
        }
      }
    }
  ],
  "producer": "https://mycompany.com/lineage/collector"
}
```

Example of OpenLineage event

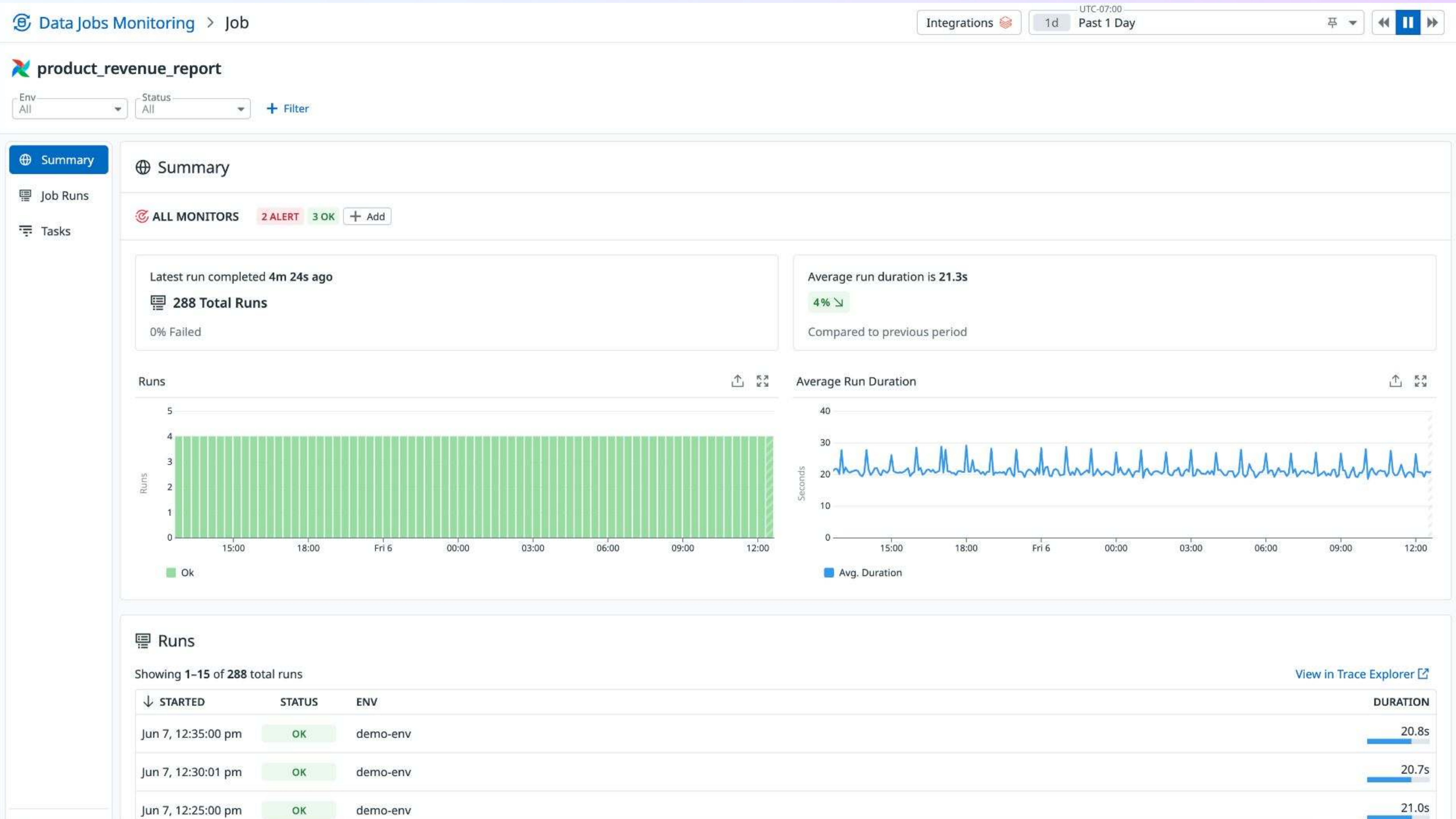


```
{
  "eventType": "COMPLETE",
  "eventTime": "2025-06-08T12:00:00.000Z",
  "run": {
    "runId": "123e4567-e89b-12d3-a456-426614174000",
    "facets": {
      "parent": {
        "run": {
          "runId": "111e4567-e89b-12d3-a456-426614174abc",
          "namespace": "my_company_etl"
        },
        "job": {
          "namespace": "my_company_etl",
          "name": "parent_job"
        }
      }
    }
  }
  "job": {
    "namespace": "my_company_etl",
    "name": "daily_sales_etl",
    "facets": {
      "documentation": {
        "description": "ETL job to aggregate daily sales for dashboard reporting"
      }
    }
  },
```

JOB



```
  "inputs": [
    {
      "namespace": "my_data_warehouse",
      "name": "raw_sales_data",
      "facets": {
        "schema": {
          "fields": [
            {"name": "sale_id", "type": "string"},
            {"name": "amount", "type": "float"},
            {"name": "timestamp", "type": "timestamp"}
          ]
        }
      }
    }
  ],
  "outputs": [
    {
      "namespace": "my_data_warehouse",
      "name": "aggregated_sales_data",
      "facets": {
        "outputStatistics": {
          "rowCount": 2500,
          "size": 102400,
          "columnStats": {
            "amount": {
              "max": 999.99,
              "min": 1.23,
              "nullCount": 0
            }
          }
        }
      }
    }
  ],
  "producer": "https://mycompany.com/lineage/collector"
}
```



product_revenue_report

Env: All

Status: All

Summary

ALL MONITORS 2 ALERT 3 OK + Add

Latest run completed **4m 24s ago**

288 Total Runs

0% Failed

Average run duration is **21.3s**

4% ↘

Compared to previous period

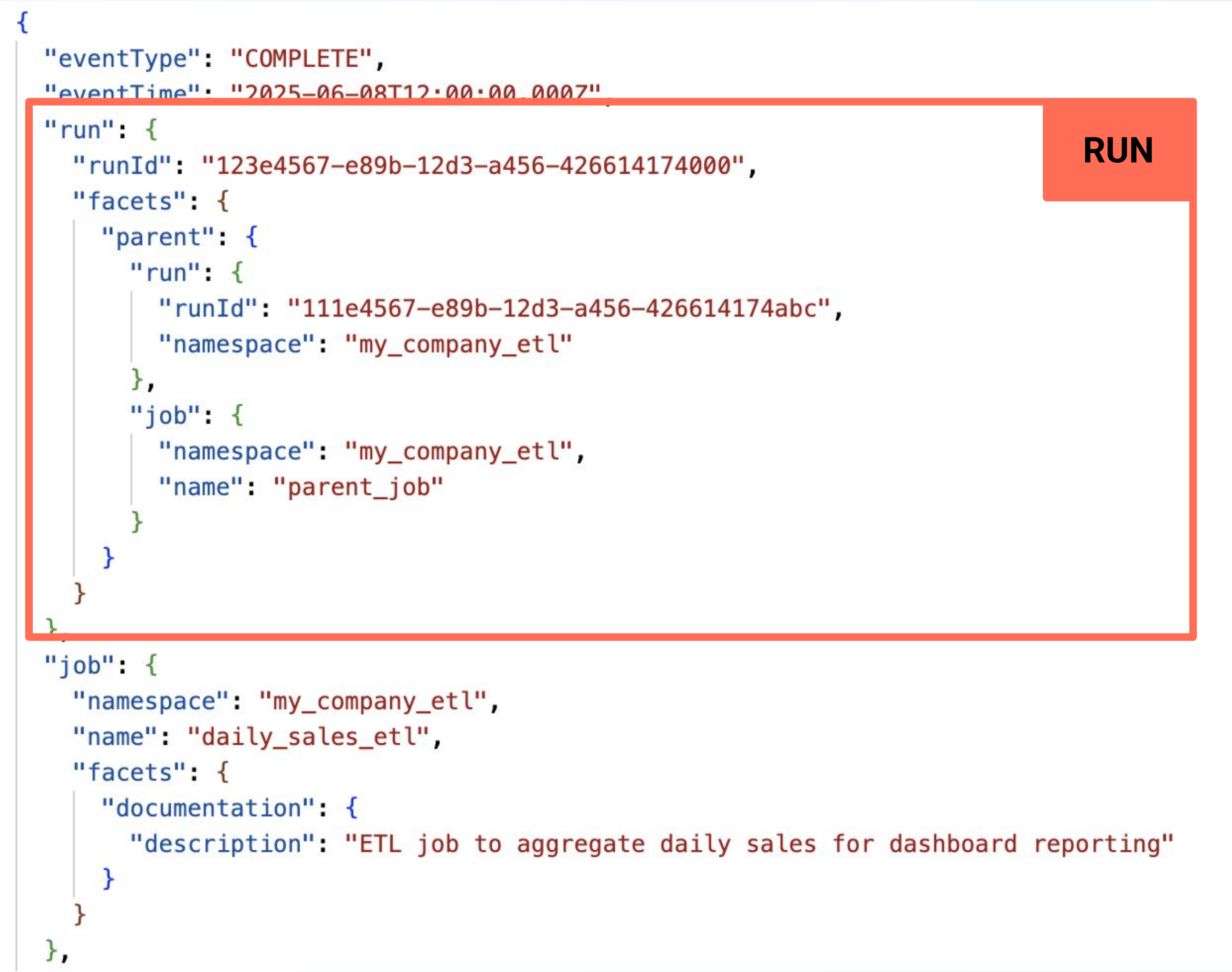
Runs

Showing **1–15** of **288** total runs

View in Trace Explorer

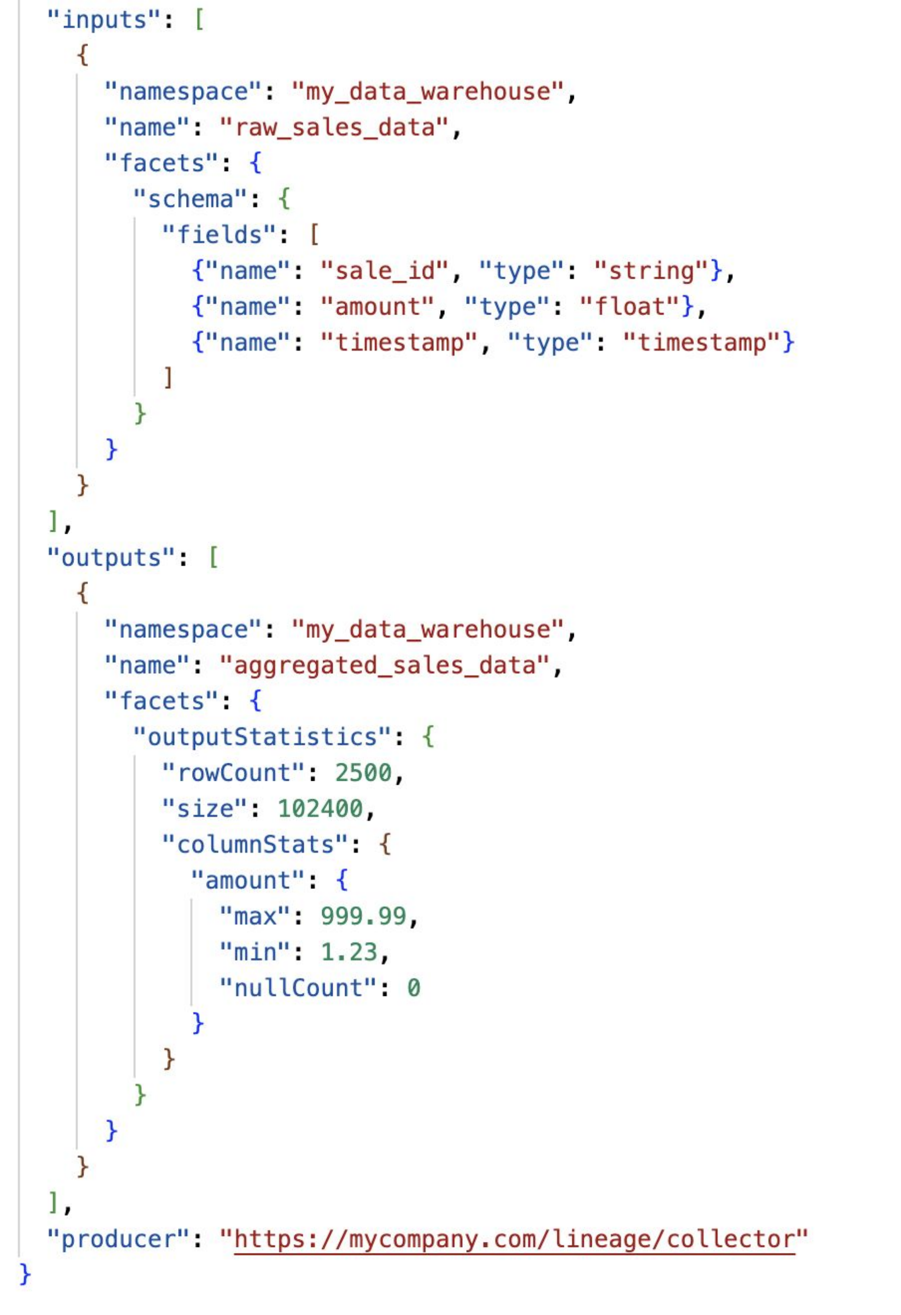
↓ STARTED	STATUS	ENV	DURATION
Jun 7, 12:35:00 pm	OK	demo-env	20.8s
Jun 7, 12:30:01 pm	OK	demo-env	20.7s
Jun 7, 12:25:00 pm	OK	demo-env	21.0s
]

Example of OpenLineage event

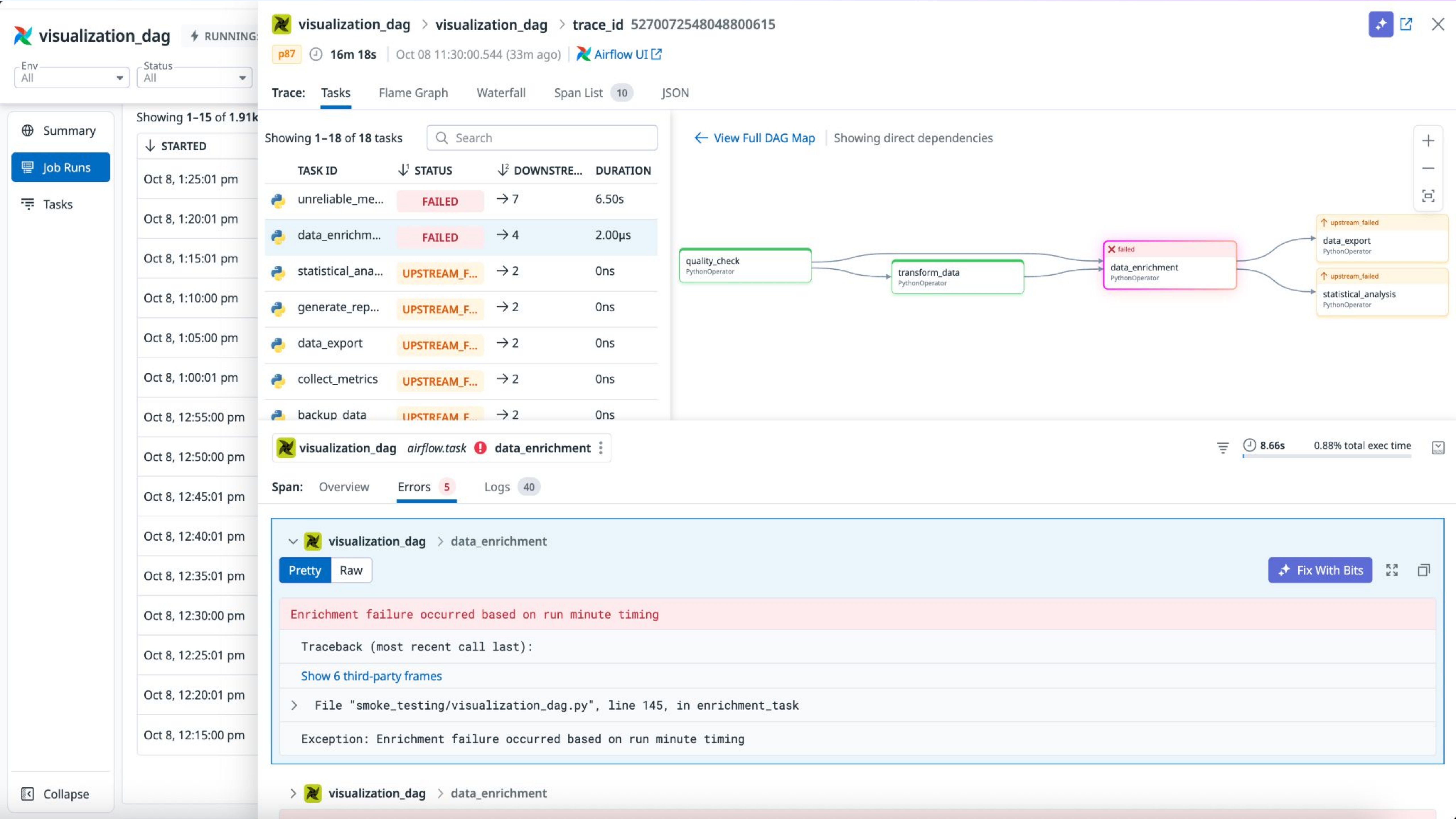


RUN

```json
{
  "eventType": "COMPLETE",
  "eventTime": "2025-06-08T12:00:00.000Z",
  "run": {
    "runId": "123e4567-e89b-12d3-a456-426614174000",
    "facets": {
      "parent": {
        "run": {
          "runId": "111e4567-e89b-12d3-a456-426614174abc",
          "namespace": "my_company_etl"
        },
        "job": {
          "namespace": "my_company_etl",
          "name": "parent_job"
        }
      }
    }
  },
  "job": {
    "namespace": "my_company_etl",
    "name": "daily_sales_etl",
    "facets": {
      "documentation": {
        "description": "ETL job to aggregate daily sales for dashboard reporting"
      }
    }
  },
```



```json
  "inputs": [
    {
      "namespace": "my_data_warehouse",
      "name": "raw_sales_data",
      "facets": {
        "schema": {
          "fields": [
            {"name": "sale_id", "type": "string"},
            {"name": "amount", "type": "float"},
            {"name": "timestamp", "type": "timestamp"}
          ]
        }
      }
    }
  ],
  "outputs": [
    {
      "namespace": "my_data_warehouse",
      "name": "aggregated_sales_data",
      "facets": {
        "outputStatistics": {
          "rowCount": 2500,
          "size": 102400,
          "columnStats": {
            "amount": {
              "max": 999.99,
              "min": 1.23,
              "nullCount": 0
            }
          }
        }
      }
    }
  ],
  "producer": "https://mycompany.com/lineage/collector"
}
```



Showing **1–15** of **1.91k**

↓ STARTED
Oct 8, 1:25:01 pm
Oct 8, 1:20:01 pm
Oct 8, 1:15:01 pm
Oct 8, 1:10:00 pm
Oct 8, 1:05:00 pm
Oct 8, 1:00:01 pm
Oct 8, 12:55:00 pm
Oct 8, 12:50:00 pm
Oct 8, 12:45:01 pm
Oct 8, 12:40:01 pm
Oct 8, 12:35:01 pm
Oct 8, 12:30:00 pm
Oct 8, 12:25:01 pm
Oct 8, 12:20:01 pm
Oct 8, 12:15:00 pm

visualization_dag > visualization_dag > trace_id 5270072548048800615

p87 16m 18s | Oct 08 11:30:00.544 (33m ago) | Airflow UI

Showing **1–18** of **18** tasks

TASK ID	STATUS	DOWNSTRE...	DURATION
unreliable_me...	FAILED	→ 7	6.50s
data_enrichm...	FAILED	→ 4	2.00µs
statistical_ana...	UPSTREAM_F...	→ 2	0ns
generate_rep...	UPSTREAM_F...	→ 2	0ns
data_export	UPSTREAM_F...	→ 2	0ns
collect_metrics	UPSTREAM_F...	→ 2	0ns
backup_data	UPSTREAM_F...	→ 2	0ns

← View Full DAG Map | Showing direct dependencies

visualization_dag *airflow.task* data_enrichment

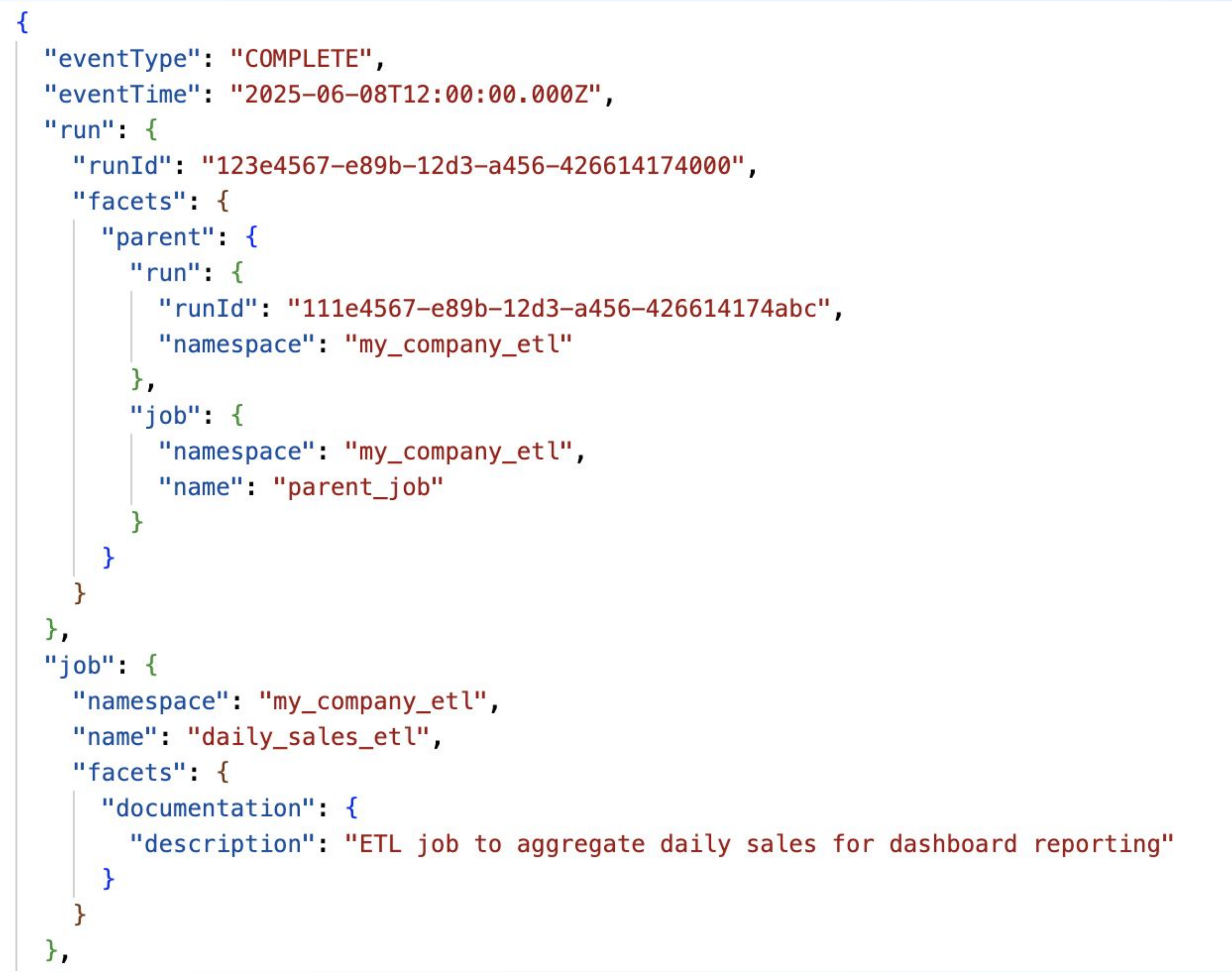
8.66s 0.88% total exec time

visualization_dag > data_enrichment

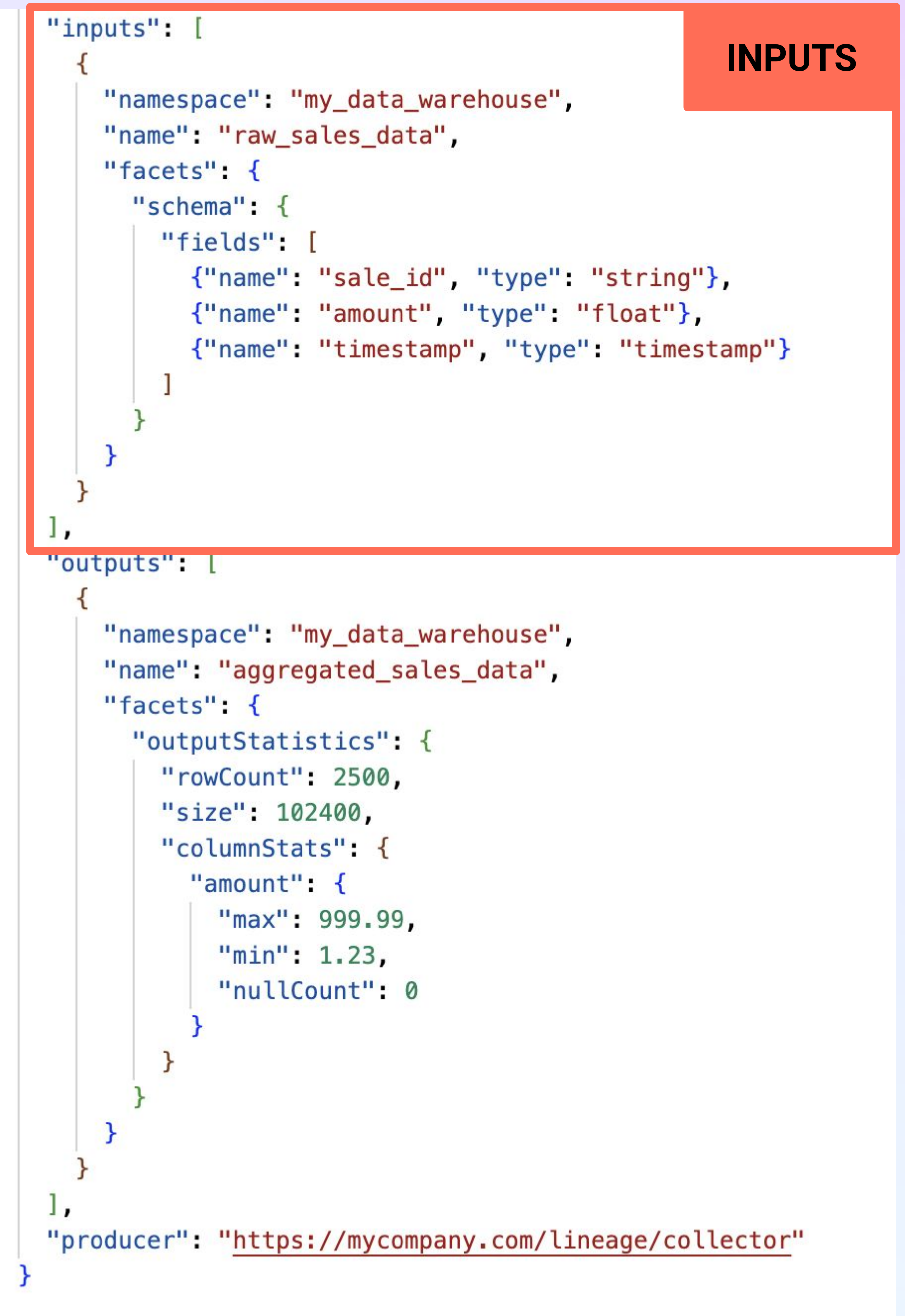
```
Enrichment failure occurred based on run minute timing
Traceback (most recent call last):
Show 6 third-party frames
> File "smoke_testing/visualization_dag.py", line 145, in enrichment_task
Exception: Enrichment failure occurred based on run minute timing
```

visualization_dag > data_enrichment

Example of OpenLineage event



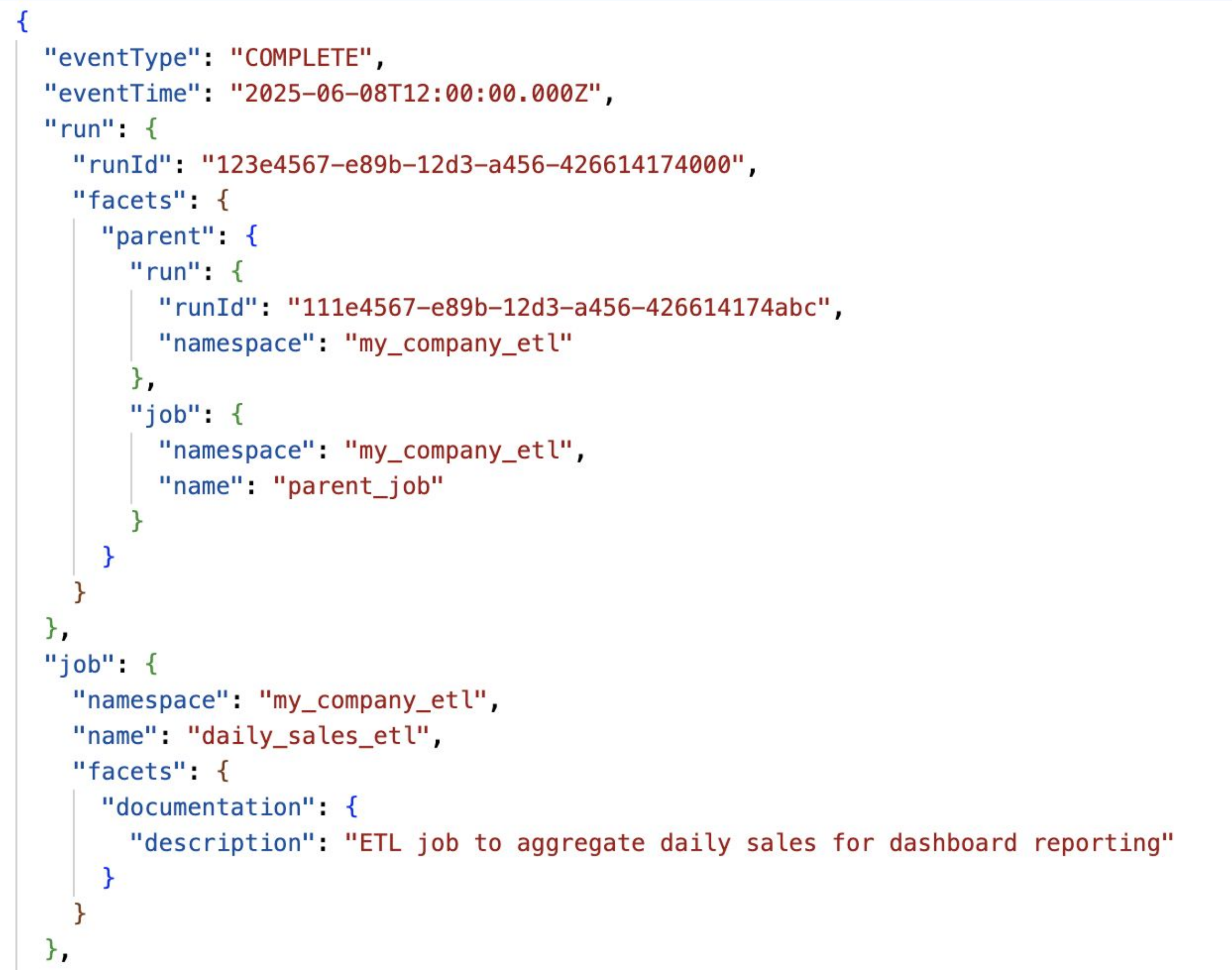
```
{
  "eventType": "COMPLETE",
  "eventTime": "2025-06-08T12:00:00.000Z",
  "run": {
    "runId": "123e4567-e89b-12d3-a456-426614174000",
    "facets": {
      "parent": {
        "run": {
          "runId": "111e4567-e89b-12d3-a456-426614174abc",
          "namespace": "my_company_etl"
        },
        "job": {
          "namespace": "my_company_etl",
          "name": "parent_job"
        }
      }
    }
  },
  "job": {
    "namespace": "my_company_etl",
    "name": "daily_sales_etl",
    "facets": {
      "documentation": {
        "description": "ETL job to aggregate daily sales for dashboard reporting"
      }
    }
  },
```



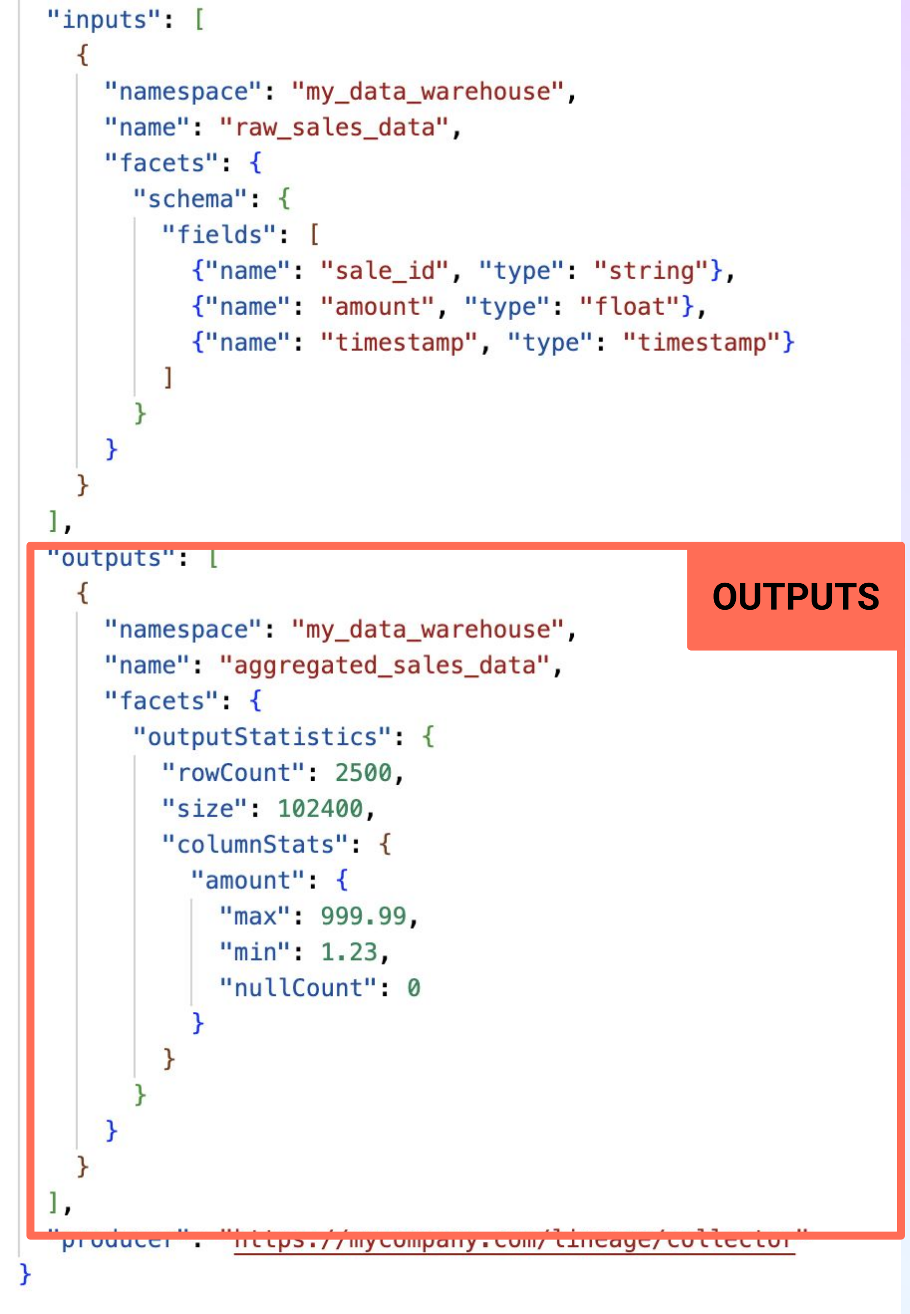
INPUTS

```
  "inputs": [
    {
      "namespace": "my_data_warehouse",
      "name": "raw_sales_data",
      "facets": {
        "schema": {
          "fields": [
            {"name": "sale_id", "type": "string"},
            {"name": "amount", "type": "float"},
            {"name": "timestamp", "type": "timestamp"}
          ]
        }
      }
    }
  ],
  "outputs": [
    {
      "namespace": "my_data_warehouse",
      "name": "aggregated_sales_data",
      "facets": {
        "outputStatistics": {
          "rowCount": 2500,
          "size": 102400,
          "columnStats": {
            "amount": {
              "max": 999.99,
              "min": 1.23,
              "nullCount": 0
            }
          }
        }
      }
    }
  ],
  "producer": "https://mycompany.com/lineage/collector"
}
```

Example of OpenLineage event



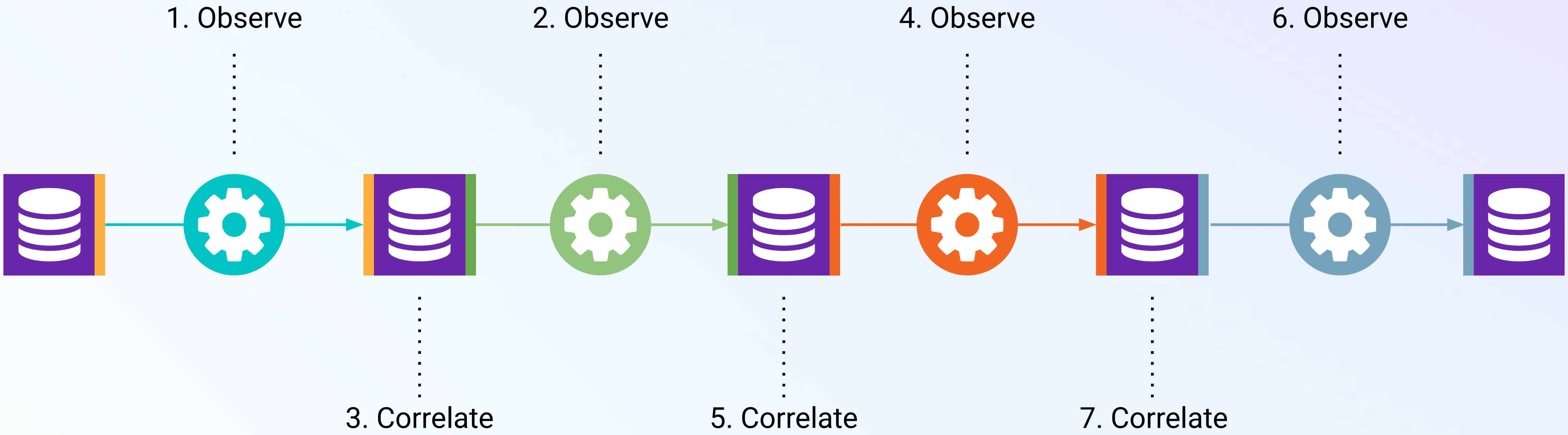
```
{
  "eventType": "COMPLETE",
  "eventTime": "2025-06-08T12:00:00.000Z",
  "run": {
    "runId": "123e4567-e89b-12d3-a456-426614174000",
    "facets": {
      "parent": {
        "run": {
          "runId": "111e4567-e89b-12d3-a456-426614174abc",
          "namespace": "my_company_etl"
        },
        "job": {
          "namespace": "my_company_etl",
          "name": "parent_job"
        }
      }
    }
  },
  "job": {
    "namespace": "my_company_etl",
    "name": "daily_sales_etl",
    "facets": {
      "documentation": {
        "description": "ETL job to aggregate daily sales for dashboard reporting"
      }
    }
  },
```



```
  "inputs": [
    {
      "namespace": "my_data_warehouse",
      "name": "raw_sales_data",
      "facets": {
        "schema": {
          "fields": [
            {"name": "sale_id", "type": "string"},
            {"name": "amount", "type": "float"},
            {"name": "timestamp", "type": "timestamp"}
          ]
        }
      }
    }
  ],
  "outputs": [
    {
      "namespace": "my_data_warehouse",
      "name": "aggregated_sales_data",
      "facets": {
        "outputStatistics": {
          "rowCount": 2500,
          "size": 102400,
          "columnStats": {
            "amount": {
              "max": 999.99,
              "min": 1.23,
              "nullCount": 0
            }
          }
        }
      }
    }
  ],
  "producer": "https://mycompany.com/lineage/collector"
}
```

OUTPUTS

Lineage is built on correlations



1. Observe

2. Observe

3. Correlate

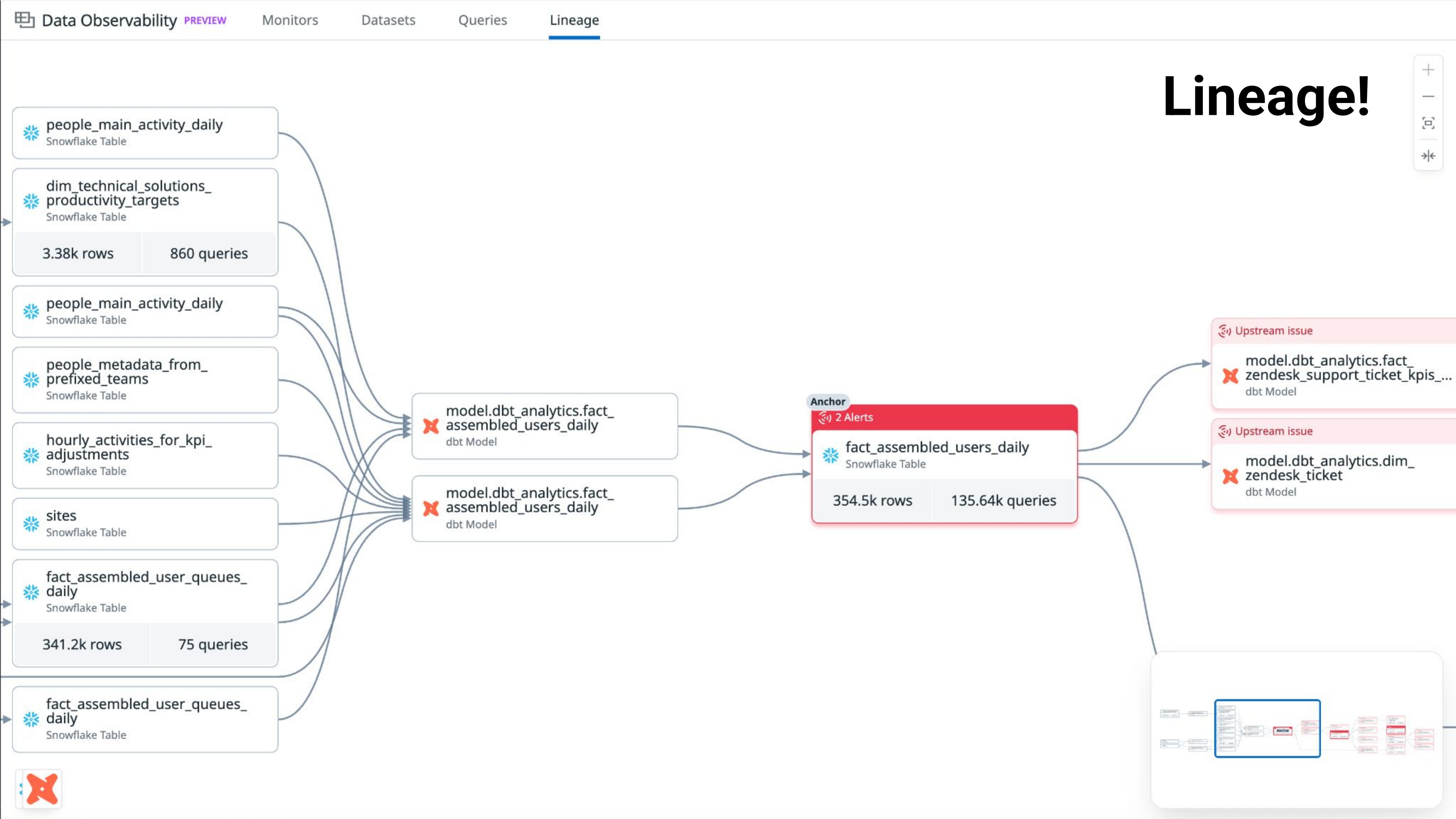
4. Observe

5. Correlate

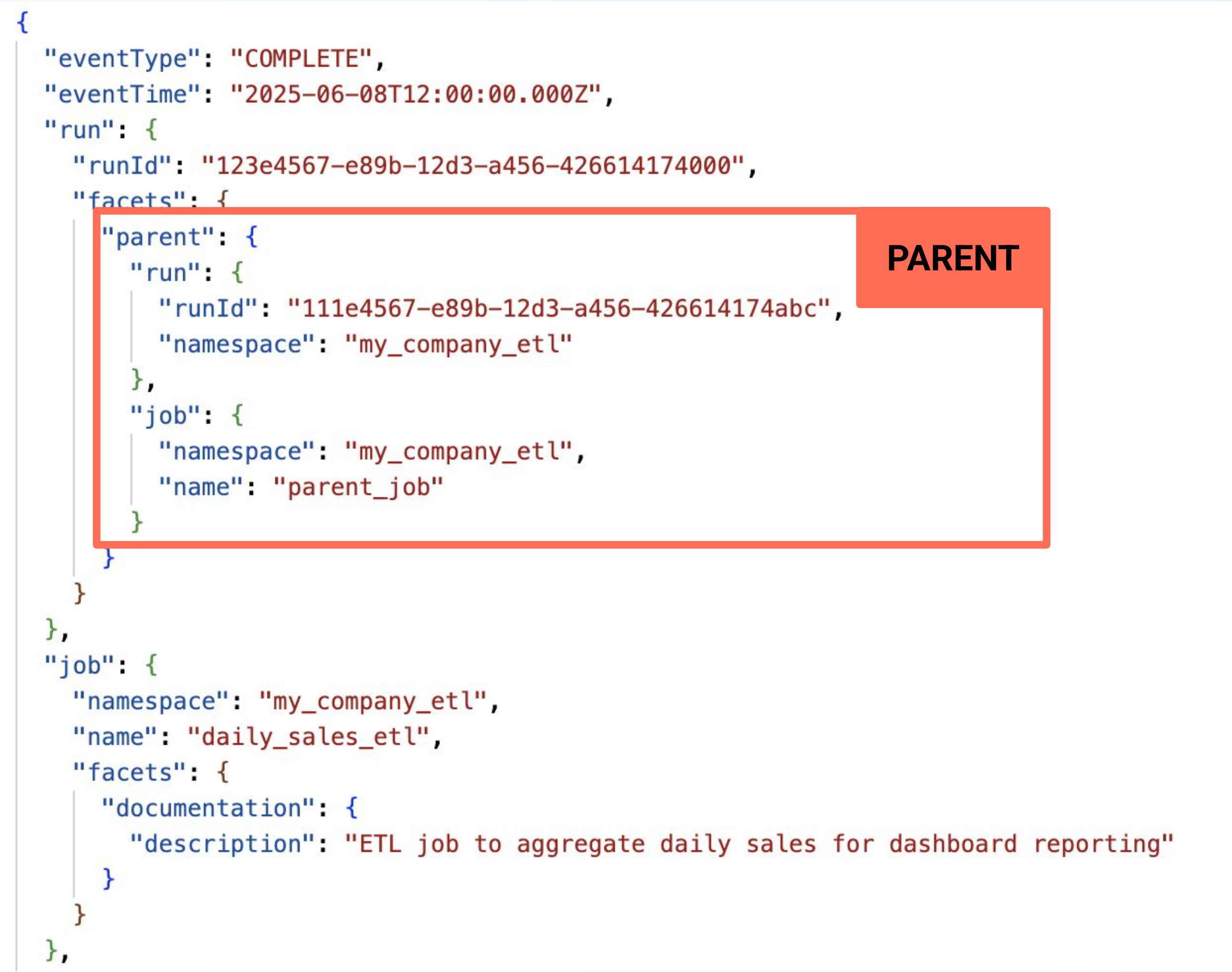
6. Observe

7. Correlate

Lineage!

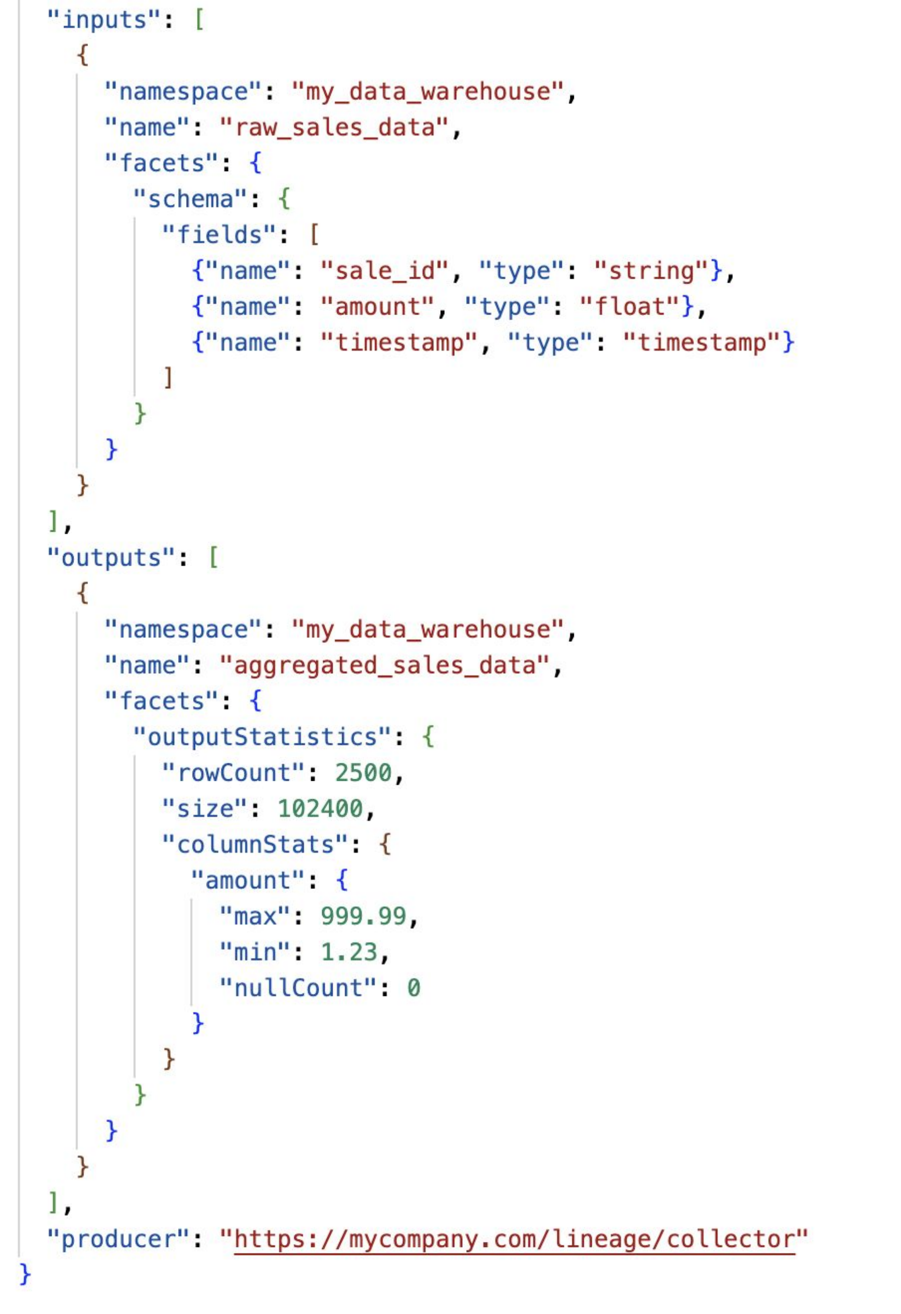

Data Observability PREVIEW
Monitors
Datasets
Queries
Lineage
people_main_activity_daily
Snowflake Table
dim_technical_solutions_productivity_targets
Snowflake Table
3.38k rows
860 queries
people_main_activity_daily
Snowflake Table
people_metadata_from_prefixed_teams
Snowflake Table
hourly_activities_for_kpi_adjustments
Snowflake Table
sites
Snowflake Table
fact_assembled_user_queues_daily
Snowflake Table
341.2k rows
75 queries
fact_assembled_user_queues_daily
Snowflake Table
model.dbt_analytics.fact_assembled_users_daily
dbt Model
model.dbt_analytics.fact_assembled_users_daily
dbt Model
Anchor
2 Alerts
fact_assembled_users_daily
Snowflake Table
354.5k rows
135.64k queries
Upstream issue
model.dbt_analytics.fact_zendesk_support_ticket_kpis_...
dbt Model
Upstream issue
model.dbt_analytics.dim_zendesk_ticket
dbt Model
Anchor

Example of OpenLineage event



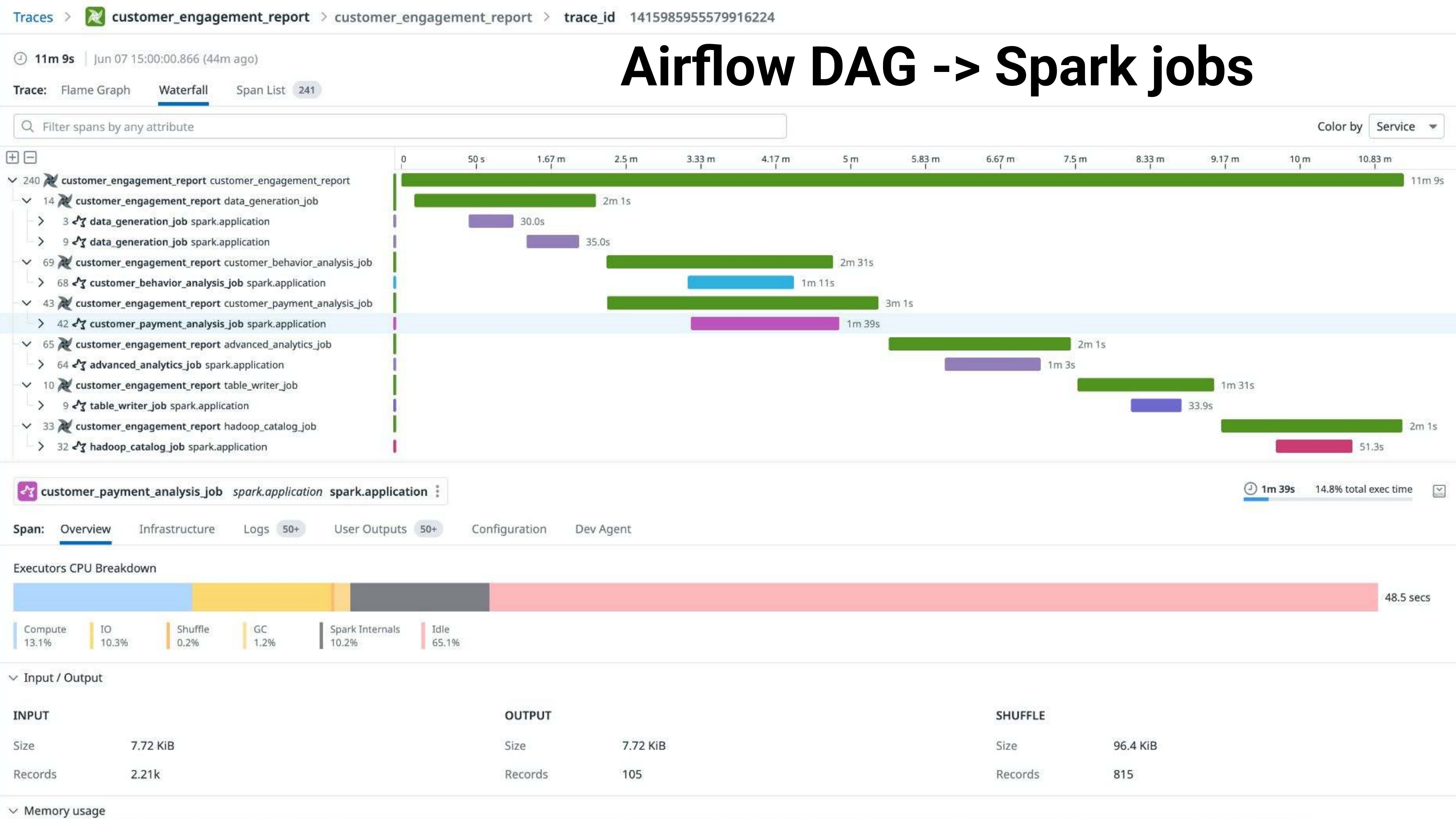
```
{
  "eventType": "COMPLETE",
  "eventTime": "2025-06-08T12:00:00.000Z",
  "run": {
    "runId": "123e4567-e89b-12d3-a456-426614174000",
    "facets": {
      "parent": {
        "run": {
          "runId": "111e4567-e89b-12d3-a456-426614174abc",
          "namespace": "my_company_etl"
        },
        "job": {
          "namespace": "my_company_etl",
          "name": "parent_job"
        }
      }
    }
  },
  "job": {
    "namespace": "my_company_etl",
    "name": "daily_sales_etl",
    "facets": {
      "documentation": {
        "description": "ETL job to aggregate daily sales for dashboard reporting"
      }
    }
  },
```

PARENT



```
  "inputs": [
    {
      "namespace": "my_data_warehouse",
      "name": "raw_sales_data",
      "facets": {
        "schema": {
          "fields": [
            {"name": "sale_id", "type": "string"},
            {"name": "amount", "type": "float"},
            {"name": "timestamp", "type": "timestamp"}
          ]
        }
      }
    }
  ],
  "outputs": [
    {
      "namespace": "my_data_warehouse",
      "name": "aggregated_sales_data",
      "facets": {
        "outputStatistics": {
          "rowCount": 2500,
          "size": 102400,
          "columnStats": {
            "amount": {
              "max": 999.99,
              "min": 1.23,
              "nullCount": 0
            }
          }
        }
      }
    }
  ],
  "producer": "https://mycompany.com/lineage/collector"
}
```

Airflow DAG -> Spark jobs



Traces > customer_engagement_report > customer_engagement_report > trace_id 1415985955579916224

11m 9s | Jun 07 15:00:00.866 (44m ago)

Trace: Flame Graph | Waterfall | Span List 241

Filter spans by any attribute

Color by Service

```
240 customer_engagement_report customer_engagement_report          11m 9s
  14 customer_engagement_report data_generation_job                2m 1s
    3 data_generation_job spark.application                        30.0s
    9 data_generation_job spark.application                        35.0s
  69 customer_engagement_report customer_behavior_analysis_job     2m 31s
    68 customer_behavior_analysis_job spark.application            1m 11s
  43 customer_engagement_report customer_payment_analysis_job      3m 1s
    42 customer_payment_analysis_job spark.application             1m 39s
  65 customer_engagement_report advanced_analytics_job             2m 1s
    64 advanced_analytics_job spark.application                    1m 3s
  10 customer_engagement_report table_writer_job                   1m 31s
    9 table_writer_job spark.application                           33.9s
  33 customer_engagement_report hadoop_catalog_job                 2m 1s
    32 hadoop_catalog_job spark.application                        51.3s
```

customer_payment_analysis_job *spark.application* **spark.application**

1m 39s 14.8% total exec time

Span: Overview | Infrastructure | Logs 50+ | User Outputs 50+ | Configuration | Dev Agent

Executors CPU Breakdown

48.5 secs

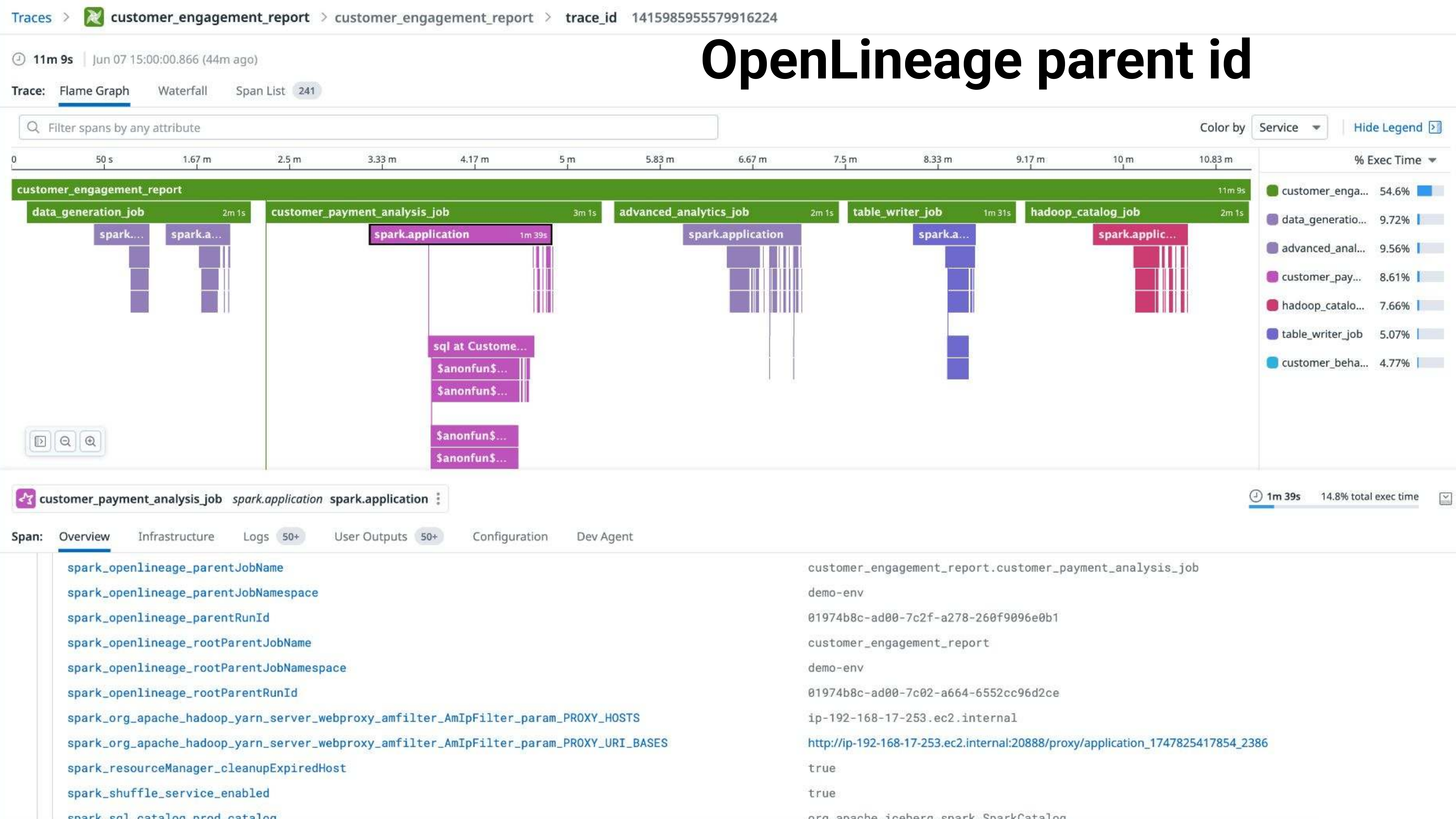
Compute	IO	Shuffle	GC	Spark Internals	Idle
13.1%	10.3%	0.2%	1.2%	10.2%	65.1%

Input / Output

INPUT		OUTPUT		SHUFFLE	
Size	7.72 KiB	Size	7.72 KiB	Size	96.4 KiB
Records	2.21k	Records	105	Records	815

Memory usage

OpenLineage parent id



Traces > customer_engagement_report > customer_engagement_report > trace_id 1415985955579916224

11m 9s | Jun 07 15:00:00.866 (44m ago)

Trace: Flame Graph Waterfall Span List 241

% Exec Time		
🟩	customer_enga...	54.6%
■	data_generatio...	9.72%
■	advanced_anal...	9.56%
■	customer_pay...	8.61%
🟥	hadoop_catalo...	7.66%
■	table_writer_job	5.07%
■	customer_beha...	4.77%

customer_payment_analysis_job *spark.application* spark.application

1m 39s 14.8% total exec time

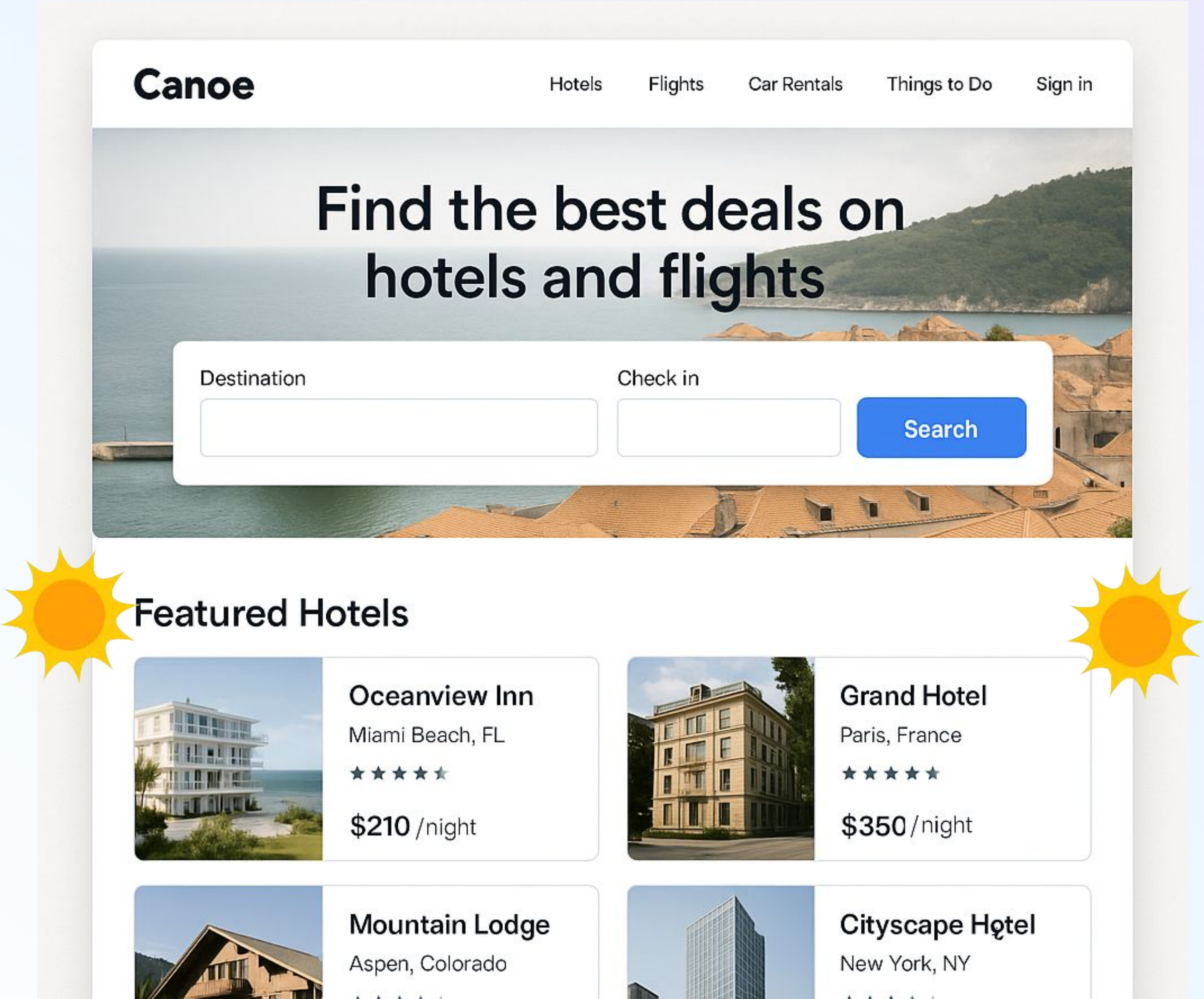
Span: Overview Infrastructure Logs 50+ User Outputs 50+ Configuration Dev Agent

spark_openlineage_parentJobName	customer_engagement_report.customer_payment_analysis_job
spark_openlineage_parentJobNamespace	demo-env
spark_openlineage_parentRunId	01974b8c-ad00-7c2f-a278-260f9096e0b1
spark_openlineage_rootParentJobName	customer_engagement_report
spark_openlineage_rootParentJobNamespace	demo-env
spark_openlineage_rootParentRunId	01974b8c-ad00-7c02-a664-6552cc96d2ce
spark_org_apache_hadoop_yarn_server_webproxy_amfilter_AmIpFilter_param_PROXY_HOSTS	ip-192-168-17-253.ec2.internal
spark_org_apache_hadoop_yarn_server_webproxy_amfilter_AmIpFilter_param_PROXY_URI_BASES	http://ip-192-168-17-253.ec2.internal:20888/proxy/application_1747825417854_2386
spark_resourceManager_cleanupExpiredHost	true
spark_shuffle_service_enabled	true
spark_sql_catalog_prod_catalog	org.apache.iceberg.spark.SparkCatalog

OMG the possibilities are endless

- Dependency tracing
- Root cause identification
- Issue prioritization
- Impact mapping
- Precision backfills
- Anomaly detection
- Change management
- Historical analysis
- Compliance

Problem fixed!



Find the best deals on hotels and flights

Destination

Check in

Featured Hotels

Oceanview Inn
Miami Beach, FL
★★★★½
$210 /night

Grand Hotel
Paris, France
★★★★★
$350 /night

Mountain Lodge
Aspen, Colorado

Cityscape Hǫtel
New York, NY

How Can OpenLineage Benefit You?

- Identify your key datasets and jobs

- Instrument with OpenLineage: Airflow, Spark, dbt, Flink, etc.

- Profit!
  … or at least have trust in your data!

Thank you!

Open Lineage

github.com/OpenLineage

OpenLineage.io

@OpenLineage

docs.datadoghq.com/data_observability