OpenLineage

Implementing OpenLineage

A work in progress

Sheeri Cabral

Meeting the OpenLineage Spec

- What does it mean to meet the OpenLineage Spec?
  - 100% compliance is not required
    - Just like “standard” SQL
  - Likely compatibility among producers and consumers

1

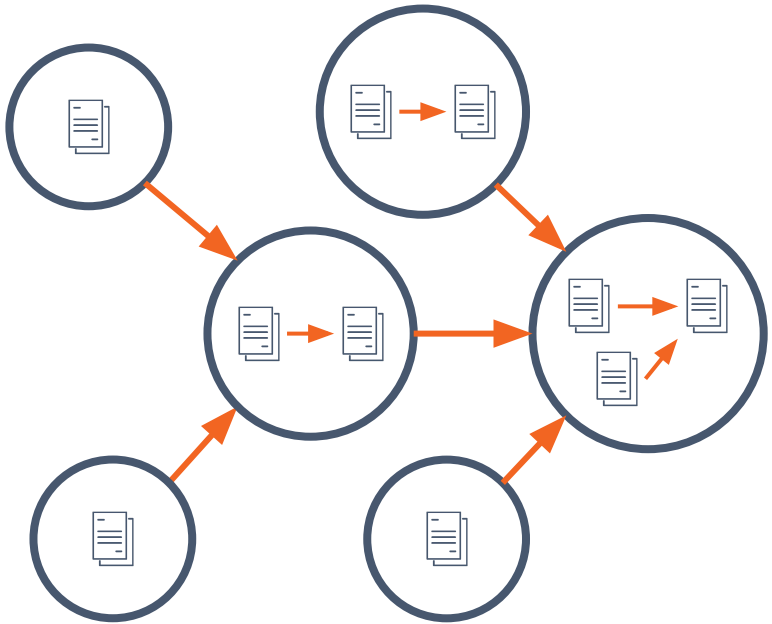
What is OpenLineage?

"OpenLineage is an open framework for data lineage collection and analysis."

What is data lineage?

Data lineage is the set of complex relationships between datasets and jobs in a pipeline.

- Producers & consumers of each dataset
- Inputs and outputs of each job



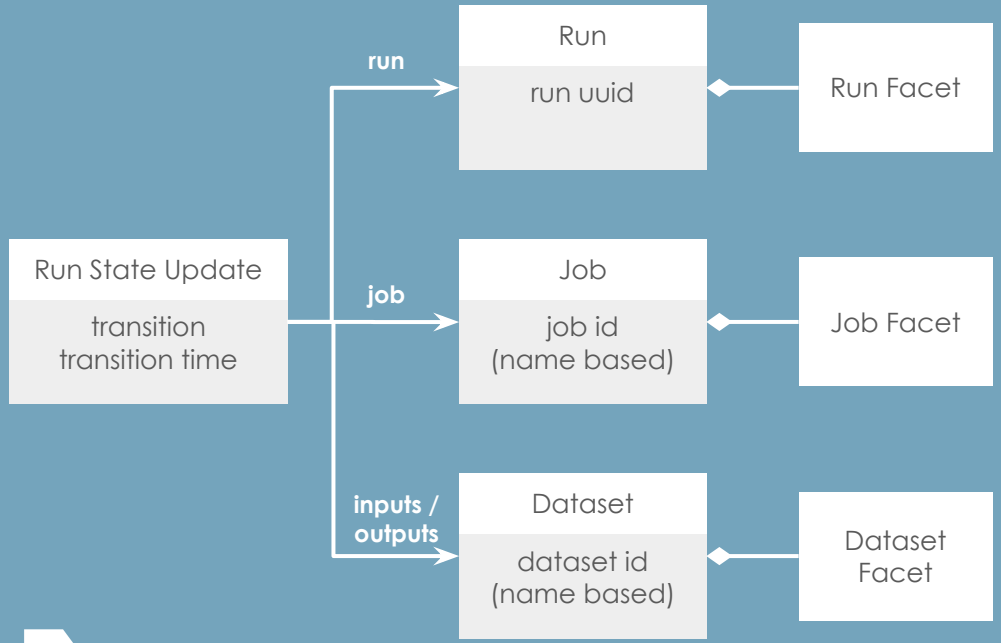
Minimum Viable Lineage

- At least one circle
- Zero or more lines
- Associated information

2

The OpenLineage Spec

Data model:
"An event runs a job on a dataset"



Built around core entities:
Datasets, Jobs, and Runs

Defined as a JSON
Schema spec

Consistent naming for:
Jobs (*scheduler.job.task*)
Datasets (*instance.schema.table*)

Required in the Spec

- Run
  - UUID
- Run State
  - Transition (type)
  - Event time
- Job
  - Namespace
  - Job name
- Datasets
  - Namespace
  - Dataset name

What is a “run”?

- All the events for a run UUID
- Must have >=1 input dataset and >=1 output dataset
- Should have associated info

OPINION: Necessary Per RUN

- At least one box
  - input/output datasets
- At least one line
  - >=1 input dataset
  - >=1 output dataset
- Everything else is optional
  - eventTime/nominalTime
  - Producer
  - Schema URL
  - Code/SQL
  - Job parent
  - errorMessage
  - outputStatistics

OpenLineage Query Example


```
{
 "eventType": "START",
 "eventTime": "2022-12-08T11:00:00.081Z",
 "run": {
  "runId": "3b452093-782c-4ef2-9c0c-aafe2aa6f34d",
 },
 "job": {
  "namespace": "airflow",
  "name": "etl_delivery_7_days.py",
  "sql": "INSERT INTO DELIVERY_7_DAYS…"
 },
 "inputs": [
  { "namespace": "OPENLINEAGE",
    "name": "FOOD_DELIVERY.ORDERS_7_DAYS" },
  { "namespace": "OPENLINEAGE",
    "name": "FOOD_DELIVERY.CUSTOMERS" }
…
 ],
 "outputs": [
  {
   "namespace": "OPENLINEAGE",
   "name": "FOOD_DELIVERY.DELIVERY_7_DAYS",
  }
 ],
  "producer":
"https://github.com/User/Project/blob/v1-0-0/client",
  "schemaURL":
"https://company.com/spec/1-0-0/RunEventLineage"
}
```

OpenLineage


```
{
 "eventType": "START",
 "eventTime": "2022-12-08T11:00:00.081Z",
 "run": {
  "runId": "3b452093-782c-4ef2-9c0c-aafe2aa6f34d",
 },
 "job": {
  "namespace": "airflow",
  "name": "etl_delivery_7_days.py",
  "sql": "INSERT INTO DELIVERY_7_DAYS…"
 },
 "inputs": [
  { "namespace": "OPENLINEAGE",
    "name": "FOOD_DELIVERY.ORDERS_7_DAYS" },
  { "namespace": "OPENLINEAGE",
    "name": "FOOD_DELIVERY.CUSTOMERS" }
…
 ],
 "outputs": [
  {
   "namespace": "OPENLINEAGE",
   "name": "FOOD_DELIVERY.DELIVERY_7_DAYS",
  }
 ],
  "producer":
"https://github.com/User/Project/blob/v1-0-0/client",
  "schemaURL":
"https://company.com/spec/1-0-0/RunEventLineage"
}
```

```
{
 "eventType": "COMPLETE",
 "eventTime": "2022-12-08T11:02:32.081Z",
 "run": { "runId": "3b452093-782c-4ef2-9c0c-aafe2aa6f34d" }
}
 "producer": "https://github.com/User/Project/blob/v1-0-0/client",
 "schemaURL": "https://company.com/spec/1-0-0/RunEventLineage"
}
```

Meeting the OpenLineage Spec

- What does it mean to meet the OpenLineage Spec?
  - 100% compliance is not required
    - Just like “standard” SQL
  - Likely compatibility among producers and consumers

The OpenLineage Community

- Discuss what it means to be compliant with the spec
  - Self-test for producers/consumers
- Producers and consumers
  - Be clear on deviations from and extensions to spec

Discussion