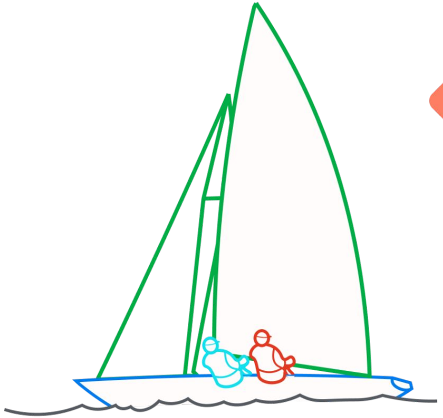
Data Lineage in Open Cloud

Michał Modras



Airflow Summit

Let’s flow together

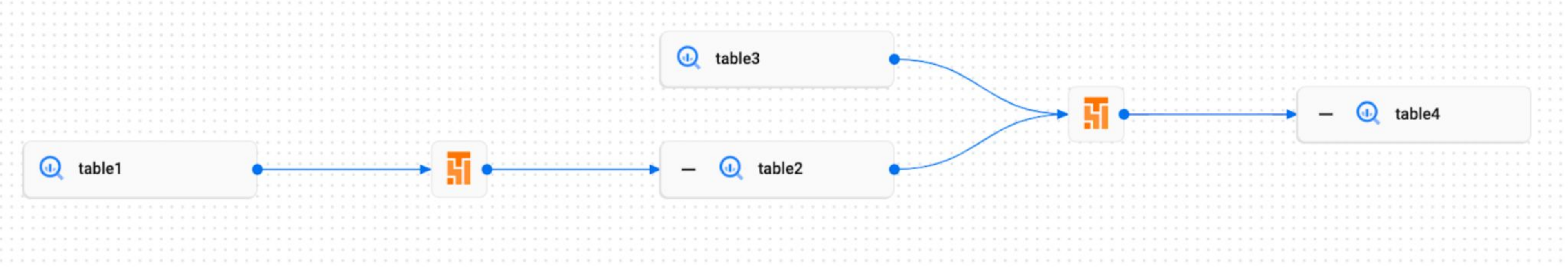
September 19-21, 2023,
Toronto, Canada

Google Cloud

Cloud Composer: Apache Airflow in Google Cloud

Michał Modras
Engineering Manager

Data lineage traces the relationship between data sources based on movement of data, explaining how data was sourced and transformed.



01

Motivation for data lineage

Customer challenges

01 | **Inability to understand and trust data**

“My manager just asked me if I am using the table from the authoritative source—how can I check this quickly?”

*“OK, I am taking a look into this dashboard, but **where is this information is coming from**? What is the **database** that is **supporting this dashboard**?*

Customer challenges

02

Inability to do deterministic change management

“What happens if I drop a table/change a column?”

*“We have huge systems, sometimes **we change** something, and we pray. **We pray** so not for someone, to go “oh, what happened”?*

“oh this table is changing, but I have queries, I have dashboards, I have infinite things plugged on this table.

***I need to map out by hand what happened.** “*

Customer challenges

“oh this table is changing, but I have queries, I have dashboards, I have infinite things plugged on this table.

***I need to map out by hand what happened.** “*

03 **Inability to do effective root cause analysis**

***“**There are issues in the data in a given table—how can I quickly zero in on the potential cause for the issue?”*

*“...they reach out to the source systems, SAP systems, third party vendors etc... (debug)**can take up to 2 weeks”***

Customer challenges

*“ to have **more faith in our system,** that what we're providing to them is correct”*

*“It’s **important** for them to **understand where everyone is getting the information**, what **transformations** are being done along the way so that they can understand if they are **comparing oranges with oranges.**”*

04

Inability to meet compliance requirements effectively

“How can I guarantee to authorities that I have not used prohibited data in my models to introduce bias?”

Customer challenges

*"It's **important** for them to **understand where everyone is getting the information**, what **transformations** are being done along the way so that they can understand if they are **comparing oranges with oranges.**"*

05

Inability to manage data estate at scale

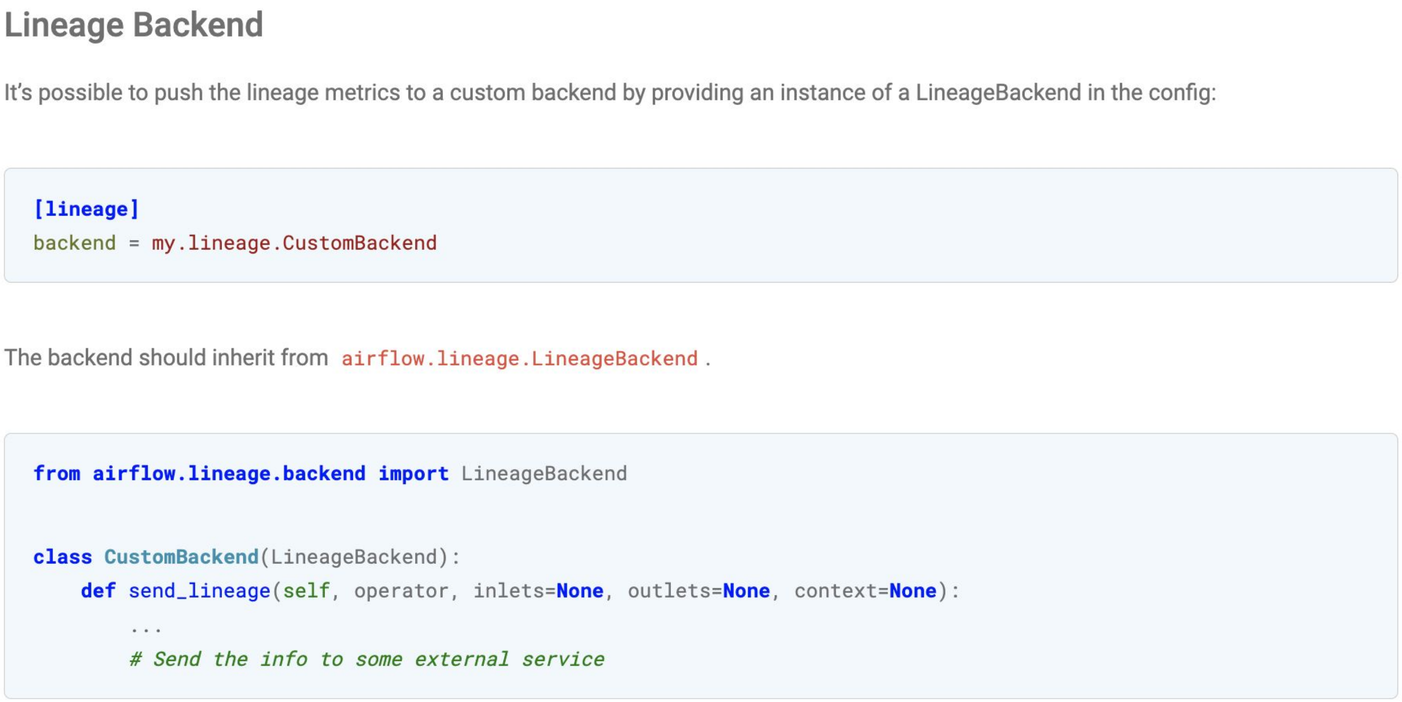
"Help me auto curate/auto apply policies based on lineage to automatically manage data"

02

Data lineage in Apache Airflow

Airflow Lineage Backends

- (until recently) Native Lineage feature in Airflow.



Lineage Backend

It's possible to push the lineage metrics to a custom backend by providing an instance of a LineageBackend in the config:

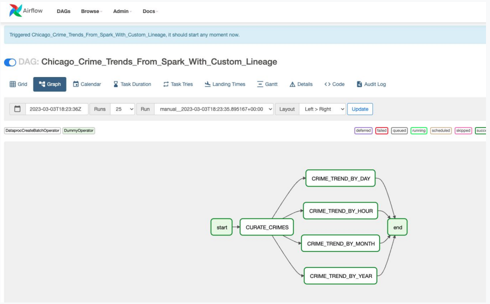
```
[lineage]
backend = my.lineage.CustomBackend
```

The backend should inherit from `airflow.lineage.LineageBackend` .

```
from airflow.lineage.backend import LineageBackend


class CustomBackend(LineageBackend):
    def send_lineage(self, operator, inlets=None, outlets=None, context=None):
        ...
        # Send the info to some external service
```

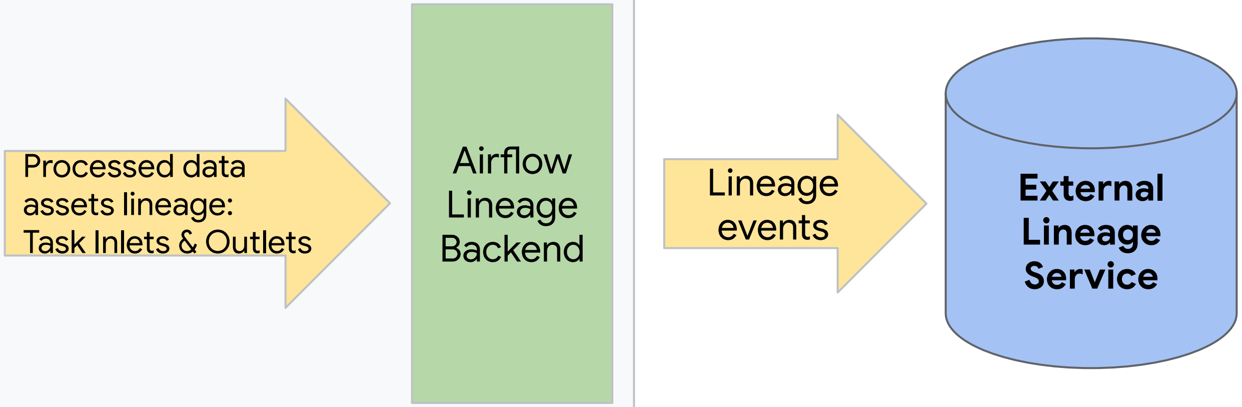
Lineage Reporting Through Airflow Lineage Backend



Triggered Chicago_Crime_Trends_From_Spark_With_Custom_Lineage, it should start any moment now.

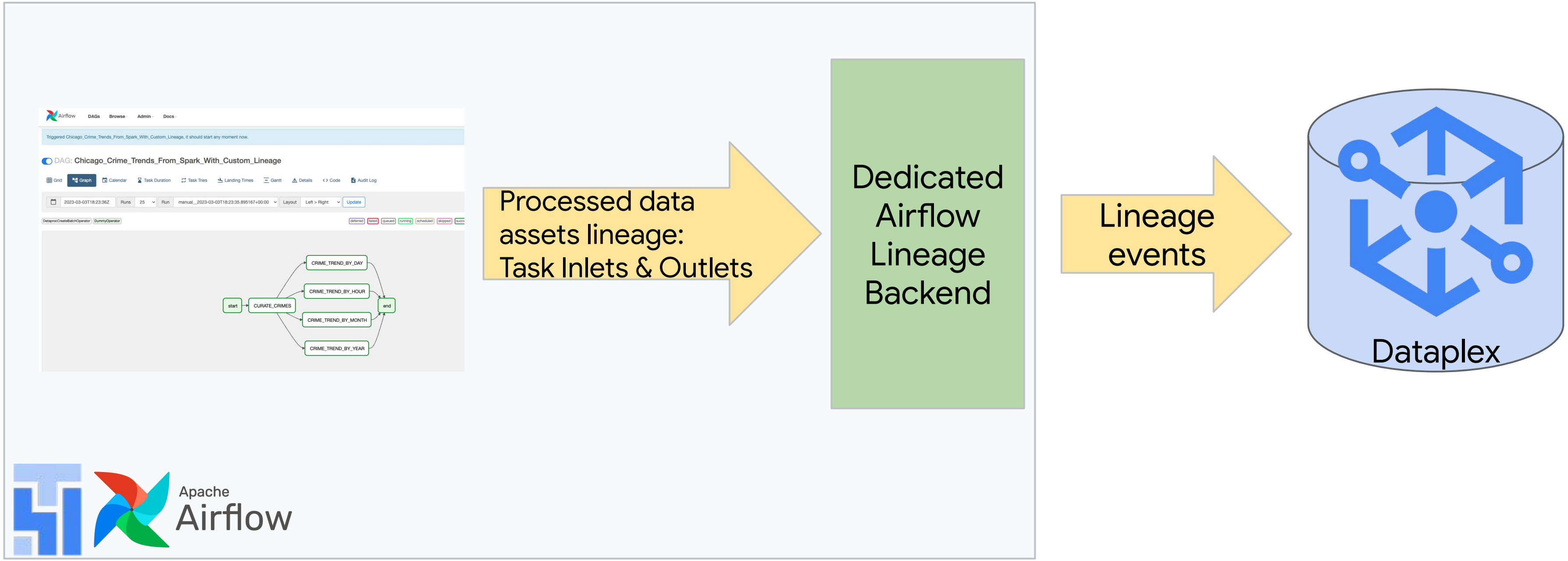
DAG: Chicago_Crime_Trends_From_Spark_With_Custom_Lineage

2023-03-03T18:23:36Z | Runs 25 | Run manual__2023-03-03T18:23:35.895167+00:00 | Layout Left > Right | Update



Apache Airflow

Cloud Composer Dataplex Data Lineage Integration



Triggered Chicago_Crime_Trends_From_Spark_With_Custom_Lineage, it should start any moment now.

DAG: Chicago_Crime_Trends_From_Spark_With_Custom_Lineage

2023-03-03T18:23:36Z | Runs 25 | Run manual__2023-03-03T18:23:35.895167+00:00 | Layout Left > Right | Update

Processed data assets lineage: Task Inlets & Outlets

Dedicated Airflow Lineage Backend

Lineage events

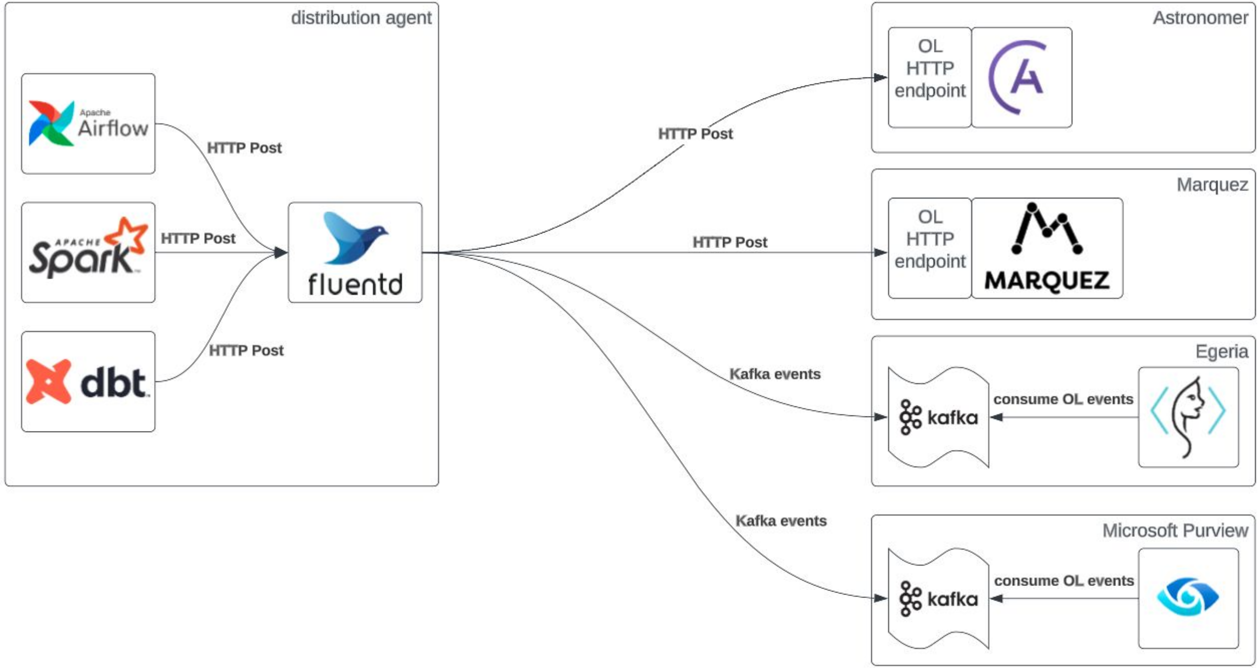
Dataplex

03

OpenLineage in Apache Airflow and Google Cloud

OpenLineage

- Emerging standard for open source lineage metadata transfer.



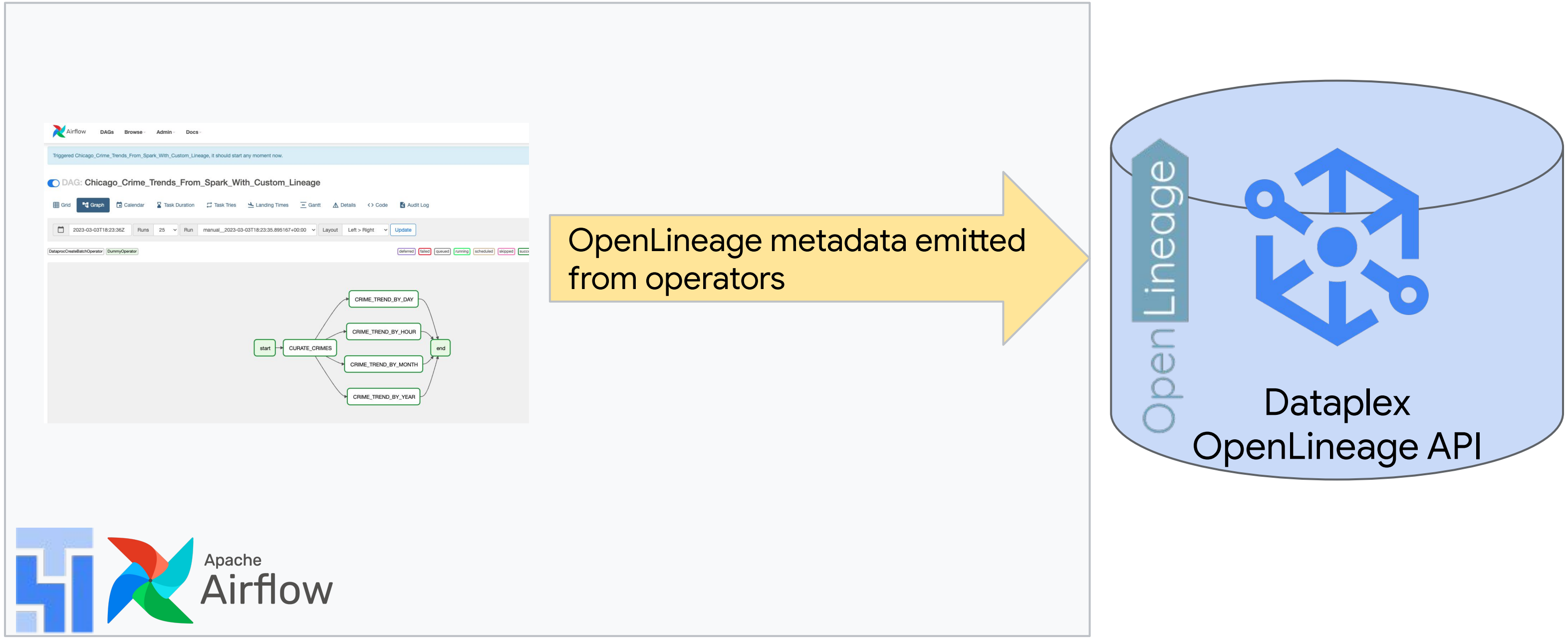
https://openlineage.io/

OpenLineage in Airflow

- Used to be ‘add-on’.
- AIP-53 Introduces native support of OpenLineage.
- Airflow OpenLineage integration modernizes its architecture - e.g. abandoning lineage metadata extractors separate to Airflow operators, making lineage metadata definition close to operators.

Cloud Composer OpenLineage Adoption

- Composer is migrating towards the new Airflow lineage architecture (OpenLineage), and leveraging Dataplex's OpenLineage API.



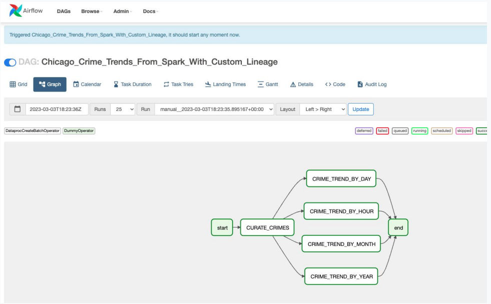
Triggered Chicago_Crime_Trends_From_Spark_With_Custom_Lineage, it should start any moment now.

DAG: Chicago_Crime_Trends_From_Spark_With_Custom_Lineage

2023-03-03T18:23:36Z | Runs 25 | Run manual__2023-03-03T18:23:35.895167+00:00 | Layout Left > Right | Update

Cloud Composer OpenLineage Adoption

- More open and pluggable Lineage in Composer's Airflow.

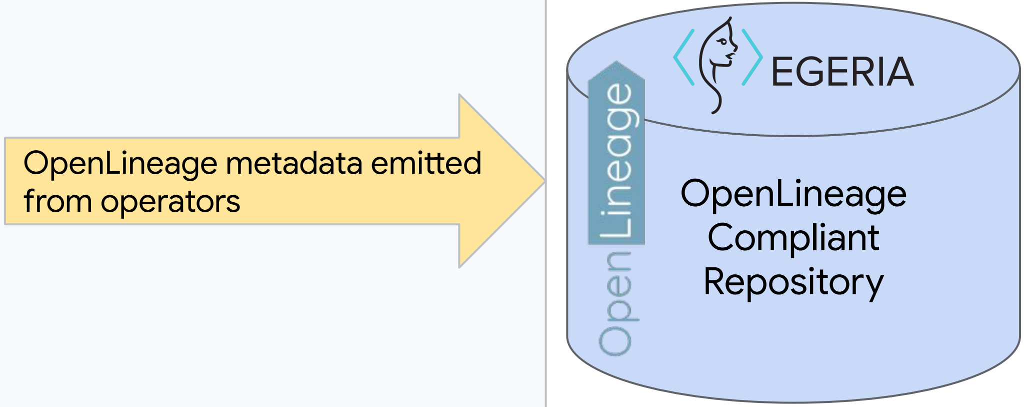


Triggered Chicago_Crime_Trends_From_Spark_With_Custom_Lineage, it should start any moment now.

DAG: Chicago_Crime_Trends_From_Spark_With_Custom_Lineage

2023-03-03T18:23:36Z Runs 25 Run manual__2023-03-03T18:23:35.895167+00:00 Layout Left > Right Update

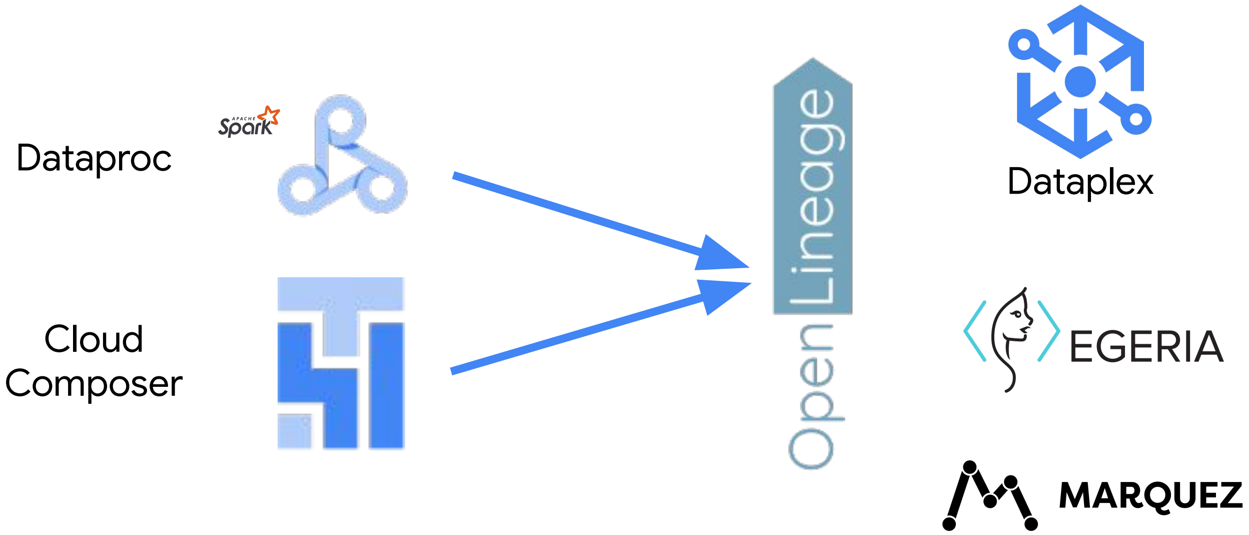
Apache
Airflow



OpenLineage metadata emitted from operators

OpenLineage Compliant Repository

Adoption of OpenLineage in Google Cloud



Dataproc

Cloud Composer

Dataplex

Thank you.

Google Cloud