DATA+AI SUMMIT 2022

Spline

Central data-lineage tracking
Not only for Spark

ORGANIZED BY databricks

Oleksandr Vayda
Lead Software Engineer, ABSA

Danil Vagapov
DevOps Engineer

Introduction

Oleksandr Vayda

Lead Software Engineer, ABSA

ABSA OSS - GitHub

An active open-source community member

Spline - Data lineage Tracking

A modular cross-platform solution for in depth data-lineage tracking.

absaoss.github.io/spline/

K8GB - Kubernetes Global Balancer

A cloud native Kubernetes Global Balancer

CNCF Sandbox Project

www.k8gb.io/

Other projects

ABRiS - Avro SerDe for Spark

Cobrix - A Mainframe/EBCDIC (COBOL) data source for Spark

Hyperdrive - Extensible streaming ingestion pipeline

Atum - A dynamic data completeness and accuracy

Enceladus - Dynamic Conformance Engine

Spot - Spark performance tuning

Data Lineage

Problem statement

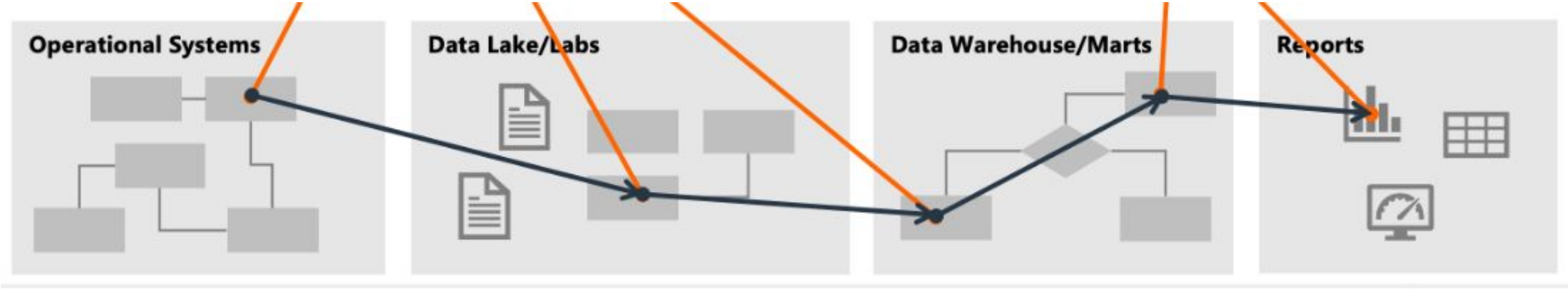
Why is data lineage important?

- Data lineage is one of the most critical components of a data governance strategy for data lakes
- Data accuracy, completeness and trustworthiness
- Regulatory compliance (GDPR, BCBS, etc.)
- Observability of data pipelines for business and tech staff
- Documentation
- Error tracing and impact analysis
- Audit trail and security

Data lineage

... in a nutshell

Data lineage is a data life cycle that includes the data's origins, where it moves over time and what happens to data as it goes through diverse processes.



Spline Project

What is Spline?

- An open-source data-lineage tracking solution.
- Originally created for Apache Spark.

Since version 0.5:

- Multi-framework lineage tracking (“Central Spline”)
- Public API, microservices, ArangoDB as a storage.

GitHub:
https://github.com/AbsaOSS/spline

IEEE paper:
https://ieeexplore.ieee.org/document/8367160

Previous session on Spark Summit:
https://databricks.com/session/spline-apache-spark-lineage-not-only-for-the-banking-industry

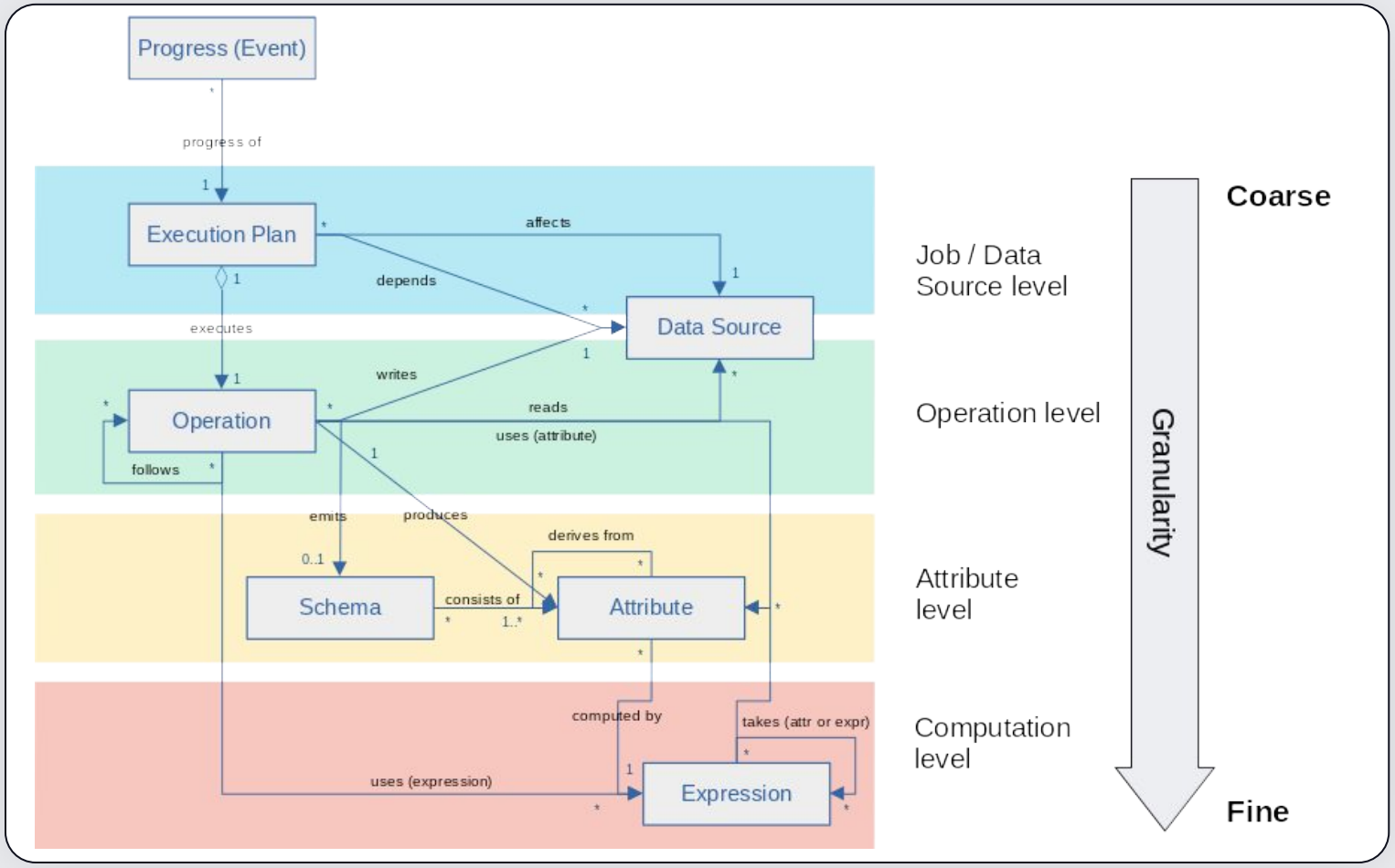
Central lineage tracking

What makes Spline capable for central lineage tracking

- Extensible, data framework independent API and data model
- Transparent for target systems
- Easy integration at different levels
- Scalable, fault-tolerant, highly available, eventually consistent
- On cloud or on-premise usage

Domain model

also persistence model



Layered structure allows for analysing data lineage on different level of granularity

Implementation

Conceptual diagram

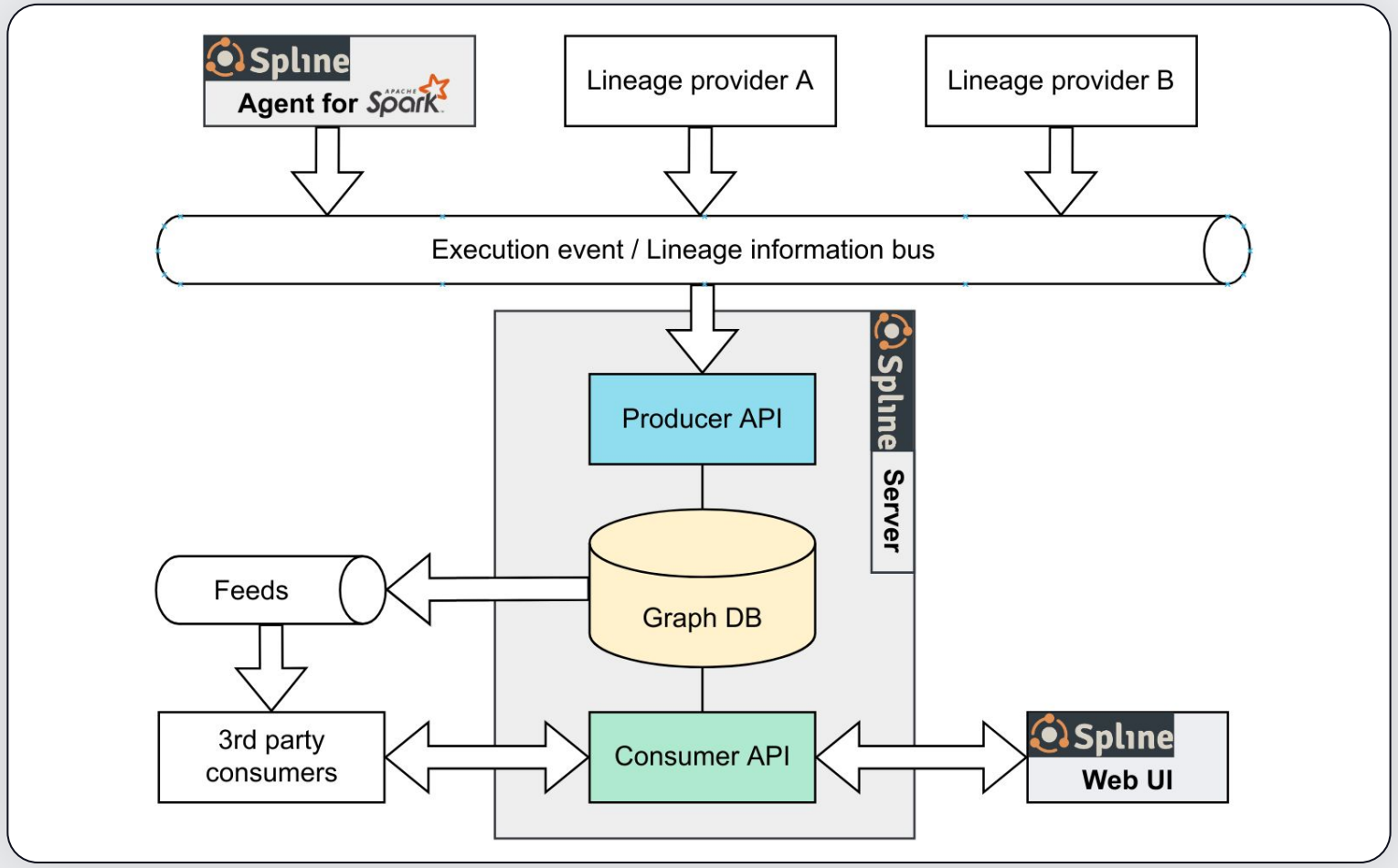
Lineage data is stored in a graph database.

Producer API is used to collect lineage data from various target systems through agents.

Data framework independent lineage model allows for collecting lineage from heterogeneous pipelines.

Consumer API is for querying lineage.

Microservice design allows for extension, integration and reusability.



Spline Agent for Spark

https://github.com/AbsaOSS/spline-spark-agent

- Supports all Spark version > 2.2,
  Both Scala 2.11 and 2.12

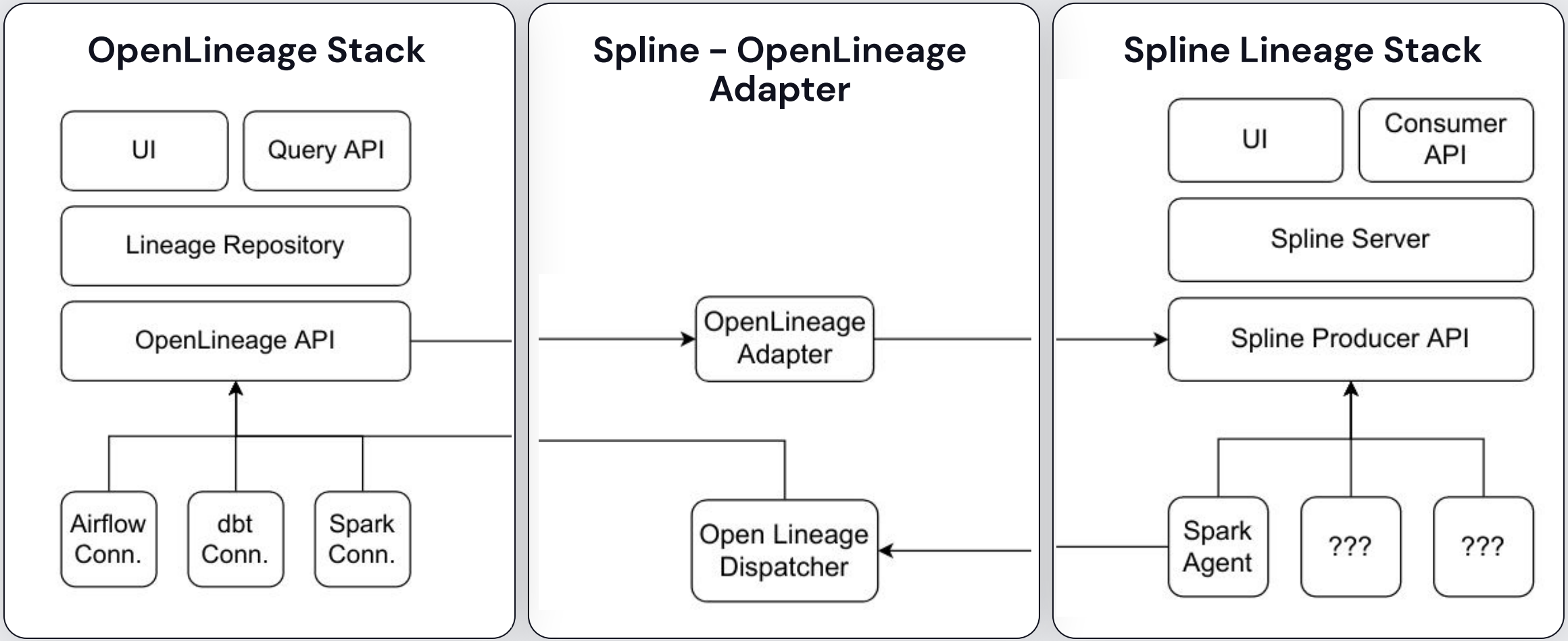
- Seamless integration:
  Shell, Python, notebooks, EMR etc

- High-fidelity and accurate lineage

- Extensible via plugins and add-ons

OpenLineage Integration

https://github.com/AbsaOSS/spline-openlineage



Usage and related work

BMW — Cloud Data Hub

Collecting and visualizing data lineage of Spark jobs

https://rdcu.be/cMds8

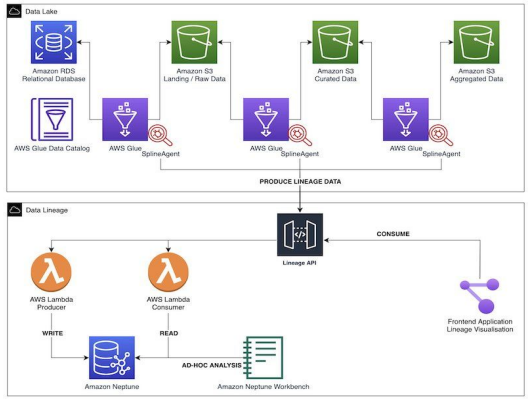
by Alexander Schoenenwald, Simon Kern, Josef Viehhauser & Johannes Schildgen

AWS — Big Data Blog

Build data lineage for data lakes using AWS Glue, Amazon Neptune, and Spline

aws3.link/halvrH

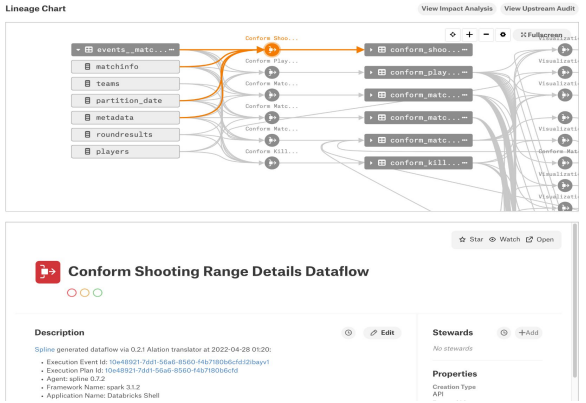
by Khoa Nguyen, Krithivasan Balasubramaniyan, and Rahul Shaurya



Riot Games — In-house

Downstream and upstream impact analysis and GDPR compliance.

We integrate with the Spline consumer API to pull direct and control lineage and publish it to our enterprise data catalog (Alation).

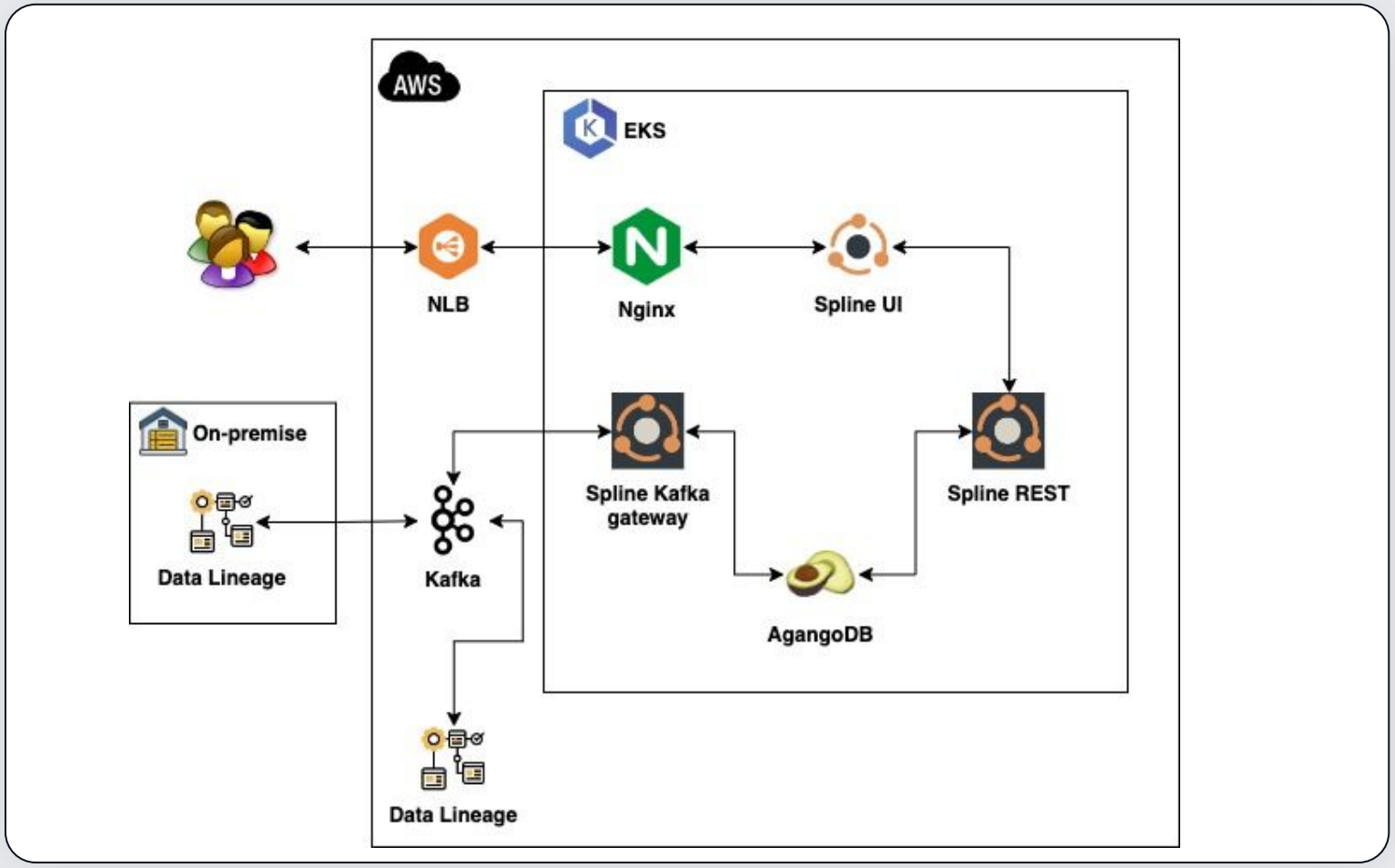


Deployment & Demo

Danil Vagapov
DevOps Engineer

Deployment diagram

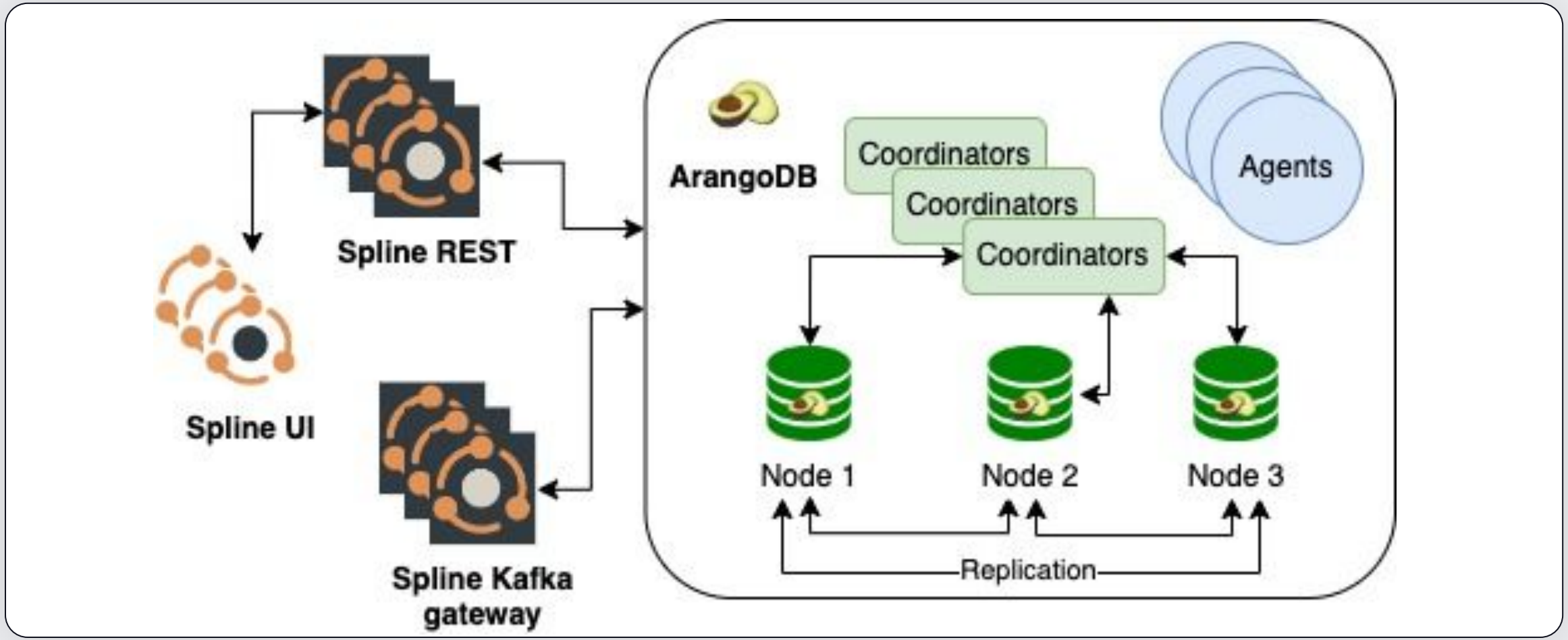
Production Spline view with Kafka



- Kafka as a transport
- Decoupled producers
- Lineage data durability
- **Important**: logically related messages have to be in the same partition

Deployment diagram

Main Spline services



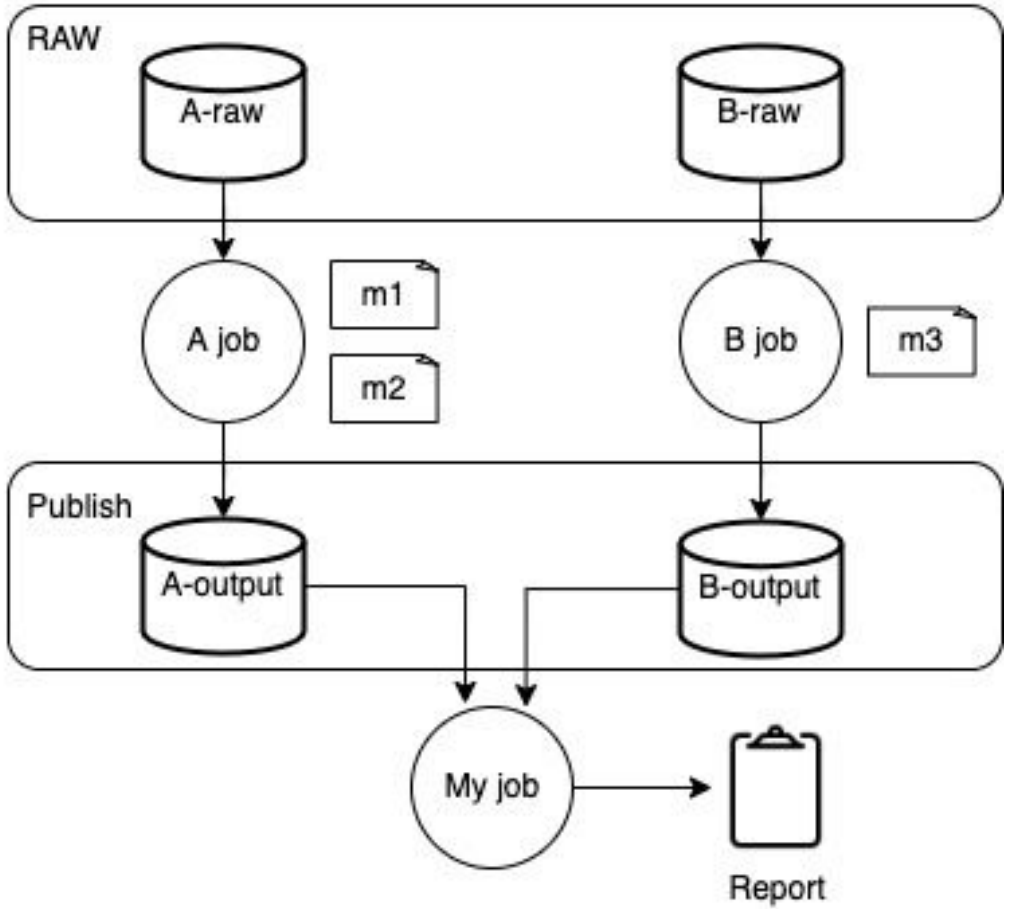
Spline Demo

Deployment

```
1
2
3
4
5
6
7
8
9
10
```

Spline Demo

Demo setup



Spline Demo (Cont.)

Explore lineage on UI

```
1
2
3
4
5
6
7
8
9
10
```

Challenges & future work

Oleksandr Vayda
Lead Software Engineer, ABSA

Assumptions & Limitations

Pillars and boundaries of the current Spline data model

1. A single output source per job
2. No partitions
3. Transactional access to data sources (including streaming based ones)
4. Execution plans never change
5. No streaming support yet
6. (Spark agent) No automatic RDD support

Future work

What comes to Spline in the near future?

1. Scaling and performance
2. Consistency and reliability
3. More UI features (lineage analysis, explore, reports, alerts etc)
4. Data source aliases
5. Streaming support
6. RDD support (for the Spark agent)

Acknowledgements

- Adam Cervenka
- Dzmitry Makatun
- Aaisha Osman
- Alexander Schoenenwald

DATA+AI
SUMMIT 2022

https://github.com/AbsaOSS

Thank you

Oleksandr Vayda
Lead Software Engineer

Danil Vagapov
DevOps Engineer