
AWS
re:Invent
NOV. 28 – DEC. 2, 2022 | LAS VEGAS, NV

OPN303

Modern data exchange using AWS data streaming

Ali Alemi

Analytics Specialist SA
Amazon Web Services

Jeremy Ber

Analytics Specialist SA
Amazon Web Services

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Agenda

Challenges with traditional data replication tools

Apache Hudi for transactional data lakes

Build a modern data replication using Amazon Kinesis Data Analytics for Apache Flink

Overview of the solution

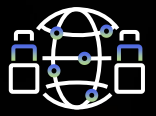
Call to action

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Challenges with traditional replication tools



Difficult to maintain



Not flexible



Centrally managed



Security risks



Expensive to scale

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Transactional data lakes meet stream data processing



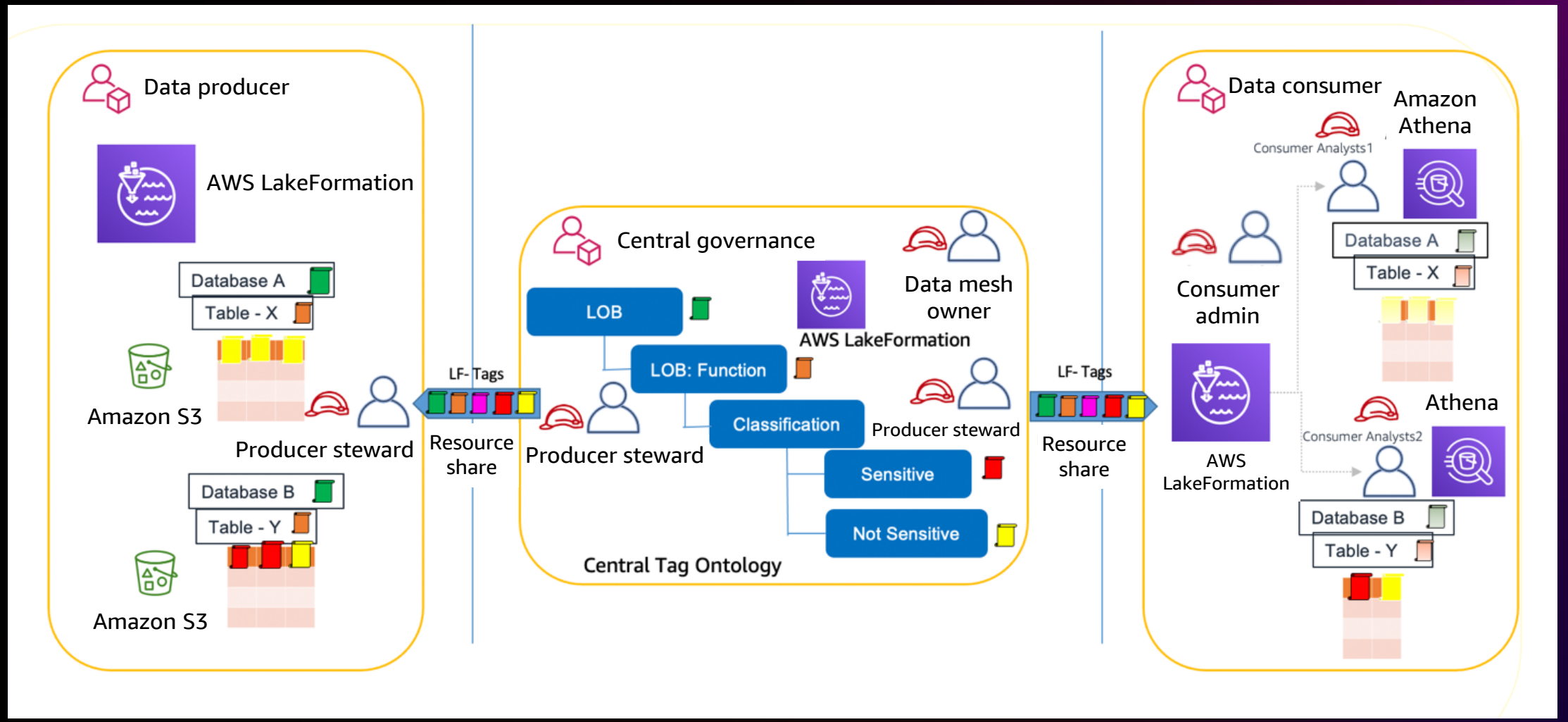
- Apache Hudi
  - Sync the data in a data lake with updates and deletes on the source systems
  - Low-latency record-level updates on Amazon S3 by utilizing data indexes
  - Provides read-optimized and real-time view to cover a variety of use cases



- Apache Flink
  - Popular framework for building stateful streaming and batch pipelines
  - Fault tolerant
  - Multi-language, multiple connectors including custom connectors and sinks
  - Exactly once processing semantics
  - Flink SQL APIs to simplify building streaming jobs (in SQL syntax)

aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

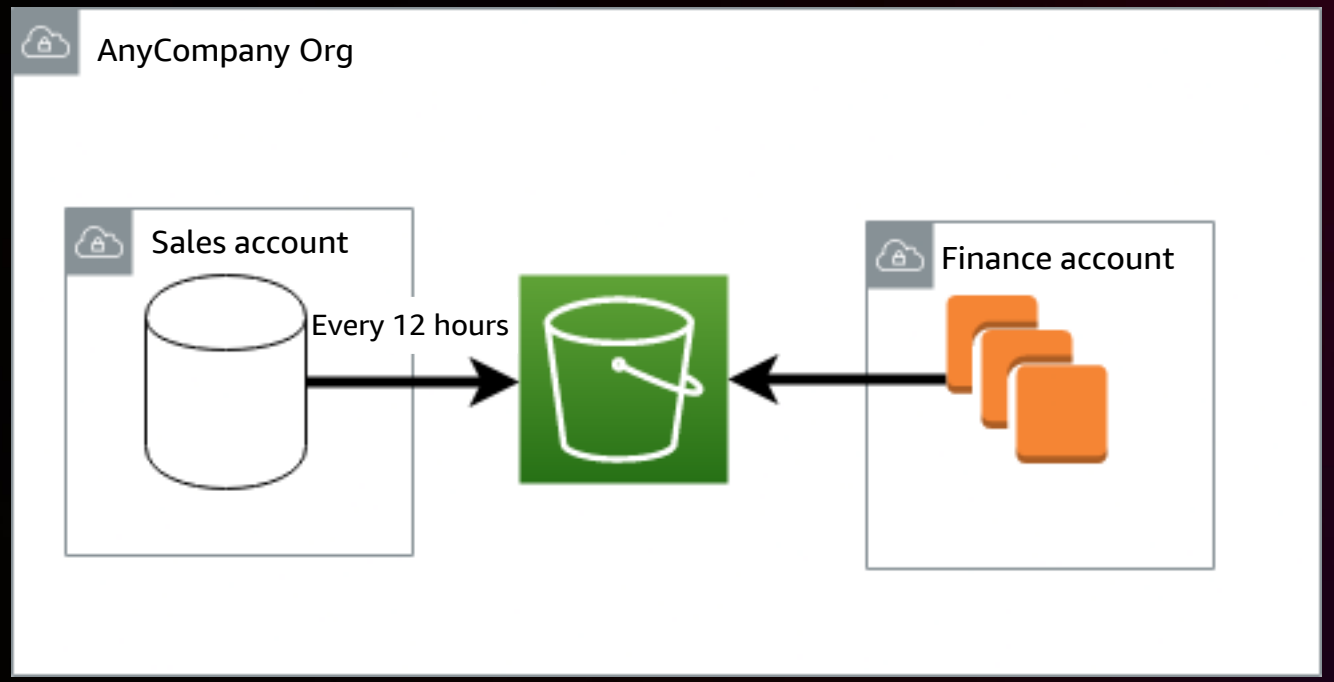
Modern data architecture on AWS



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Data replication use case

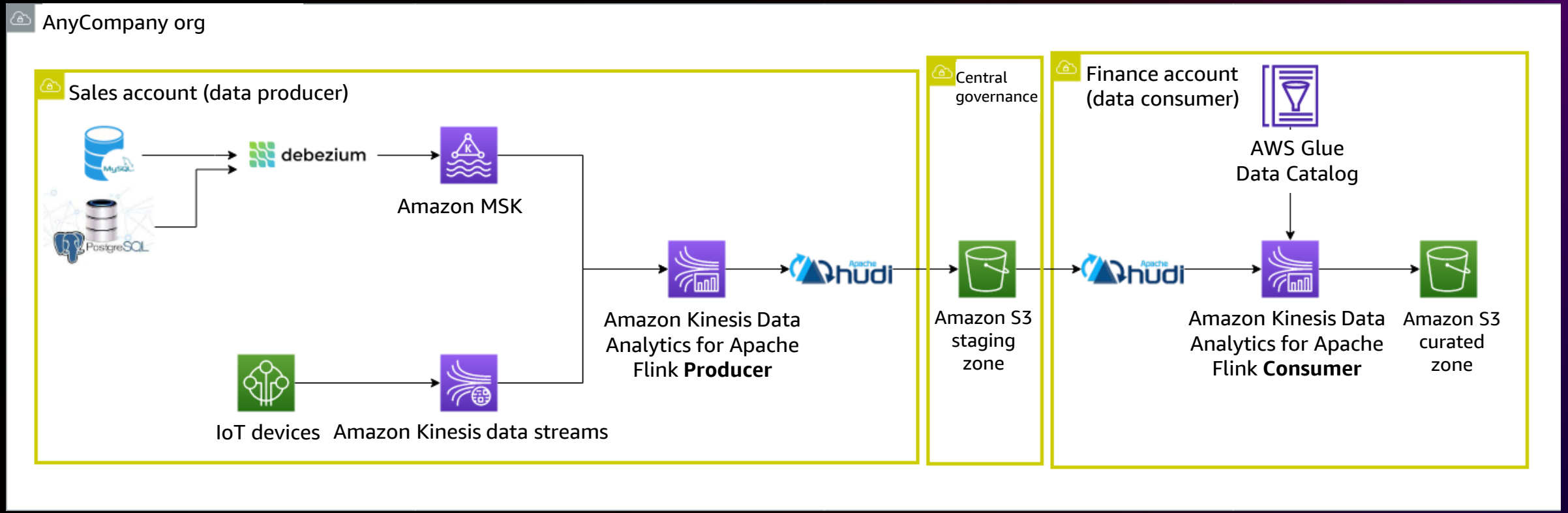
- AnyCompany sales platform tracked in database
- Need to share data with finance department
- Dump database to shared Amazon S3 bucket periodically



aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Modern data replication (setup/architecture) on AWS

USING DATA STREAMING SERVICES



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Postgres, PostgreSQL and the Slonik Logo are trademarks or registered trademarks of the PostgreSQL Community Association of Canada, and used with their permission.

Conclusion and call for action

- Traditional replication tools do not scale to the size of today's datasets
- Decentralizing the ownership of the data pipelines reduces organizational friction
- AWS modern data architecture for data mesh democratizes data and its ownership
- Apache Hudi connector for Apache Flink makes data streaming possible to and from a data lake

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Call to action

- Modern data architecture on AWS
- Apache Flink official documentation
- Apache Hudi official Flink Guide
- Create a low-latency source-to-data lake pipeline using Amazon MSK Connect, Apache Flink, and Apache Hudi
- Amazon Kinesis Data Analytics for Apache Flink, using Hudi Connector Code example
- Amazon EMR
- Amazon Kinesis Data Analytics for Apache Flink
- Introducing Amazon Kinesis Data Analytics Studio

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Thank you!

Ali Alemi
alialem@amazon.com

Jeremy Ber
jdber@amazon.com



Please complete the session survey in the **mobile app**

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.