INNOVATE

DATA EDITION

19 August 2021

Beyond batch processing - real-time analytics at scale with Apache Flink

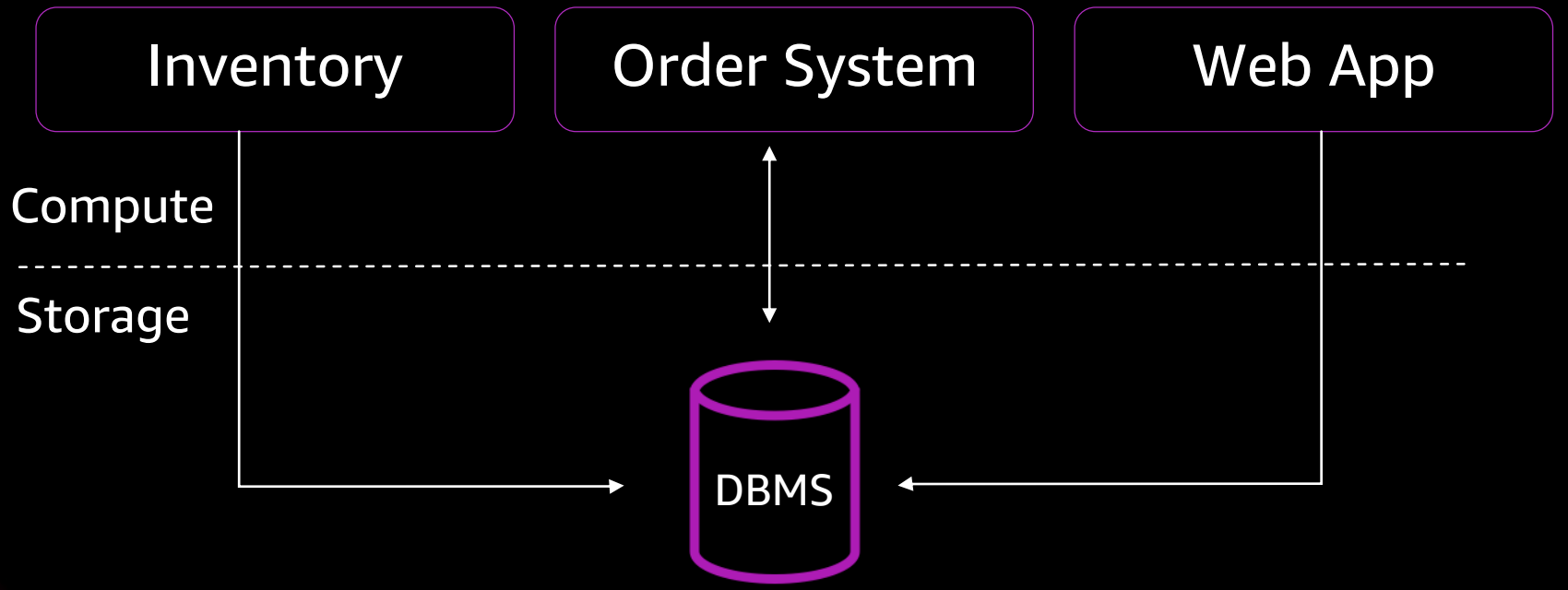
Masudur Rahaman Sayem
Analytics Specialist SA
Amazon Web Services

Agenda

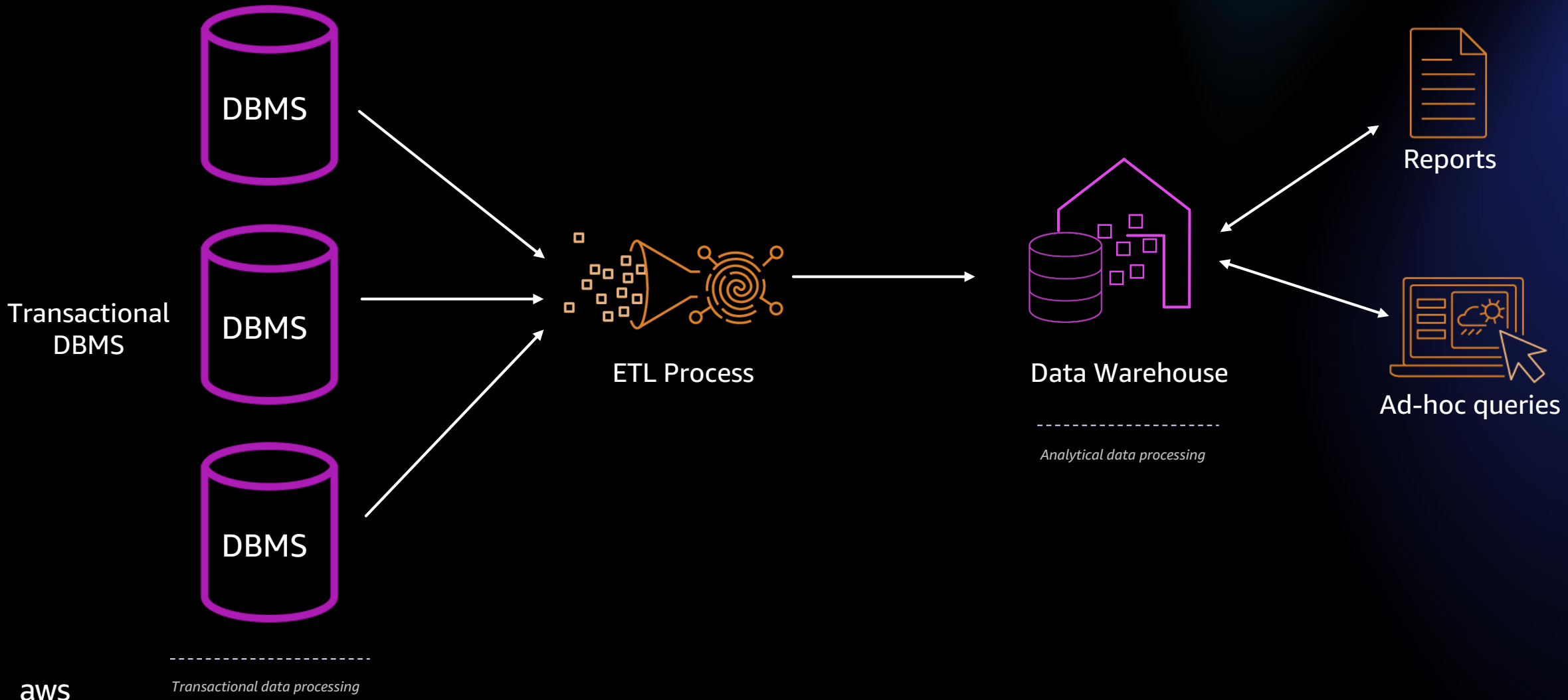
- Data processing pattern
- Overview of Apache Flink and Amazon Kinesis Data Analytics
- Streaming analytics reference architecture
- A demo with Amazon Kinesis Data Analytics Studio

The evolution of data processing

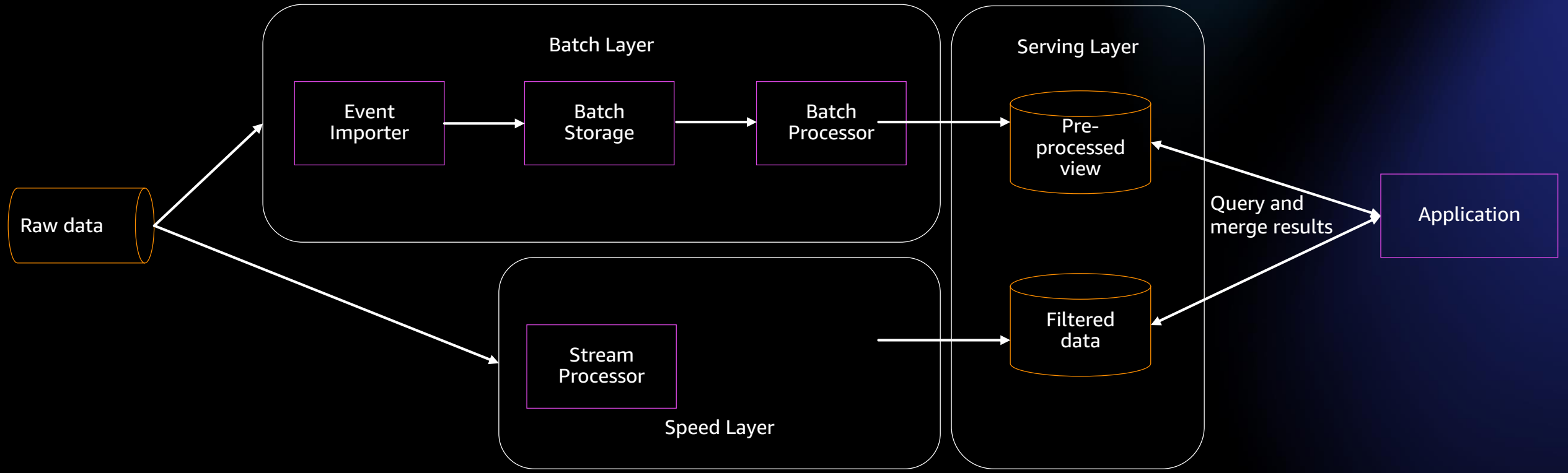
Transactional data processing



The evolution of data processing



The evolution of data processing



Why Apache Flink?

Diverse use-cases

- Event-driven Applications
- Streaming Analytics & ETL
- Batch Analytics

Expressive APIs

- SQL
- Table API
- DataStream API
- Stateful Functions

Processing guarantees

- Exactly-once state consistency
- Event-time processing
- Late data handling

Scale-out architecture

- Adapt to desired throughput
- Support for TBs of state

Community

- Vibrant open source community
- Broad set of connectors

Apache Flink deployment targets

YARN

Apache Mesos

Kubernetes

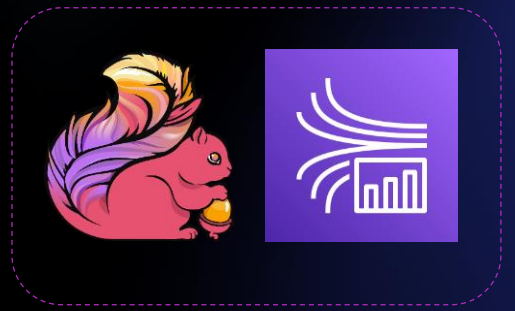
Amazon Kinesis Data Analytics

Challenges of running Apache Flink

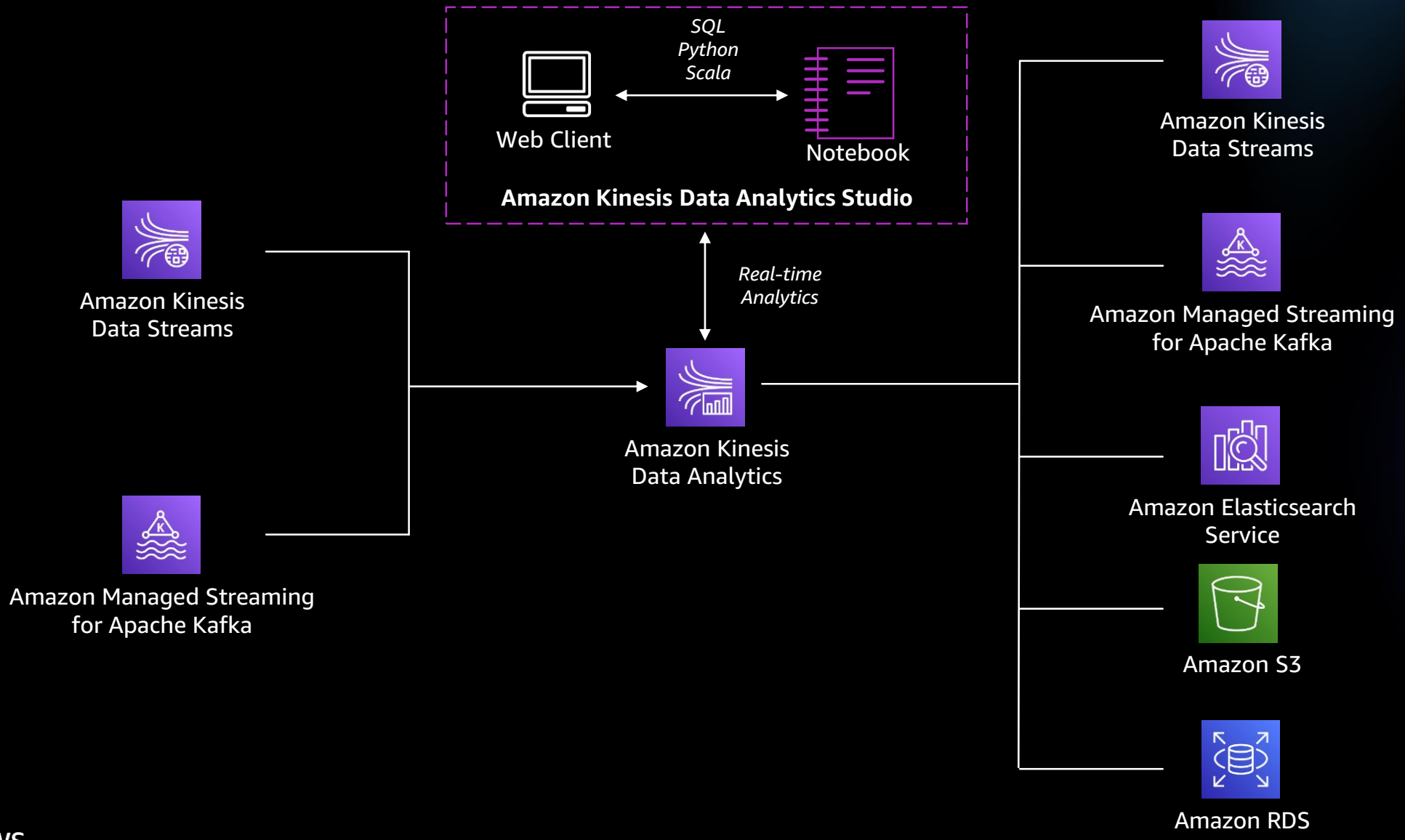
- On-premises Hadoop cluster stretched to its limits
- Not easily scalable
- Cost to upgrade unviable
- Infrastructure managed by internal team

Amazon Kinesis Data Analytics

- No servers to manage
- Sub-second latencies
- Monitoring and high availability built-in
- Automatic scaling capabilities



Integration with different sources and sinks



Amazon Kinesis Data Streams

Collect and store data streams for analytics

- Long term retention (365 days)
- Easy administration and low cost
- Real-time, elastic performance
- Secure, durable storage
- Avg 70 ms latency with Enhanced Fan Out

Amazon Managed Streaming for Apache Kafka (Amazon MSK)

Collect and store data streams for analytics

- Vertical and horizontal scaling
  Encryption in transit and at rest
- Deep integration with AWS services
  Fully compatible with Open source Apache Kafka
  Choice of 2 AZ or 3 AZ deployments
- Lowest TCO in the Industry
  IAM, mutual TLS, SASL/ SCRAM authentication

Kinesis Data Analytics other features

Checkpoints and savepoints

- Out-of-the-box RocksDb for state
- Default incremental, asynchronous state snapshots to Amazon S3

Scaling

- Ability to scale using parallelism config
- CPU-based auto scaling

Logs and metrics

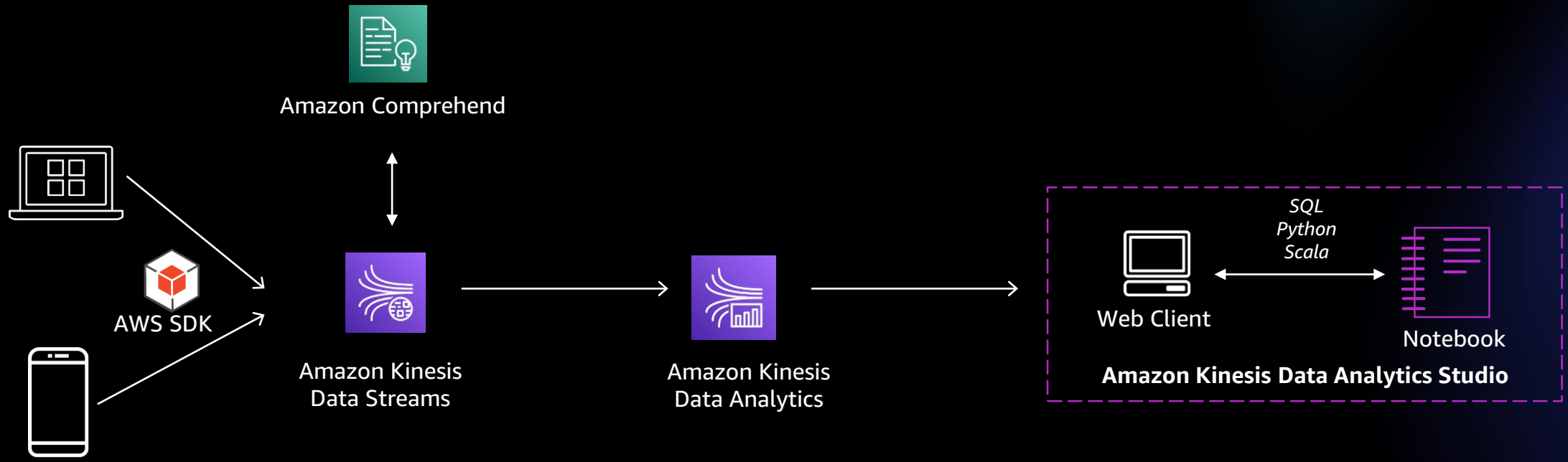
- CloudWatch integration for key Apache Flink and custom metrics
- CloudWatch logs integration for Apache Flink application logs

Other

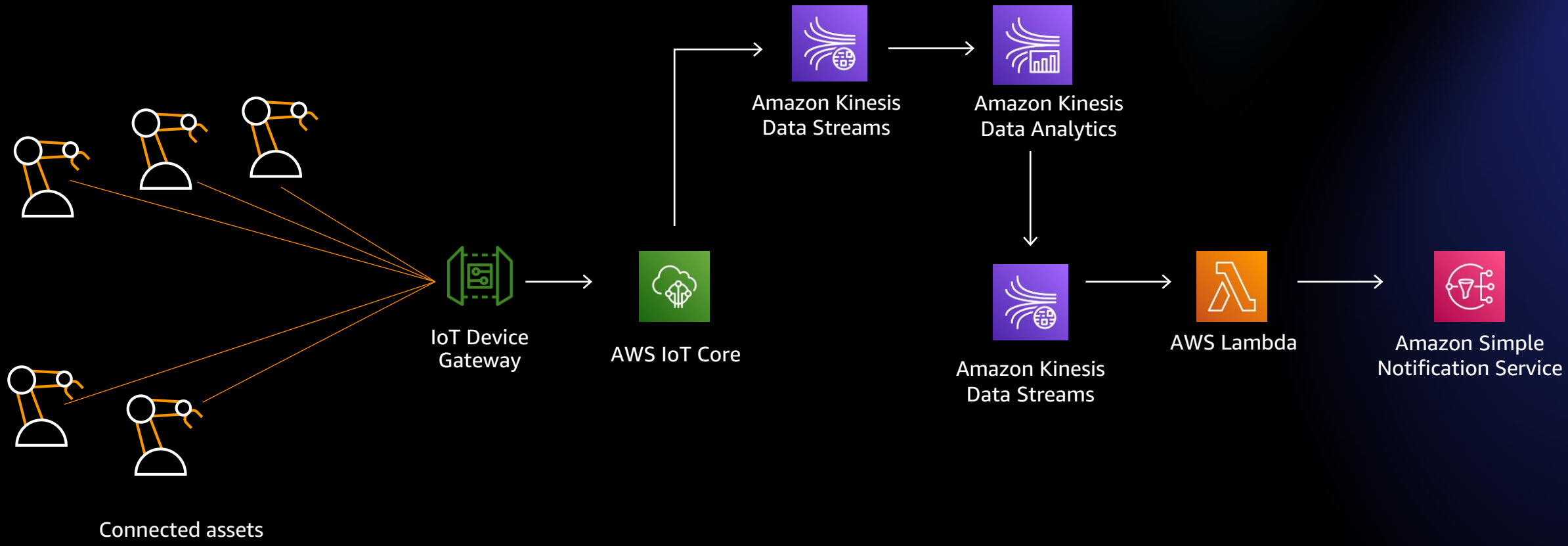
- SOC, PCI, HIPAA complaint
- single-tenant Apache Flink cluster
- JobManager in high availability mode
- Kinesis Data Analytics Studio

Real-world use cases

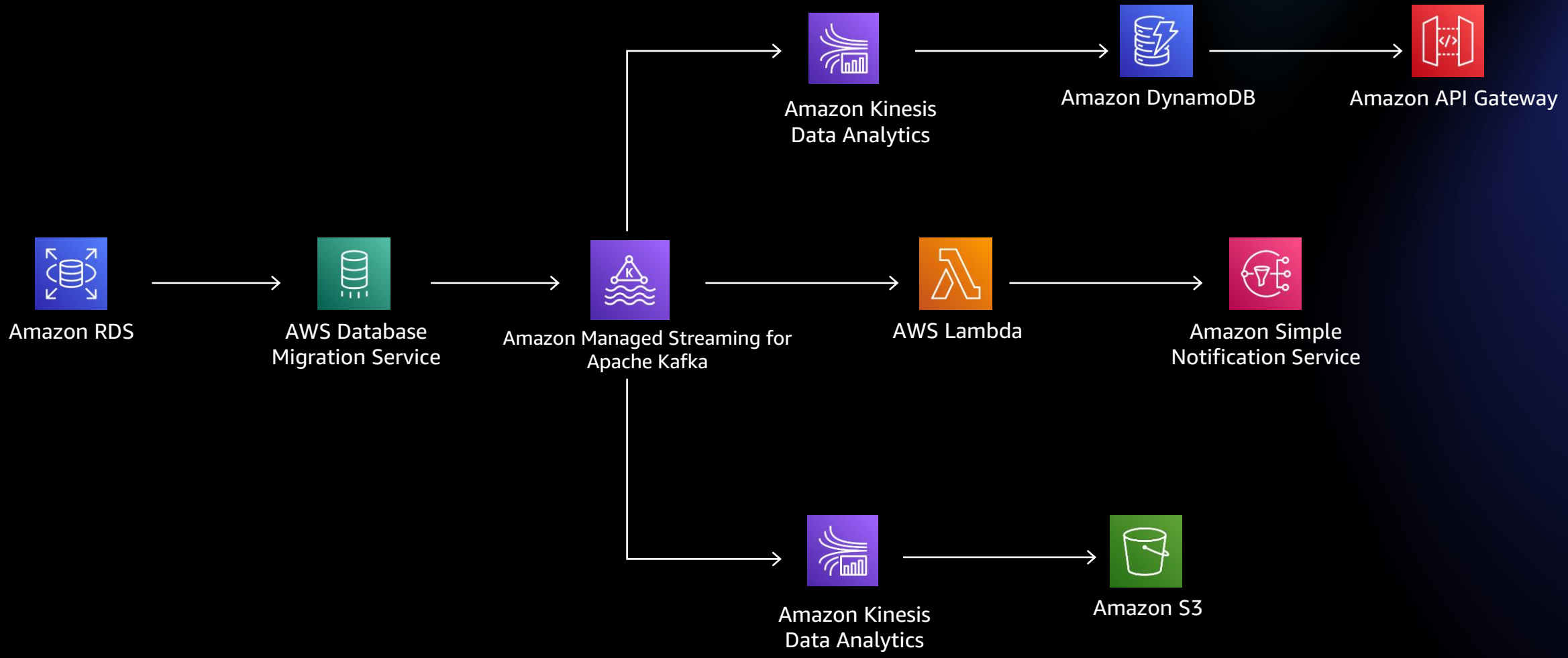
Real-time sentiment analysis on customer feedback



Asset condition monitoring



Change data capture (CDC)



Demo

Recap

- Business wants to analyze data faster; real-time streaming data analytics is the best way to get timely insight and react quickly.
- Apache Flink is a popular open-source framework for complex data processing in real-time.
- Deep integration with many sources and sinks in Kinesis Data Analytics.
- Kinesis Data Analytics is a serverless service for real-time data analytics with Apache Flink.
- With Kinesis Data Analytics Studio, you can build applications interactively in SQL, Scala, and Python using Apache Zeppelin notebook.

Additional resources

Learn

https://aws.amazon.com/blogs/big-data/

Workshop

https://streaming-analytics.workshop.aws/flink-on-kda/

Demo – try yourself!!

https://github.com/awsmasudur/real-time-sentiment-flinkSQL-KDAStudio

Visit the AWS Data Resource Hub

Dive deeper with these resources, get inspired and learn how you can use data to make better decisions and innovate faster.

- Building a winning data strategy
- The new leadership mindset for data & analytics
- Harness data to reinvent your organization
- Put your data to work with a modern analytics approach
- Breaking free from on-premises database constraints
- Cloud storage adoption: From cost optimization to agility & innovation
- A strategic playbook for data, analytics, and machine learning
- … and more!

https://tinyurl.com/aws-data-resource

Visit resource hub

AWS Training and Certification

Empower your teams with comprehensive training

By building skills with AWS Training and Certification, businesses and individuals can see the bigger picture understanding the reasoning behind every data point. As training progresses and teams become data-fluent, previously hidden insights come into view.

Leverage free digital training

Learn how to harness the world's most valuable resource: data. Access digital and virtual instructor-led courses on data analytics and databases built by the experts at AWS and start your learning journey to become data-driven.

Take a digital course »

Get certified

Earn industry-recognized credibility and set tangible goals for success with industry-recognized certifications, like *AWS Certified Data Analytics – Specialty.*

Learn more »

Ramp-up your skills

Deep dive into new topics and focus on knowledge gaps at your own pace with the *AWS Ramp-Up Guide: Database* and *AWS Ramp-Up Guide: Data Analytics*. With a wide range of whitepapers, blog posts, videos, webinars and peer resources available for data professionals to leverage for independent learning.

Download ramp-up guides »

Thank you for attending AWS Innovate – Data Edition

We hope you found it interesting! A kind reminder to **complete the survey.**
Let us know what you thought of today's event and how we can improve the event experience for you in the future.


aws-apj-marketing@amazon.com


twitter.com/AWSCloud


facebook.com/AmazonWebServices


youtube.com/user/AmazonWebServices


slideshare.net/AmazonWebServices


twitch.tv/aws

Thank you!

https://www.linkedin.com/in/awssayem/