aws INNOVATE

DATA EDITION

19 August 2021

Analytics services optimized for your use case

Managed analytics on AWS Analytics

Francis McGregor-Macdonald

Principal Analytics Specialist Solutions Architect
Amazon Web Services

Agenda

- AWS Lake House architecture
- Why managed AWS analytics services
- Big data
- Operational analytics
- Real-time insights
- Event streams

Customers want more value from their data

Growing Exponentially

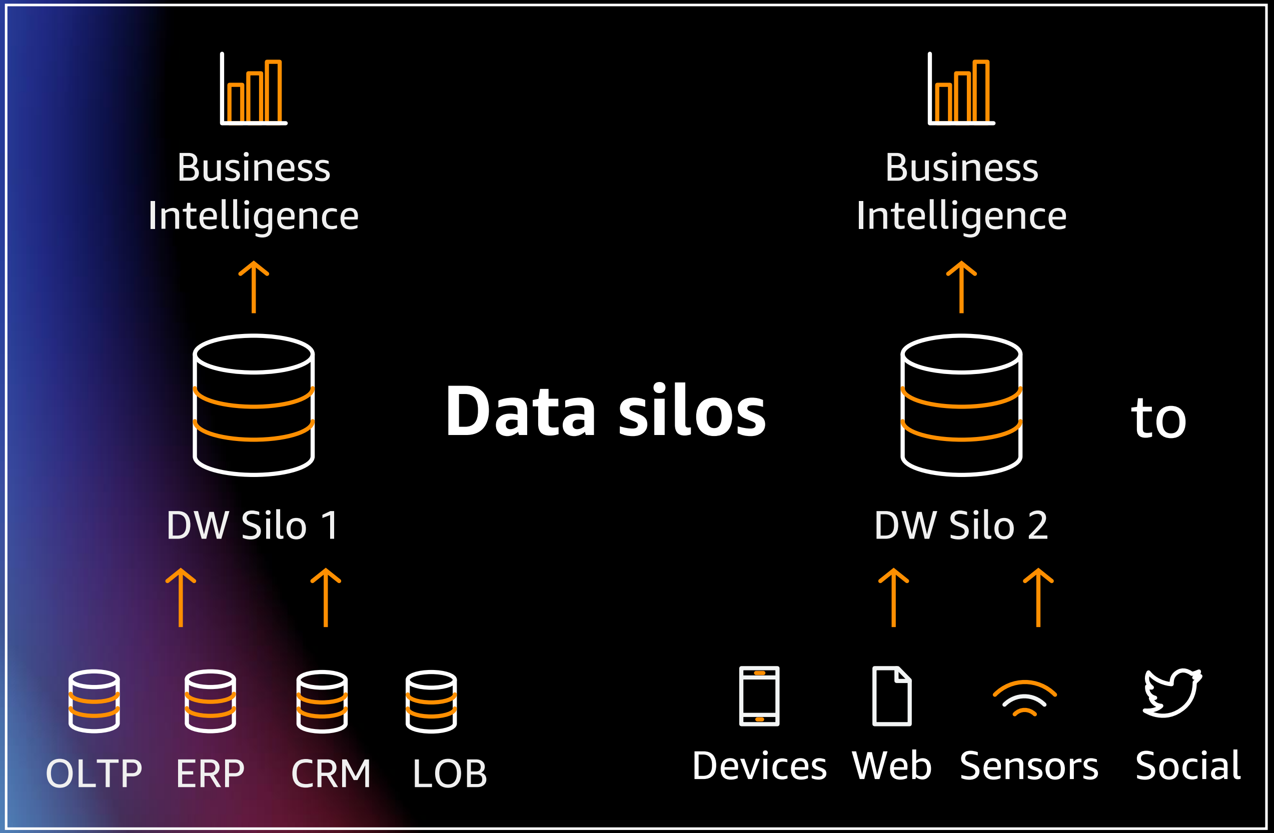
From new sources

Increasingly diverse

Used by many people

Analyzed by many applications

Customers moving from traditional silo approach to Lake House architecture



Business Intelligence

Business Intelligence

Data silos

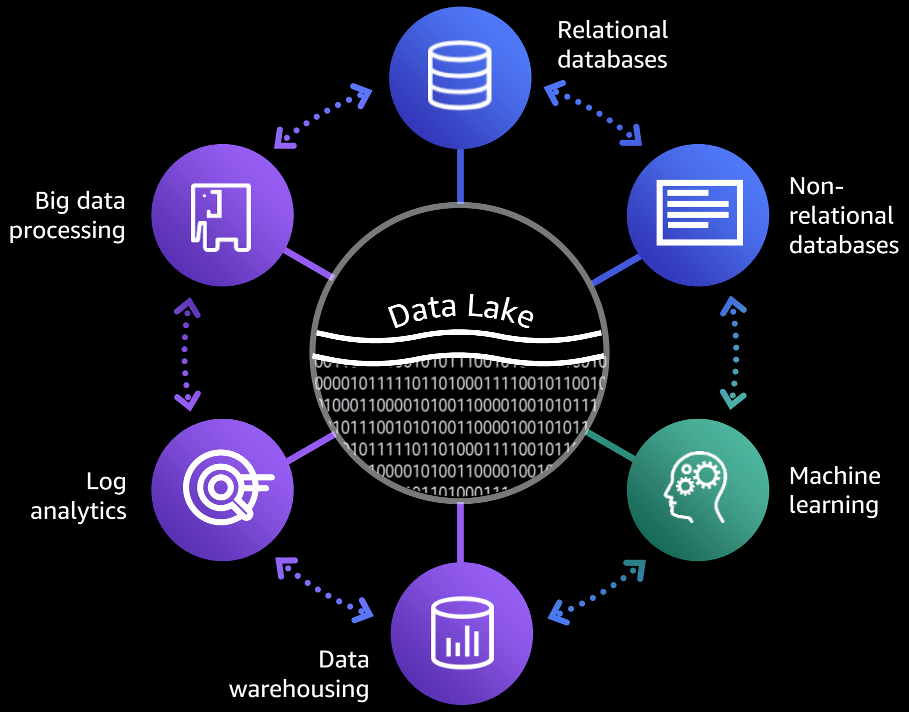
DW Silo 1

DW Silo 2

to

OLTP ERP CRM LOB

Devices Web Sensors Social



Relational databases

Non-relational databases

Machine learning

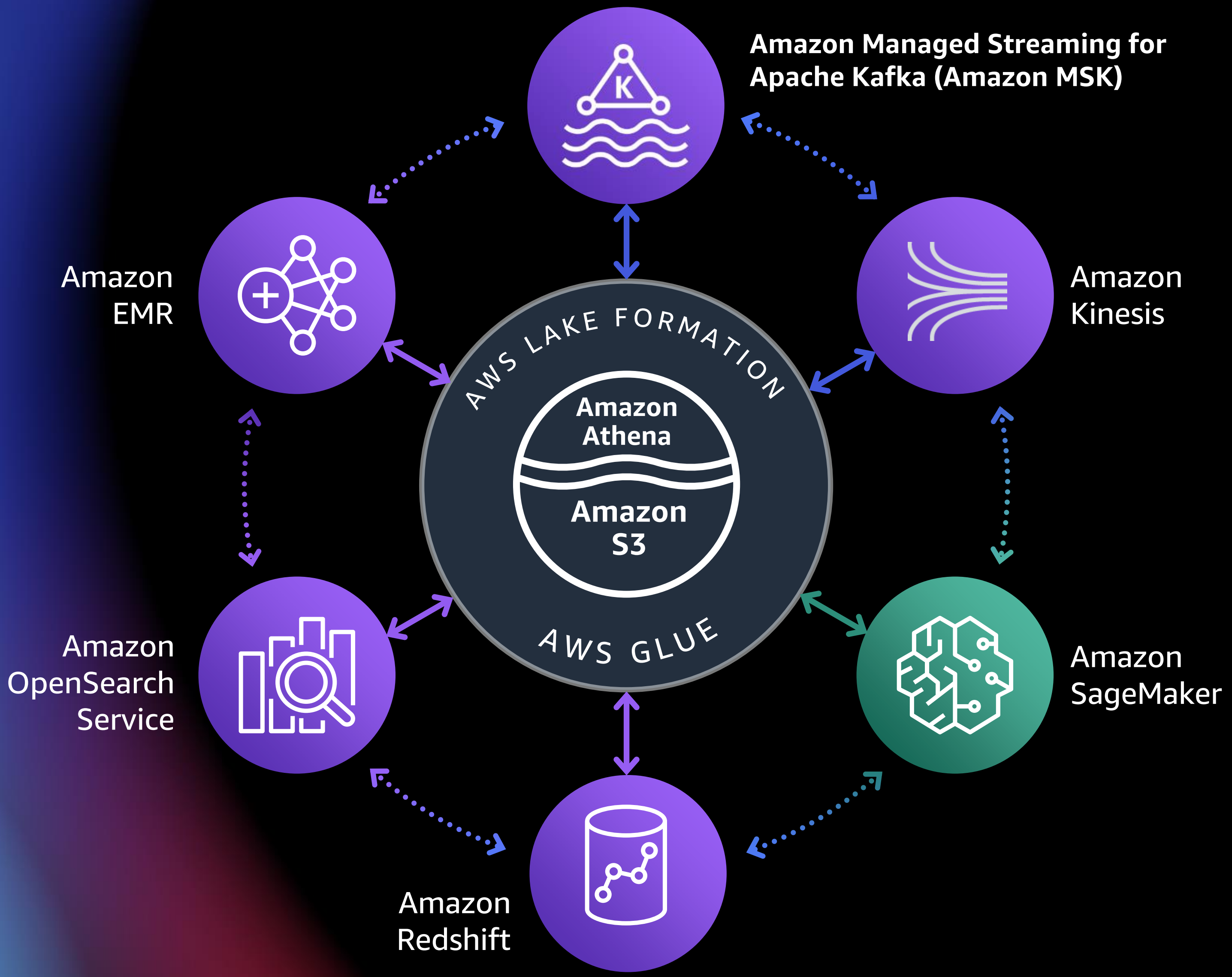
Data warehousing

Log analytics

Big data processing

Data Lake

Lake House architecture on AWS



Scalable data lakes

Purpose-built data services

Seamless data movement

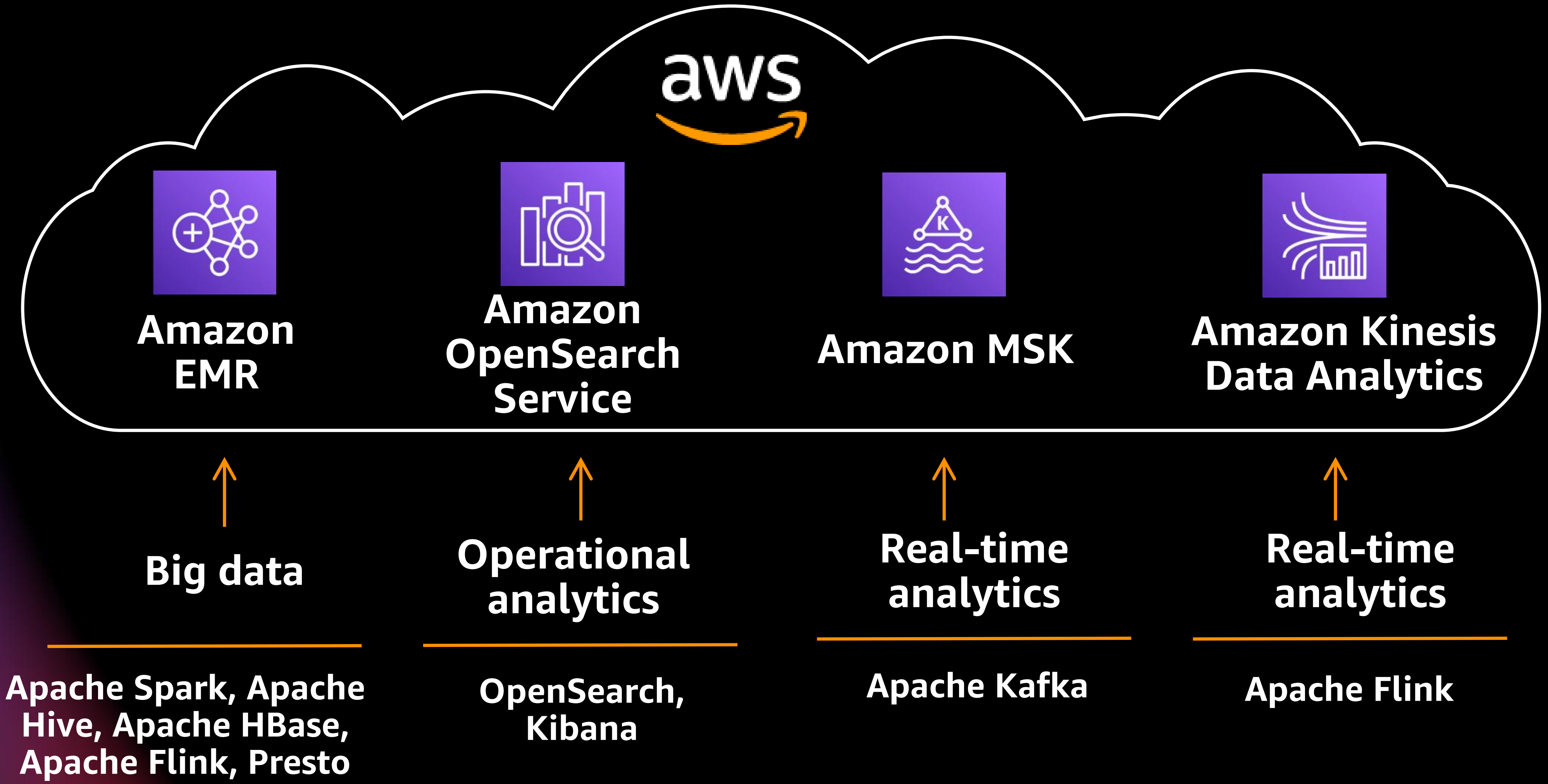
Unified governance

Performant and cost-effective

Self-managing open source analytics services is time consuming, complex, and expensive

- Hardware and software installation, configuration, patching, backups
- Capacity planning and scaling clusters for compute and storage
- Performance, throughput, latency, and high availability issues
- Security and compliance

Focus on the use case with managed analytics on AWS



Amazon EMR

Big data

Apache Spark, Apache Hive, Apache HBase, Apache Flink, Presto

Amazon OpenSearch Service

Operational analytics

OpenSearch, Kibana

Amazon MSK

Real-time analytics

Apache Kafka

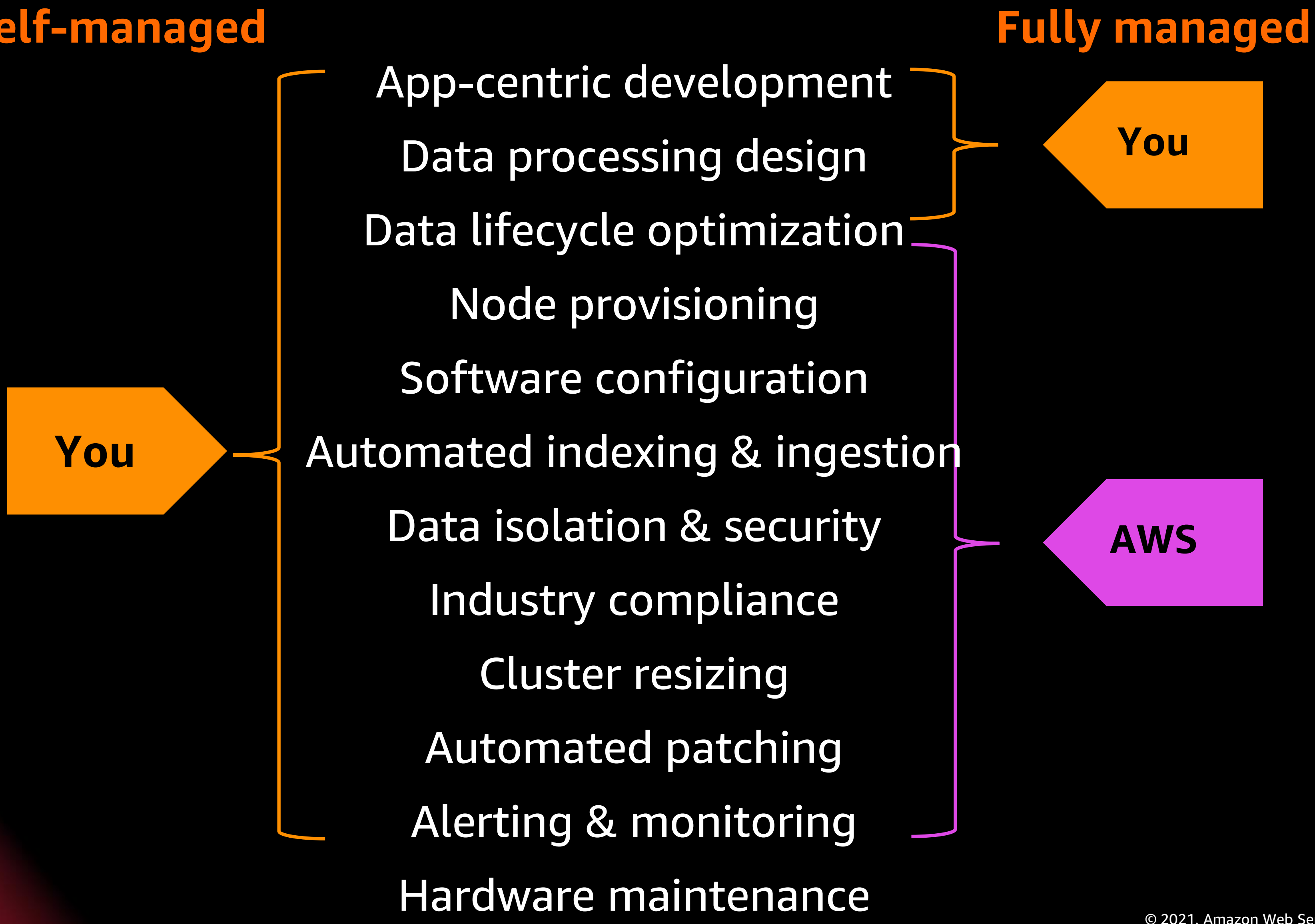
Amazon Kinesis Data Analytics

Real-time analytics

Apache Flink

Fully managed services on AWS analytics

SPEND TIME INNOVATING AND BUILDING NEW APPLICATIONS, NOT MANAGING INFRASTRUCTURE



Self-managed

Fully managed

App-centric development

Data processing design

Data lifecycle optimization

Node provisioning

Software configuration

Automated indexing & ingestion

Data isolation & security

Industry compliance

Cluster resizing

Automated patching

Alerting & monitoring

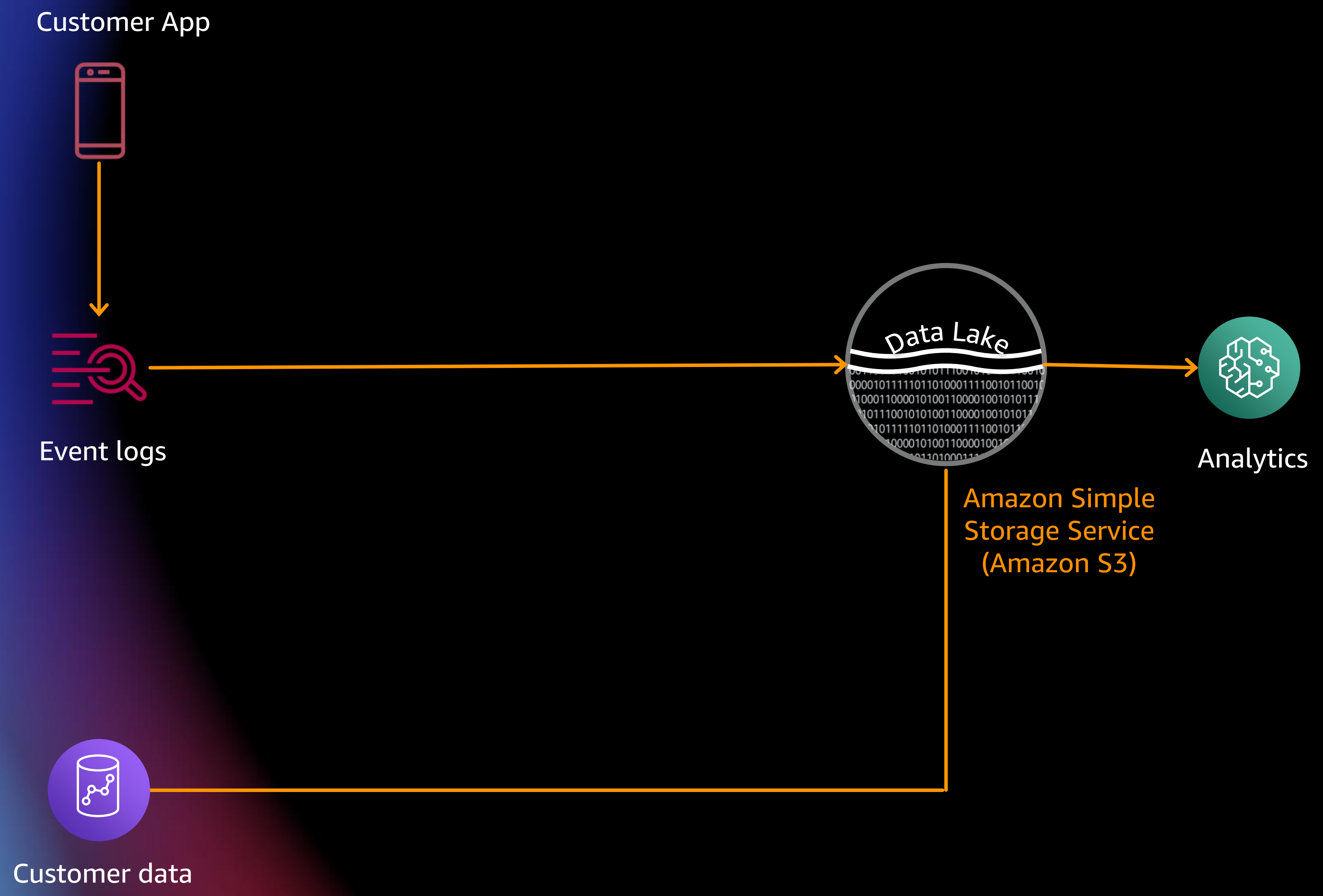
Hardware maintenance

You

You

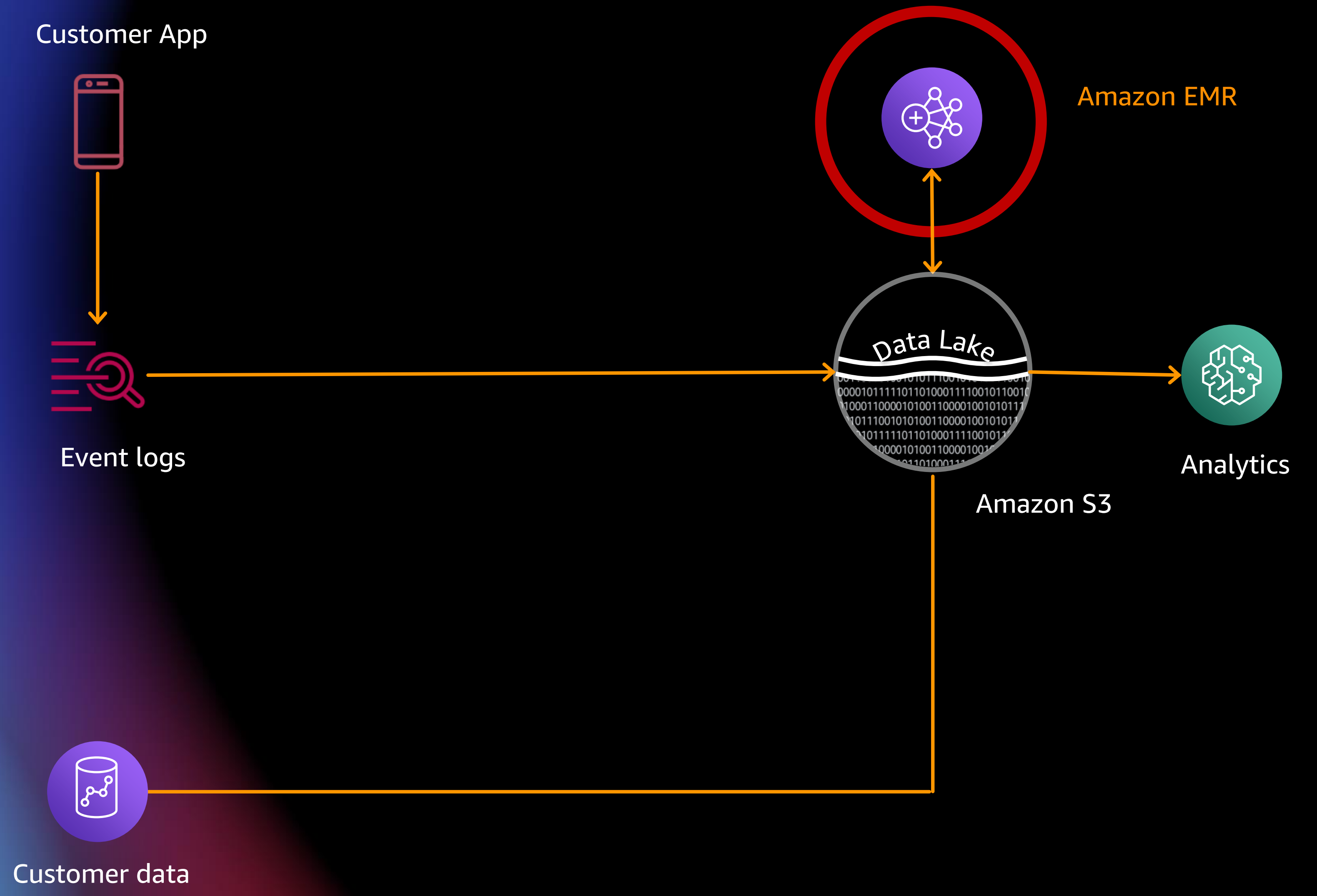
AWS

New customer app



1. Amazon S3 stores app logs and customer records

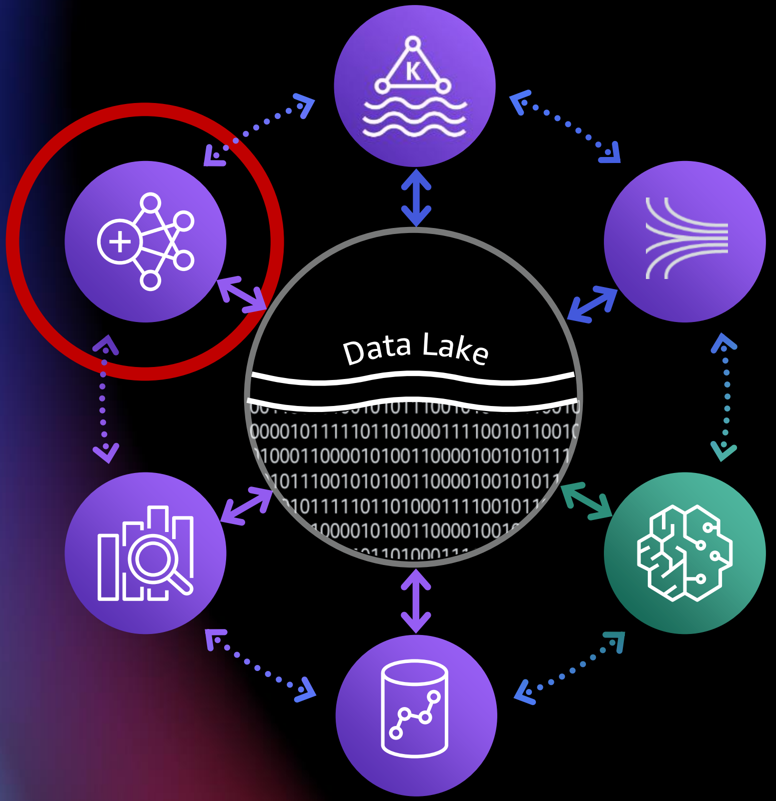
Personalized customer app



1. Amazon S3 stores app logs and customer records
2. Apache Spark on Amazon EMR creates Customer 360 insights

Amazon EMR

EASILY RUN APACE SPARK, HADOOP, HIVE, HBASE, PRESTO, AND OTHER BIG DATA FRAMEWORKS



https://aws.amazon.com/big-data/datalakes-and-analytics/
https://aws.amazon.com/emr/

Automate provisioning, configuring, and tuning
Easy setup, management, and monitoring, with latest open-source framework updates within 30 days

Run workloads faster and more cost-effectively
1.7x faster than standard Apache Spark 3.0 at 40% of the cost, and 2.6x faster than open-source Presto 0.238 at 80% of the cost

Automatically scale up and down
Manage cluster size based on utilization to reduce costs

Simple and predictable pricing
Per-second pricing, and save 50%–80% with Amazon EC2 Spot and Reserved Instances

Amazon EMR flexible deployment options

Amazon EMR on Amazon EC2

AWS offers more instance options than any other cloud provider. Choose the instance that gives the best performance or cost for your workload, including Graviton2. Take advantage of on-demand, reserved, and spot instances to optimize costs.

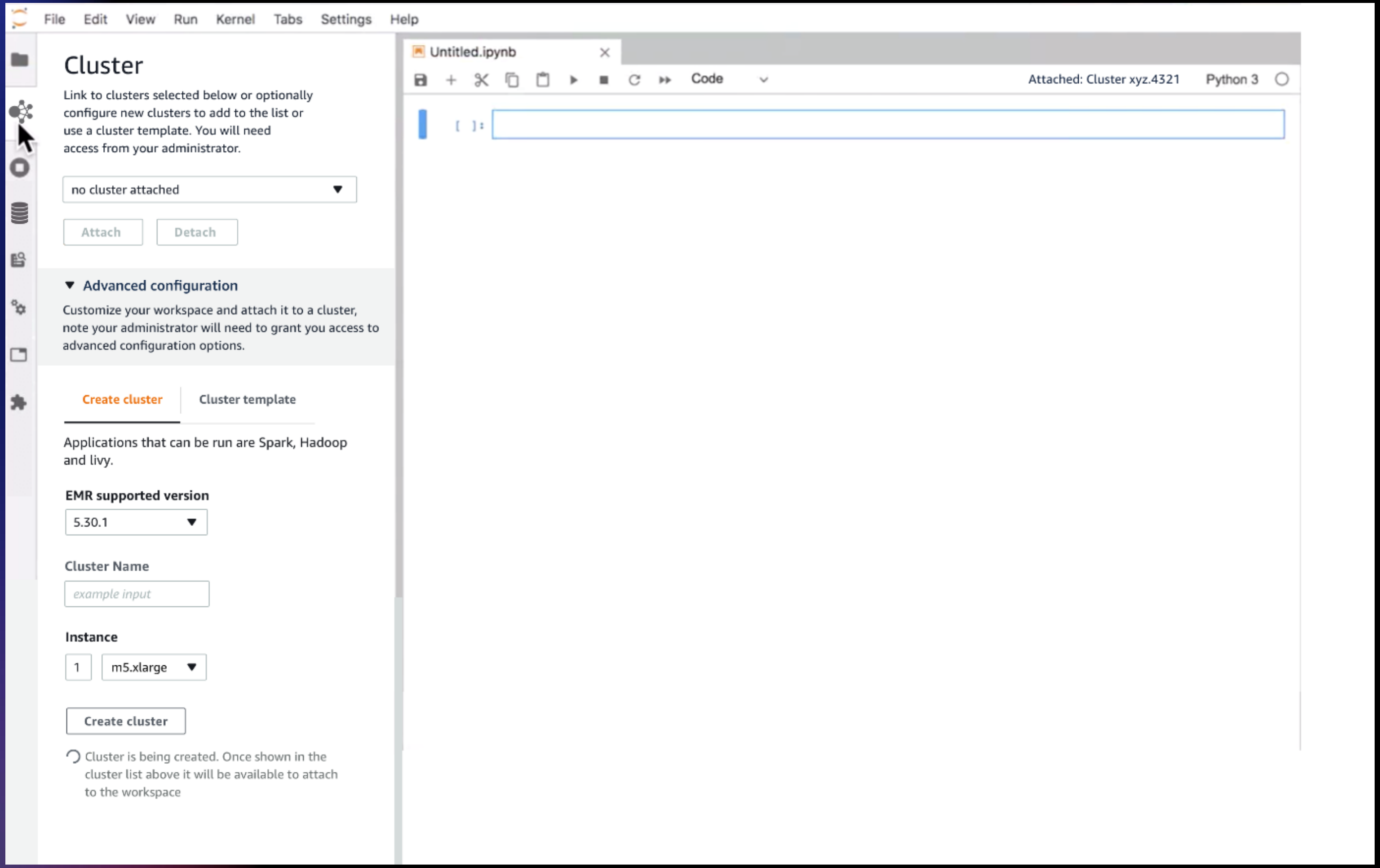
Amazon EMR on Amazon EKS

Use Amazon EMR to automate the provisioning, management, and scaling of Apache Spark jobs on Amazon Elastic Kubernetes Service (Amazon EKS), and take advantage of the optimized Amazon EMR runtime.

Amazon EMR on AWS Outposts

Set up, deploy, manage, and scale Amazon EMR in your on-premises environments, just as you would in the cloud. AWS Outposts brings AWS services, infrastructure, and operating models to virtually any data center, co-location space, or on-premises facility.

EMR Studio – Simplify Running Jobs



Cluster

Link to clusters selected below or optionally configure new clusters to add to the list or use a cluster template. You will need access from your administrator.

no cluster attached

Advanced configuration

Customize your workspace and attach it to a cluster, note your administrator will need to grant you access to advanced configuration options.

Applications that can be run are Spark, Hadoop and livy.

EMR supported version: 5.30.1

Cluster Name: example input

Instance: 1 m5.xlarge

Cluster is being created. Once shown in the cluster list above it will be available to attach to the workspace

Easily build and deploy data science code without logging in to AWS Management Console

Start notebooks in seconds, run jobs later

Build production pipelines simply and flexibly

Save debugging time with native application UIs in one place

FINRA

Challenge

FINRA’s legacy system was not able to scale to handle 150 billion events per day. They needed to run complex surveillance queries over 20+ PB of data to detect and analyze illegal market activity.

Solution

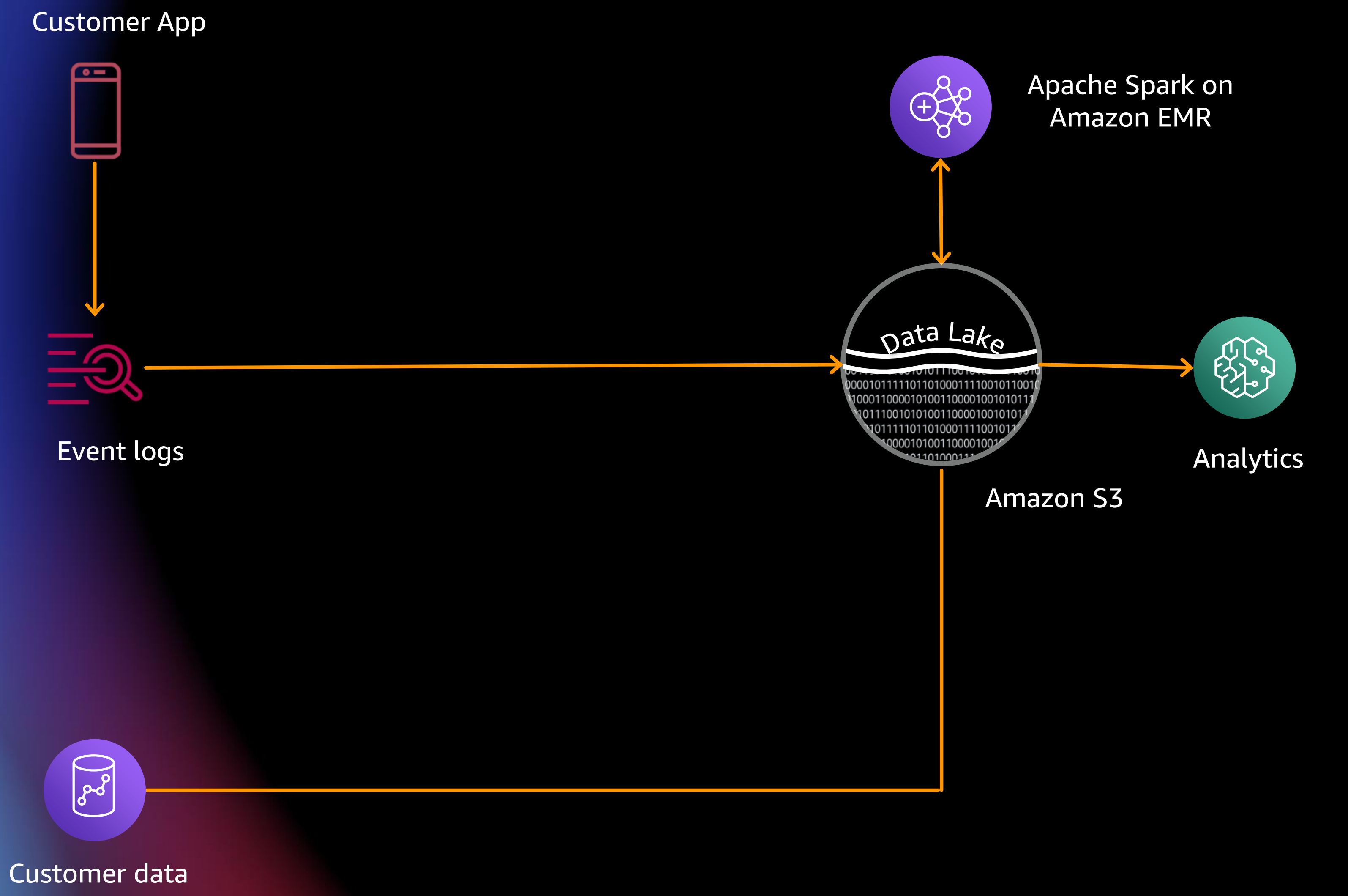
FINRA migrated their big data appliance to an Amazon S3 data lake and use AWS Lambda and Amazon EMR for data ingestion and Amazon EMR and Amazon Redshift for data processing.

Result

FINRA has been able to increase agility, speed, and cost savings while allowing them to operate at scale. The company estimates it will save $10–$20 million annually by using AWS.

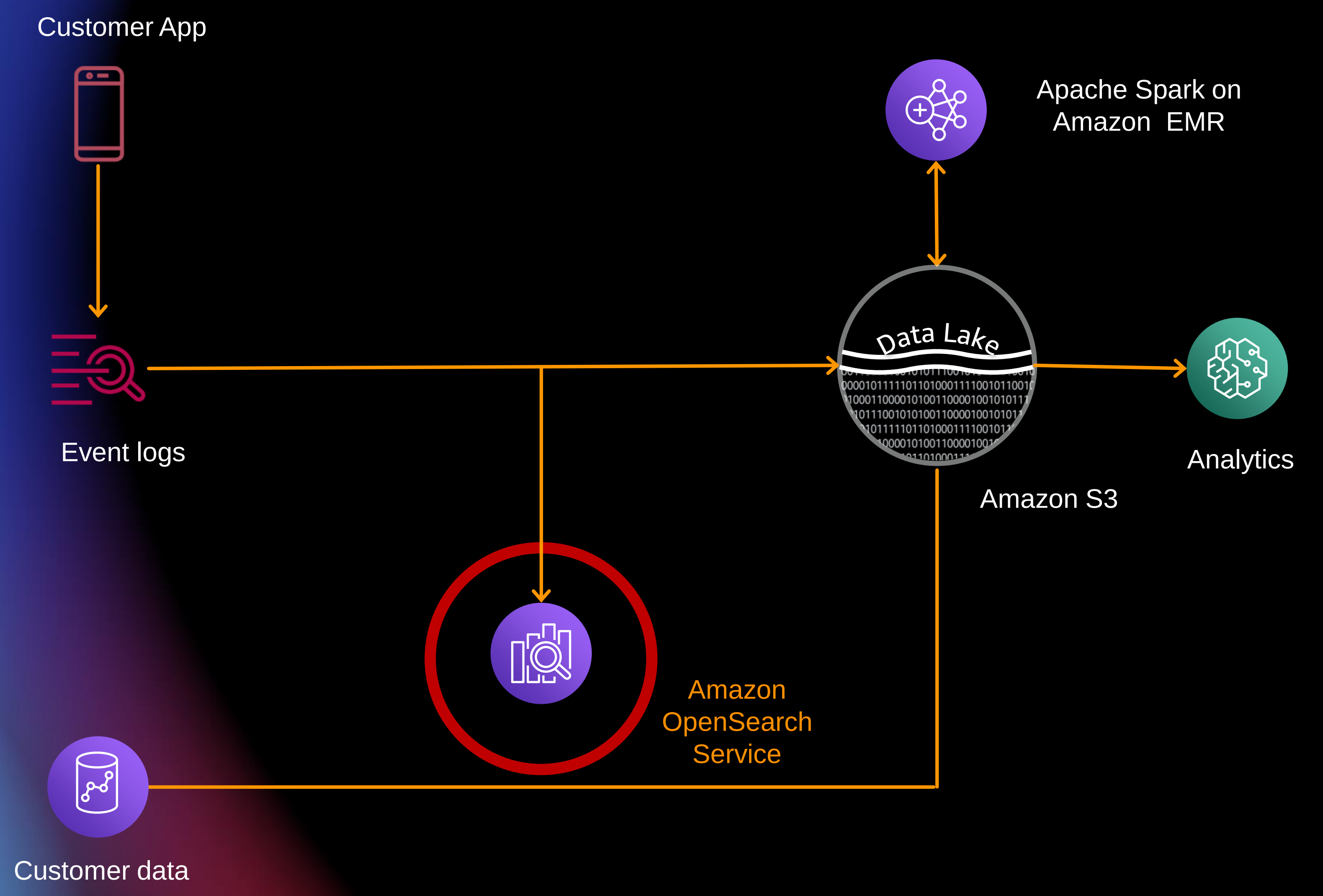
https://aws.amazon.com/solutions/case-studies/finra-data-validation/

Personalized customer app



1. Amazon S3 stores app logs and customer records
2. Apache Spark on Amazon EMR creates Customer 360 insights

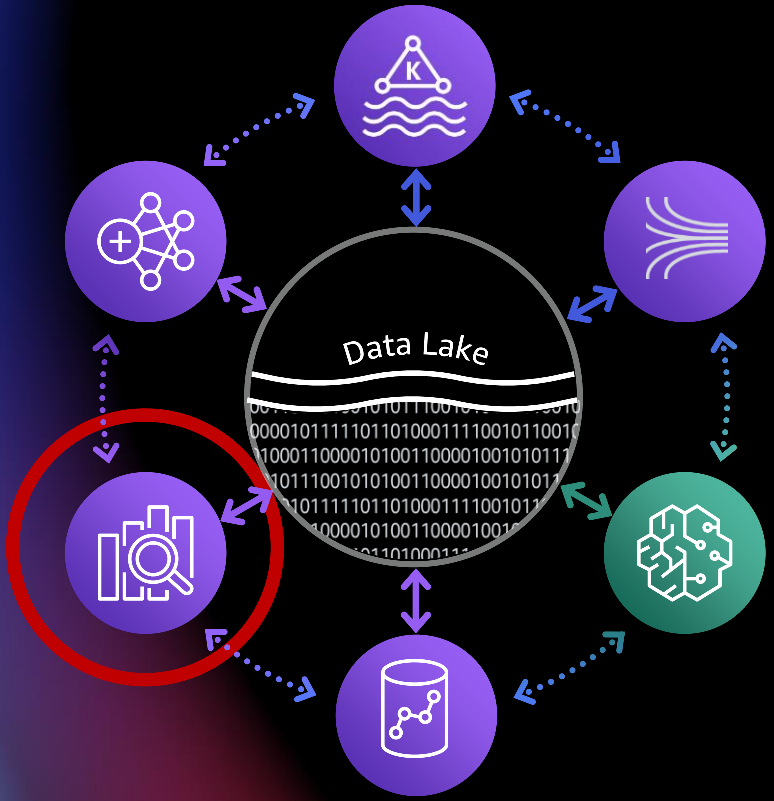
Personalized performance enhanced customer app



1. Amazon S3 stores app logs and customer records
2. Apache Spark on Amazon EMR creates Customer 360 insights
3. Amazon OpenSearch Service receives logs for SRE

Amazon OpenSearch Service

FULLY MANAGED, SCALABLE, AND SECURE OPENSEARCH SERVICE



https://aws.amazon.com/about-aws/whats-new/2020/05/aws-announces-amazon-elasticsearch-service-ultrawarm-general-availability/

Easy integration
Open source OpensSearch APIs, managed Kibana, integration with Logstash

Fully managed
Deployment in minutes, software installation and patching, failure recovery, backups, and monitoring

Scalable, secure, and compliant
Network isolation with Amazon VPC, encryption at-rest and in transit, and compliant with HIPPA PCI DSS, and ISO

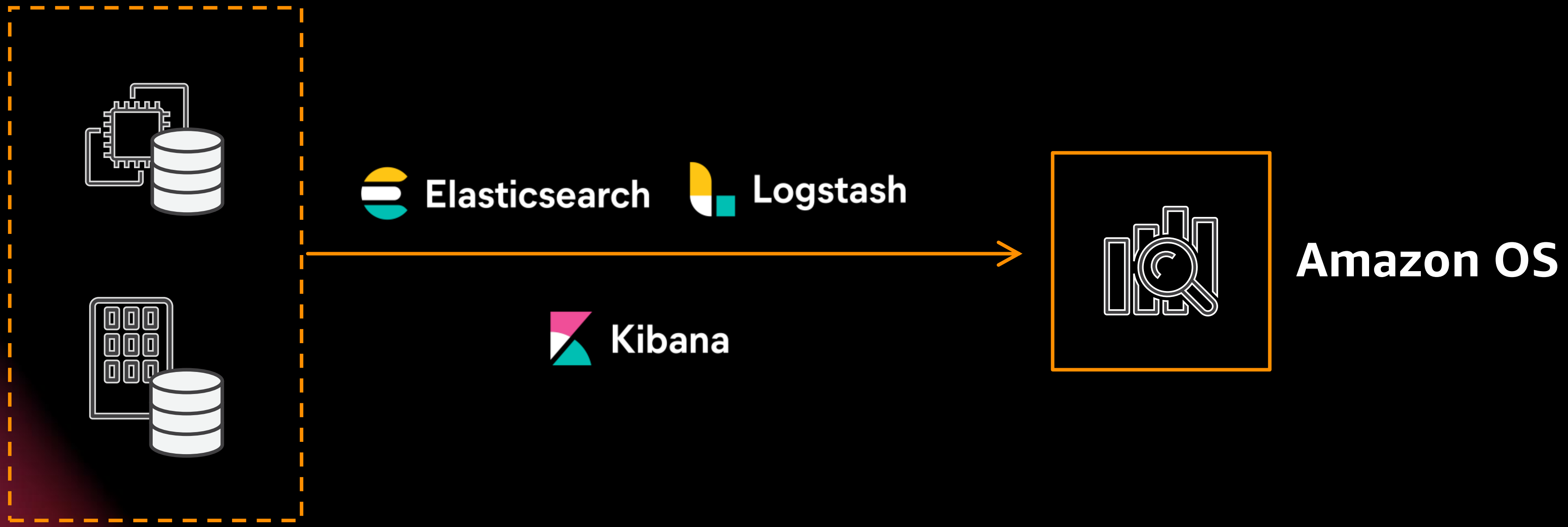
Cost-effective
Pay only for resources used with choice of on-demand and Reserved Instance compute pricing, and save up to 90% with Ultrawarm low-cost storage tier

Move to managed operational analytics

MIGRATE ON-PREMISES OR SELF-MANAGED OPENSEARCH TO A FULLY MANAGED SERVICE

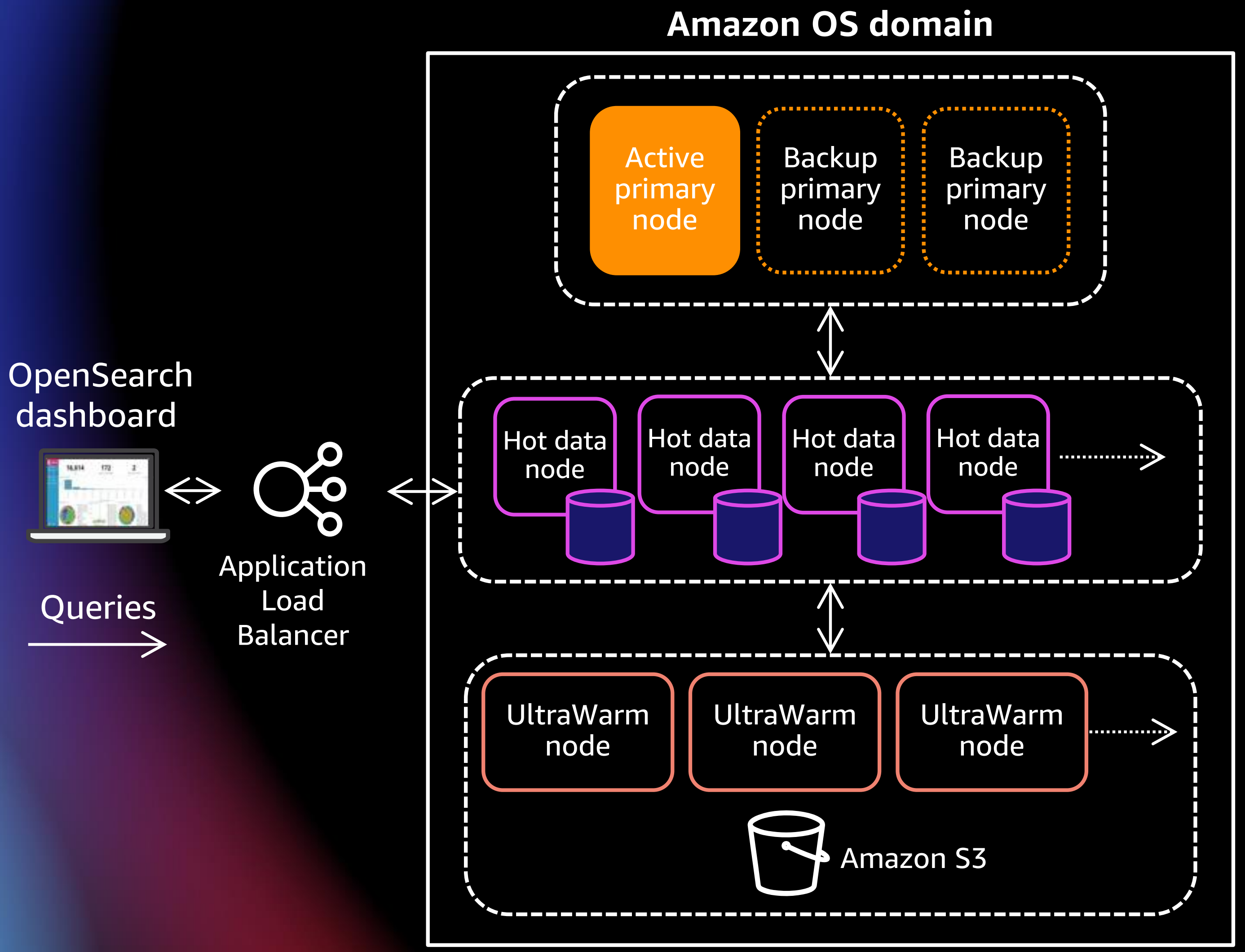
Typical challenges

1. A team of OpenSearch experts is needed to support provisioning, operations, software patches and updates, 24/7 monitoring for hardware and application failures, and resizing clusters – all driving up the total cost of ownership
2. Customers need to build or pay for advanced features such as security, encryption, alerting, anomaly detection, index lifecycle automation, and more



UltraWarm for Amazon OpenSearch Service

A NEW WARM STORAGE TIER FOR AMAZON OS



90% lower cost

Scale up to 3 PB per domain

Analyze years of operational data

Interactive log analytics and visualization

https://aws.amazon.com/about-aws/whats-new/2020/05/aws-announces-amazon-elasticsearch-service-ultrawarm-general-availability/

Pinterest scales daily log search and analytics to 1.7 TB and reduces costs by 30% using Amazon OpenSearch Service

Challenge:

As one of the largest visual discovery engines in the world and with 400 million monthly active users, Pinterest sought a solution to address the growing volume of data it needed to ingest for its engineers to effectively analyze log data.

Solution:

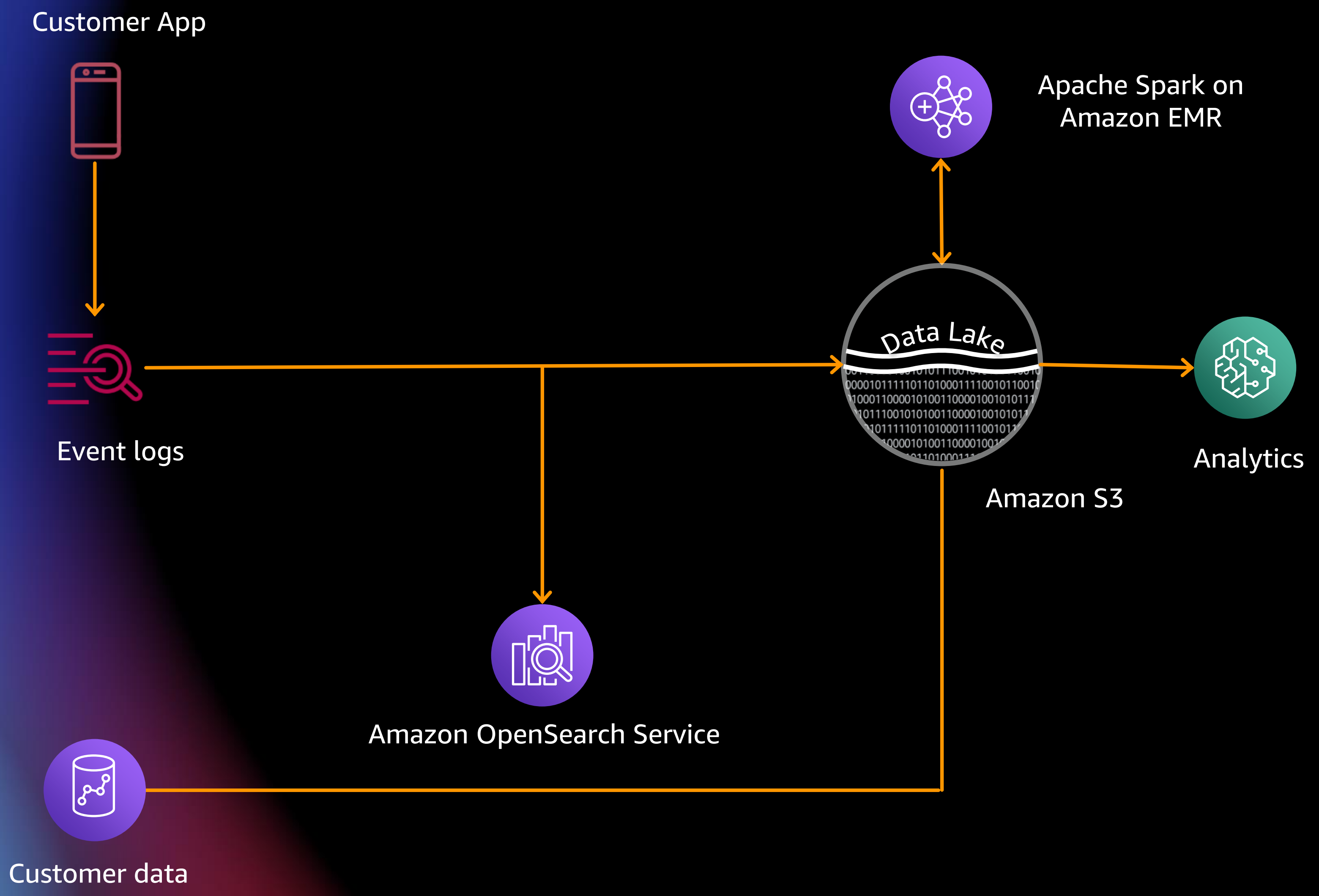
Pinterest moved to managed analytics using Amazon OpenSearch Service (Amazon OS), enabling it not only to scale its log-analysis capabilities but also to reduce operational burdens, improve security, and save costs.

Result:

- Scaled monitoring and alerting capabilities for software deployment
- Reduced operational costs by 30%, with 40–50% expected
- Increased productivity by freeing software engineers from low-value work

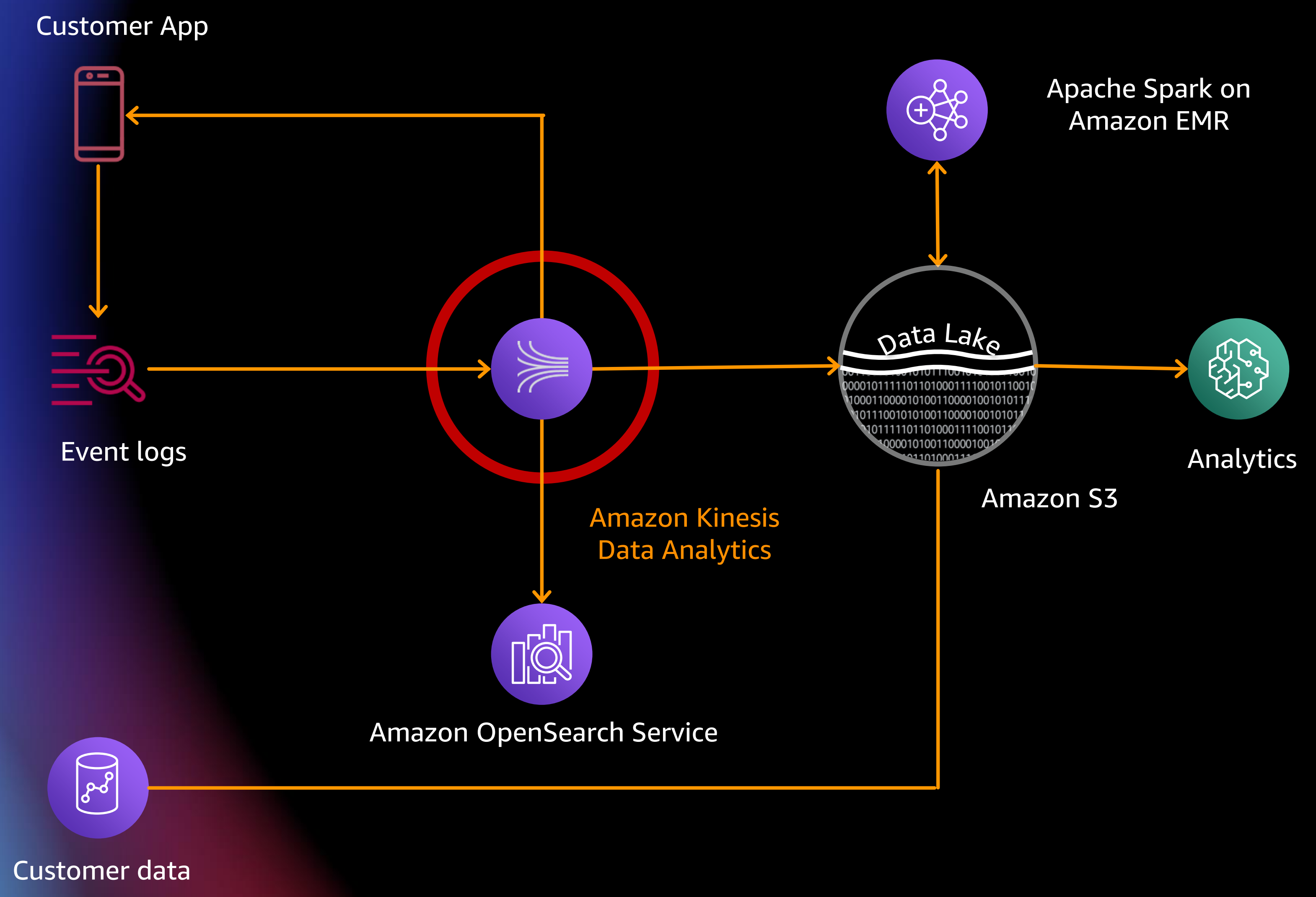
https://aws.amazon.com/solutions/case-studies/pinterest-elasticsearch-case-study/

Personalized performance enhanced customer app



1. Amazon S3 stores app logs and customer records
2. Apache Spark on Amazon EMR creates Customer 360 insights
3. Amazon OpenSearch Service receives logs for SRE

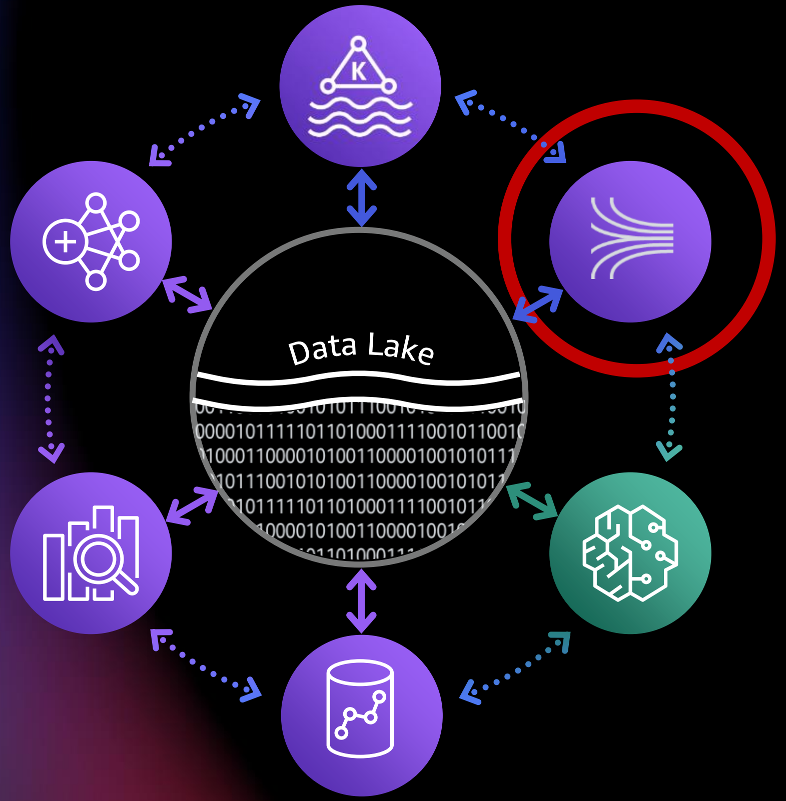
Personalized performance enhanced recommending customer app


Customer App
Event logs
Amazon Kinesis Data Analytics
Amazon OpenSearch Service
Customer data
Apache Spark on Amazon EMR
Data Lake
Amazon S3
Analytics

1. Amazon S3 stores app logs and customer records
2. Apache Spark on Amazon EMR creates Customer 360 insights
3. Amazon OpenSearch Service receives logs for SRE
4. Amazon Kinesis Data Analytics returns real-time decisions to app

Amazon Kinesis

EASILY COLLECT, PROCESS, AND ANALYZE DATA AND VIDEO STREAMS IN REAL TIME



Amazon Kinesis Data Analytics

Analyze data streams with serverless Apache Flink

Amazon Kinesis Data Streams

Capture, process, and store data streams

Amazon Kinesis Data Firehose

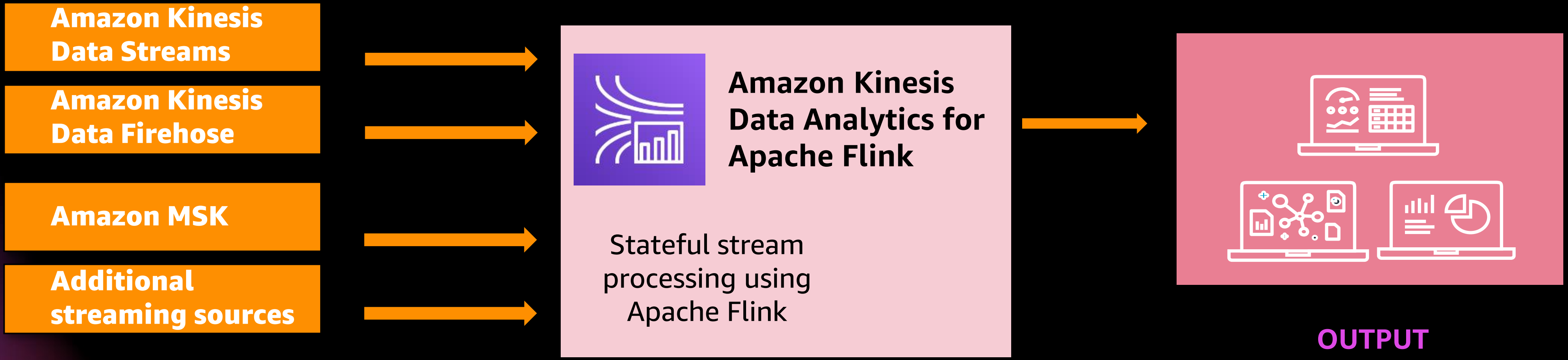
Load data streams into AWS data stores

Move to managed real-time analytics

MIGRATE APACHE FLINK APPLICATIONS TO A FULLY MANAGED SERVICE

Typical challenges

1. Apache Flink is an open-source framework and engine for processing data streams
2. Building, managing, and integrating streaming applications is complex, and streaming data flows at an incredible rate that can vary up and down all the time – Streaming analytics services need to process this data when it arrives, often at speeds of millions of events per hour

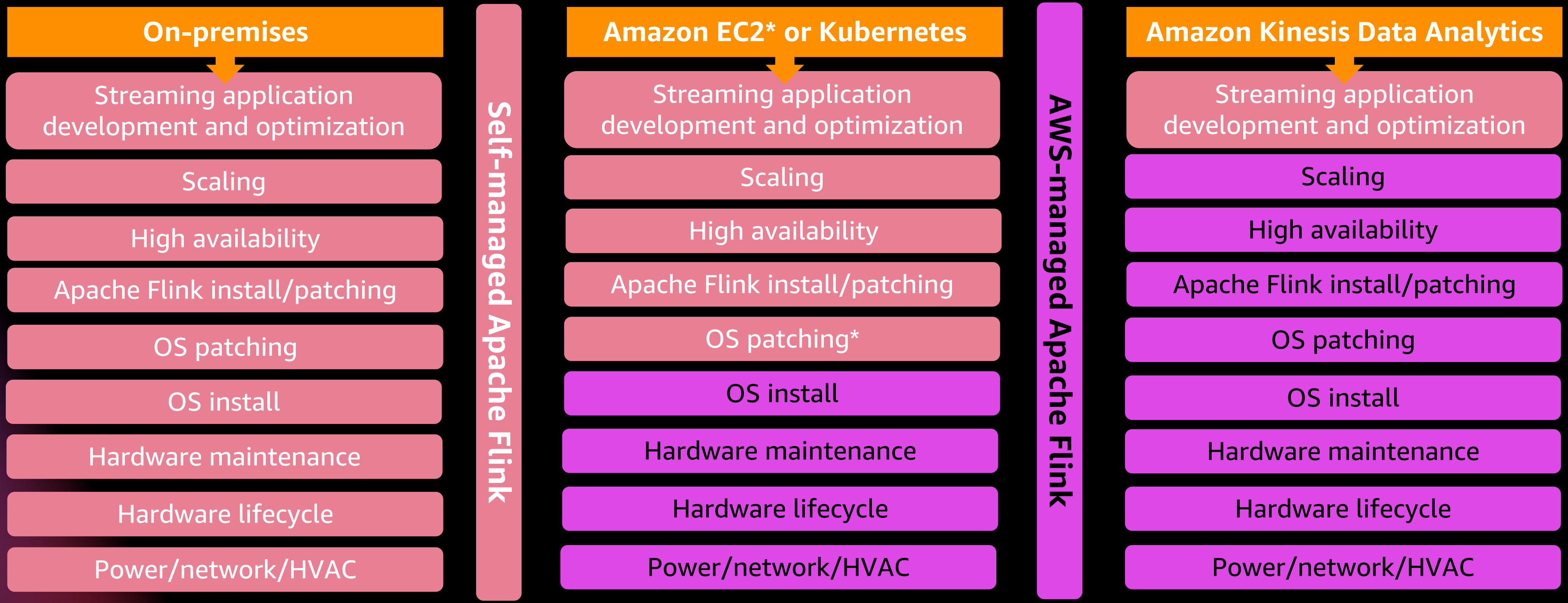

Amazon Kinesis Data Streams
Amazon Kinesis Data Firehose
Amazon MSK
Additional streaming sources
Amazon Kinesis Data Analytics for Apache Flink
Stateful stream processing using Apache Flink

INPUT
Captured streaming data

OUTPUT
Amazon Kinesis Data Analytics can send processed data to analytics tools so you can create alerts and respond in real time

Amazon Kinesis Data Analytics is fully managed

More focus on creating streaming applications than managing Apache Flink infrastructure



On-premises	Amazon EC2* or Kubernetes	Amazon Kinesis Data Analytics
Streaming application development and optimization	Streaming application development and optimization	Streaming application development and optimization
Scaling	Scaling	Scaling
High availability	High availability	High availability
Apache Flink install/patching	Apache Flink install/patching	Apache Flink install/patching
OS patching	OS patching*	OS patching
OS install	OS install	OS install
Hardware maintenance	Hardware maintenance	Hardware maintenance
Hardware lifecycle	Hardware lifecycle	Hardware lifecycle
Power/network/HVAC	Power/network/HVAC	Power/network/HVAC

Amazon Kinesis Data Analytics customers

"With Amazon Kinesis Data Analytics for Apache Flink, we are able help our customers build better connections and view more relevant content in real time."

Nextdoor

"Ultimately, we are improving our software products and offering better service to our customers because of the real-time visibility we're getting into log data."

Autodesk

"Using Apache Flink in Amazon Kinesis Data Analytics enables us to respond in real time to user actions across our suite of mobile and web apps."

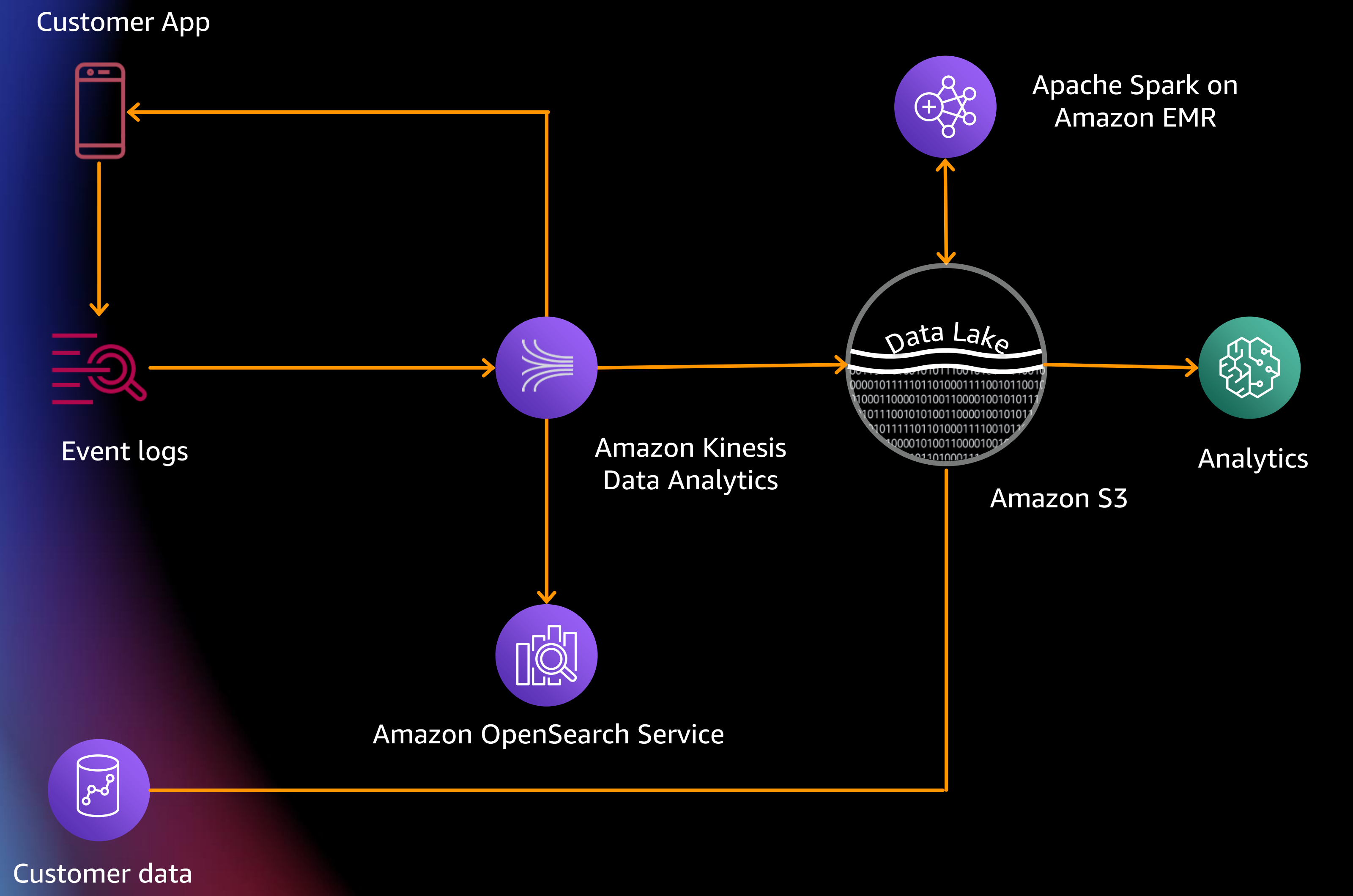
Lightricks

"Our senior management were bowled over. They've never had the ability to see what games were being played in real time, and it was all done in a few days with a small amount of code using the streaming data platform provided by Amazon Kinesis."

Palringo

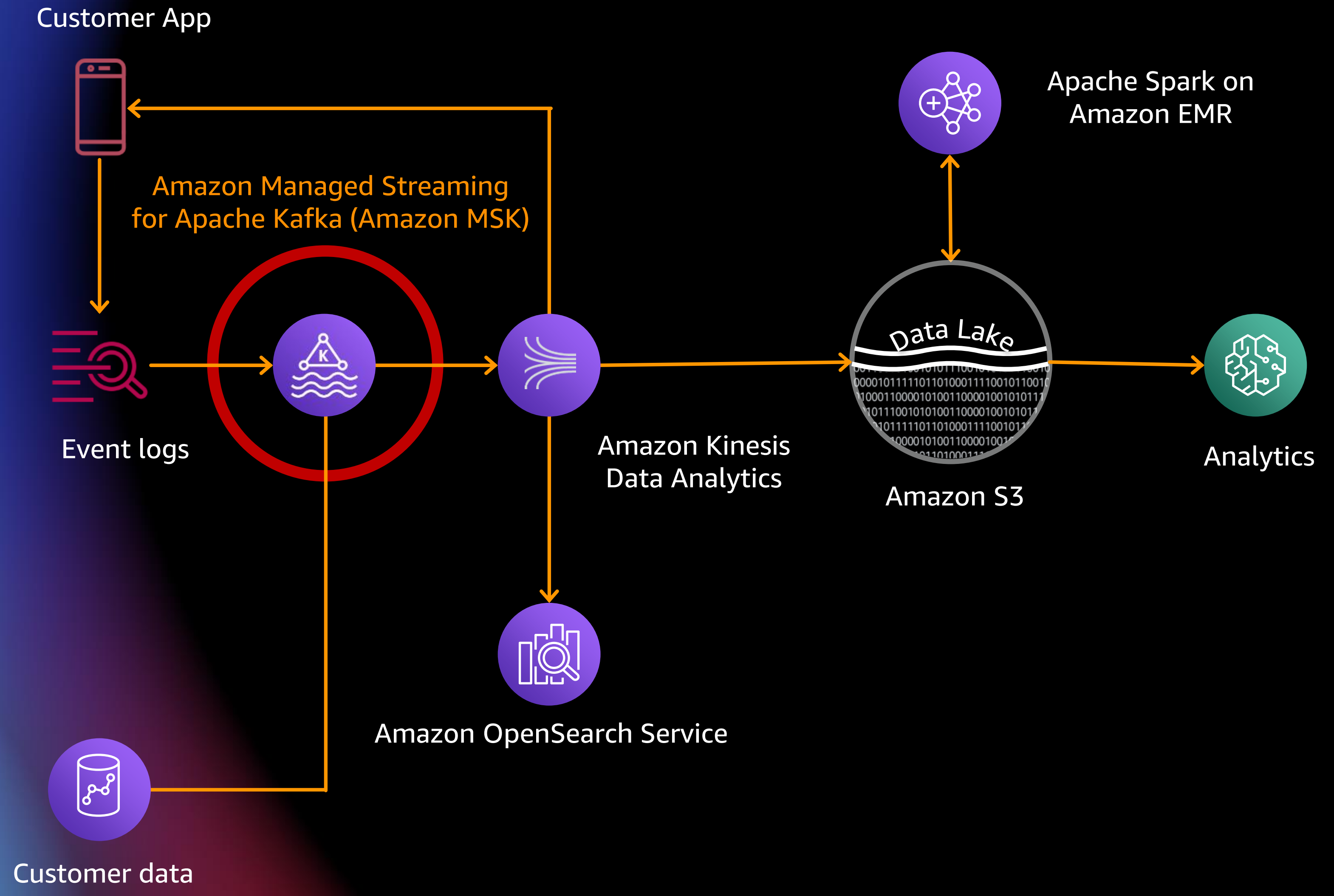
https://aws.amazon.com/kinesis/data-analytics/customers/

Personalized performance enhanced recommending customer app


Customer App
Apache Spark on Amazon EMR
Data Lake
Event logs
Amazon Kinesis Data Analytics
Amazon S3
Analytics
Amazon OpenSearch Service
Customer data

1. Amazon S3 stores app logs and customer records
2. Apache Spark on Amazon EMR creates Customer 360 insights
3. Amazon OpenSearch Service receives logs for SRE
4. Amazon Kinesis Data Analytics returns real-time decisions to app

Personalized performance enhanced recommending customer app ready for the next innovation

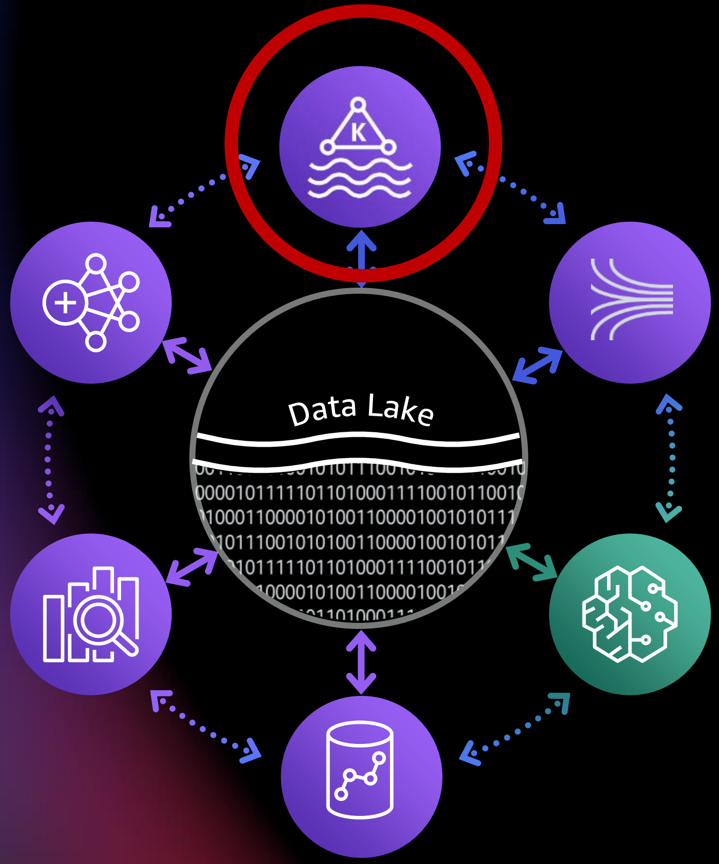


Customer App
Amazon Managed Streaming for Apache Kafka (Amazon MSK)
Event logs
Amazon Kinesis Data Analytics
Apache Spark on Amazon EMR
Data Lake
Amazon S3
Analytics
Amazon OpenSearch Service
Customer data

1. Amazon S3 stores app logs and customer records
2. Apache Spark on Amazon EMR creates Customer 360 insights
3. Amazon OpenSearch Service receives logs for SRE
4. Amazon Kinesis Data Analytics returns real-time decisions to app
5. Amazon MSK provides a standard event stream processing framework

Amazon Managed Streaming for Apache Kafka (Amazon MSK)

FULLY MANAGED, HIGHLY AVAILABLE, AND SECURE APACHE KAFKA SERVICE



Fully compatible
Run your existing Apache Kafka applications on AWS without changes to source code

Fully managed
Focus on creating applications not managing your Apache Kafka environment

Elastic stream processing
Run Apache Flink applications written in SQL, Java, or Scala that elastically scale to process data streams

Highly available
Take advantage of multi-AZ replication within an AWS region

Highly secure
Protect your data with multiple levels of security, including VPC network isolation, encryption at-rest and in-transit, and more

aws

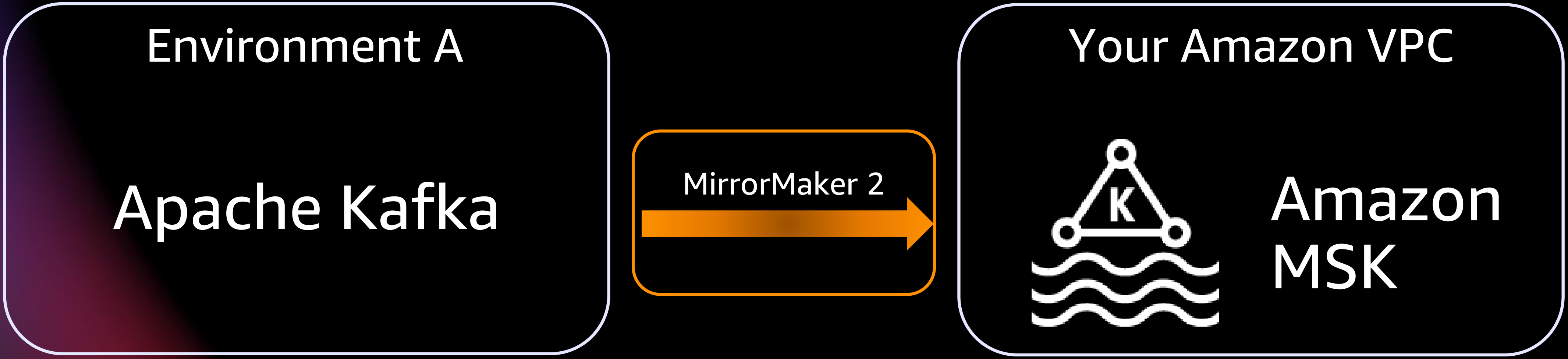
https://aws.amazon.com/msk/

Move to managed real-time analytics

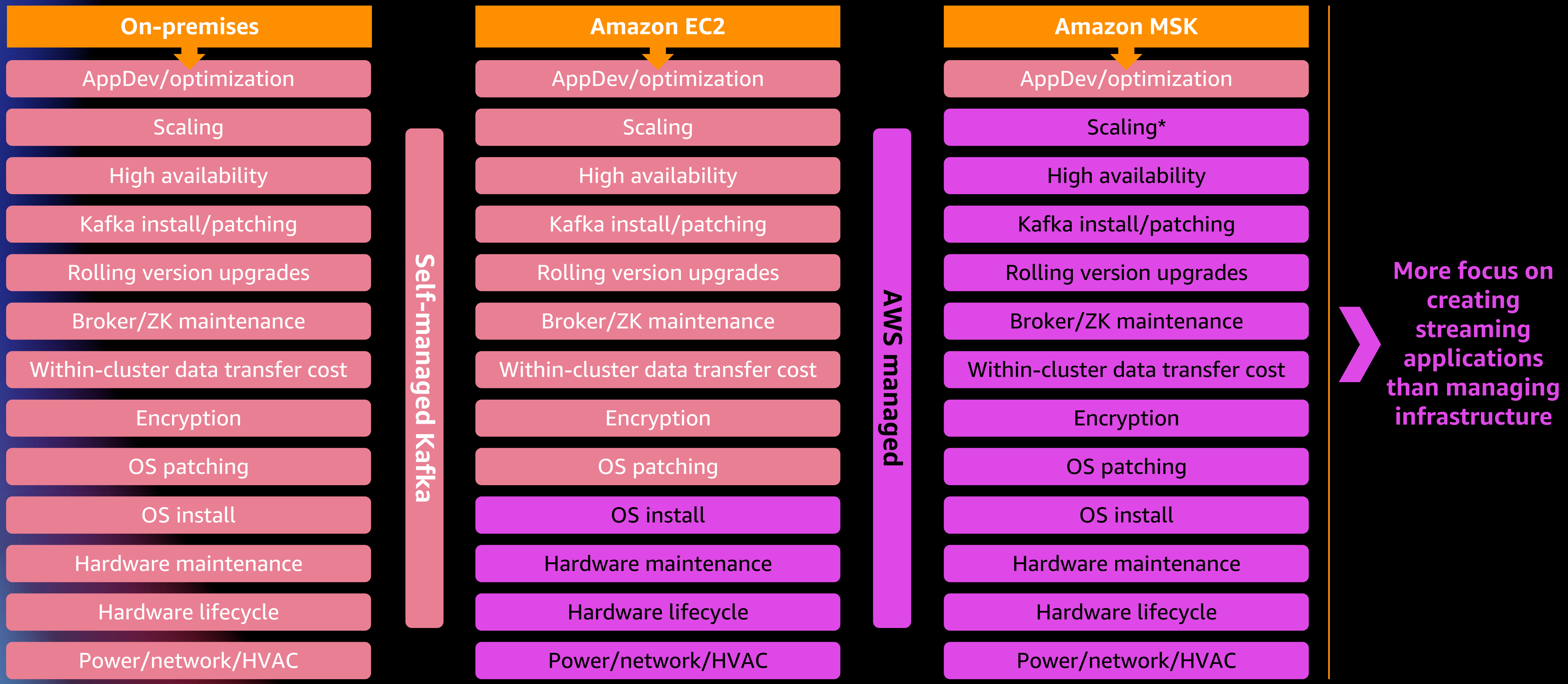
MIGRATE ON-PREMISES OR SELF-MANAGED APACHE KAFKA TO A FULLY MANAGED SERVICE

Typical challenges

1. Apache Kafka is an open-source platform for building real-time streaming data pipelines and applications
2. Apache Kafka clusters are challenging to set up, scale, and manage in production; you need to provision servers, configure Apache Kafka manually, replace servers when they fail, orchestrate server patches and upgrades, architect the cluster for high availability, ensure data is durably stored and secured, set up monitoring and alarms, and carefully plan scaling events to support load changes



Amazon MSK is fully managed



On-premises	Amazon EC2	Amazon MSK
AppDev/optimization	AppDev/optimization	AppDev/optimization
Scaling	Scaling	Scaling*
High availability	High availability	High availability
Kafka install/patching	Kafka install/patching	Kafka install/patching
Rolling version upgrades	Rolling version upgrades	Rolling version upgrades
Broker/ZK maintenance	Broker/ZK maintenance	Broker/ZK maintenance
Within-cluster data transfer cost	Within-cluster data transfer cost	Within-cluster data transfer cost
Encryption	Encryption	Encryption
OS patching	OS patching	OS patching
OS install	OS install	OS install
Hardware maintenance	Hardware maintenance	Hardware maintenance
Hardware lifecycle	Hardware lifecycle	Hardware lifecycle
Power/network/HVAC	Power/network/HVAC	Power/network/HVAC

More focus on creating streaming applications than managing infrastructure

Amazon MSK customers

"We can now confidently and frequently update our applications at scale, run complex Kafka streams topologies with ease, and debug applications instead of infrastructure."

Maksym Schipka, CTO
Vortexa

"Amazon MSK not only helped us offload infrastructure overhead, but we also maintained high throughput and performance for our business-critical metadata pipelines in a more secure and reliable manner."

Kapil Bharati, CTO, and
Akashdeep Verma, Technical Architect
Delhivery

"Amazon MSK has removed the complexity of setup and maintenance, allowing us to focus on what's most important – building innovative new capabilities for our customers."

Venkatesh Ennala, Software Engineer
Vonage

https://aws.amazon.com/msk/customer-success/

aws

Move to AWS analytics to break free from the undifferentiated heavy lifting

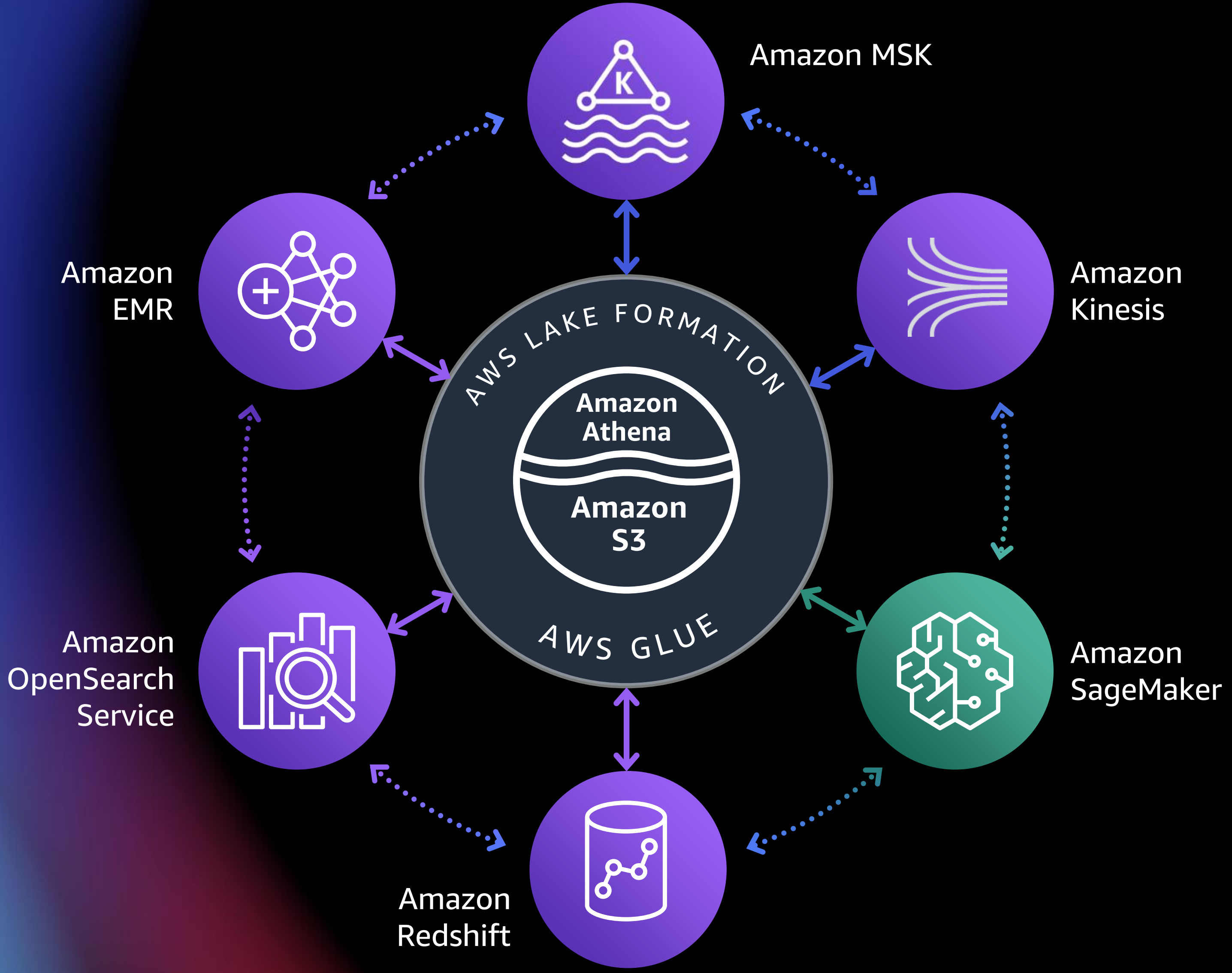
Easiest to build data lakes and analytics

Most secure infrastructure for analytics

Most comprehensive and open

Most scalable and cost effective

Lake House architecture on AWS



Scalable data lakes

Purpose-built data services

Seamless data movement

Unified governance

Performant and cost-effective

Want to build a data vision and strategy?

aws data-driven everything

- ✓ Joint engagements with business and technology stakeholder alignment
- ✓ Create an organizational vision for innovation with data to drive business outcomes
- ✓ Define the first pilot, learn, and build

Jump-start the data flywheel

Have a strategy and need help executing it?

aws data lab

- ✓ Joint engineering engagements between customers and AWS technical resources
- ✓ Create tangible deliverables to accelerate strategic databases, analytics, and ML initiatives
- ✓ Leave with an architecture, working prototype, path to production, and deeper knowledge of AWS services

Come with an idea, leave with a solution

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!