AWS re:Invent

NOV. 29 – DEC. 3, 2021 | LAS VEGAS, NV

accenture

DAT216-S

SPONSORED BY ACCENTURE

Jumping into an AWS-native data lake

Jonathan Steinert
Associate Director
Accenture

Daniel Grube
Senior Manager
Accenture

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Market situation

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Digital Disruption

Organizations need to harness the power & speed of data . . .

or get left behind



1.7MB of data will be created every second for 7.5B people on the planet[1]

The amount of data coming into organizations has increased by **25%** Every year over the last five years[2]

[1] Source: DOMO, Data Never Sleeps Report
[2] Aberdeen, Big Data Pro-essentials, bit.ly/3Do6AI1

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

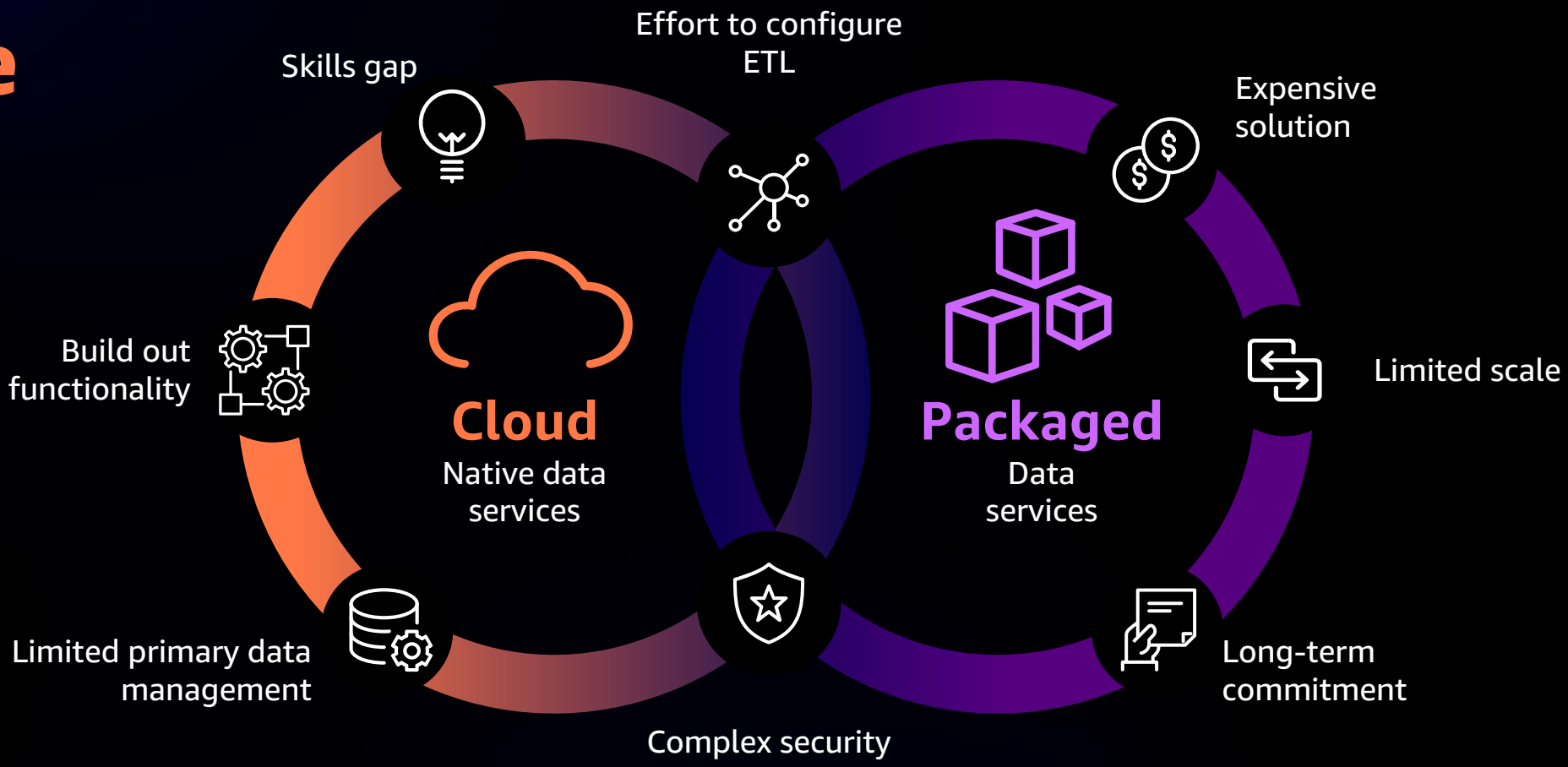
The global data lake market is projected to reach **$20.1B** by 2024[1]

With data lakes in place, best-in-class organizations see improvements in their bottom line with **71%** Reducing IT costs for storage and data management[2]

[1] MarketsandMarkets, "Data Lake Market by Component," bit.ly/30xOBR7
[2] Aberdeen, Big Data Pro-essentials, bit.ly/3Do6AI1

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

The data lake challenge



Results

Long time to value creation

Unconsumed data

Suboptimal value realization

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

The Data Lake Accelerator

Accenture's cloud-native Data Lake Accelerator provides a cost-effective, scalable solution that streamlines, automates, and expedites the build and delivery of enterprise-class data lakes on AWS

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AABG's AWS Cloud-native

data lake asset



Automated zone creation based on best practices

4 zones (landing, raw, curated, transformed & provisioned) are created; scripts are run that create the **Amazon S3 buckets** and associated **folder structures** automatically in minutes



Scheduled data crawling

AWS Glue crawlers are scheduled to automatically crawl the data source folders and capture metadata to store them in **AWS Glue Data Catalog**; **t**his has been configured to automatically capture metadata for new sources without any manual intervention



Parameterized framework

The entire creation of data lake is parameterized using **CFTs and AWS Lambda functions** (i.e., users just have to enter the details either in a **Microsoft Excel or Visio** file, and the data lake is created as per their needs)



Reduced deployment lifecycle

Resource provisioning is completely automated using **CFTs and AWS Lambda functions**; **DataOps cartridge** for data lake through **AWS CodeCommit and AWS CodeBuild** helps ensure CI/CD across environments



Automated file movement

AWS Glue jobs are used to automatically move files across folders and zones; this minimizes the manual intervention for data movement



File-level quality controls

Scripts are run through **AWS Lambda** to check the file quality in the landing zone (e.g., null files); poor quality files are rejected, and good files are moved to the raw zone

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.



Data Lake Accelerator

The AWS-native data lake

10% faster
ETL workflow creations

4 weeks
To set up entire platform

30% faster
To transfer data with preconfigured connectors

On demand
Pay only when used

Up to 95%
Cheaper than traditional packaged data services

10 TB store
less than $300/month

Thousands
Of ETL jobs/day

99.99%
Availability out of the box

11 9s
Of durability at the store level

80% faster
To build access controls

3 levels
Of predefined data controls

Compliance
Supports CJIS, PII, HIPAA, and more

The benefits

Speed to market

Cost-effective

Scale and resilience

Built-in security

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Data Lake Accelerator

PHASED APPROACH: BUILDING A MODERN DATA PLATFORM

Rapid data Lake build-out

Phase 1

Quickly scale
a team with the right DevSecOps skills

Architect
a modular, scalable data lake, separate from core IT systems

Ingest raw data
into landing zones

- ▲ Speed to market
- ▼ Time to insights
- ▲ Flexible architectures
- ▼ TCO

Scaled data lake

Phase 2

Adopt governance
policies to help secure and protect data

Define access controls
to provide the right data to the right people

Enhance
free-flowing analytics and intelligence created from data pool

- ▲ Data trust-worthiness
- ▲ Innovation with polyglot data
- ▲ Data-led decision-making

Evolved data lake

Phase 3

Manage fully-governed data
with low friction and on-demand consumption

Collaborate
through data marketplace

Transform data
into economic insights for new revenue streams and improved customer experiences

- ▲ Governed data access & delivery
- ▲ On-demand data consumption
- ▲ Data monetization

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Global pharmaceutical example

Data platform solution:
Capturing, curating,
and consuming data
using a data lake

Creating a cohesive enterprise data platform

25K
ETL executions
a day

30 MINUTES
to execute AI models
(instead of taking days)

25+
AWS Glue
integrations

$48M
Savings in
software licenses

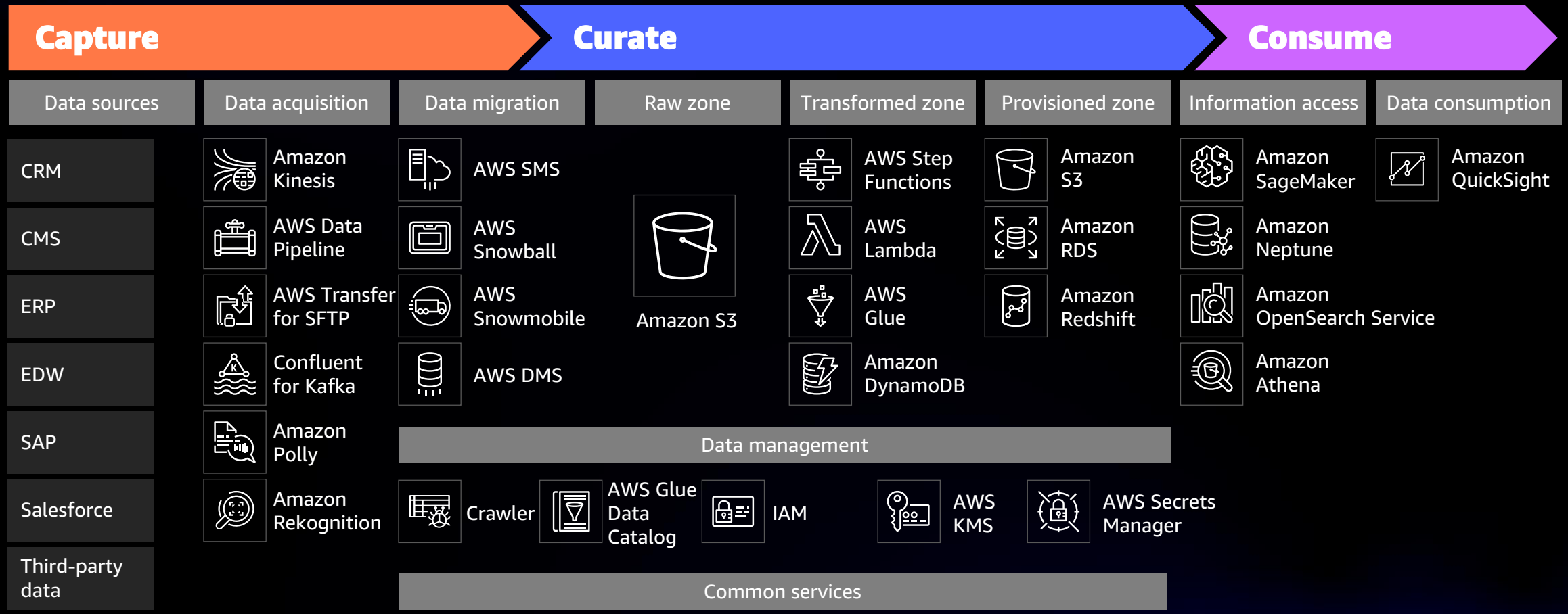
3 HOURS
to execute financial reports
(used to take 4 days)

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

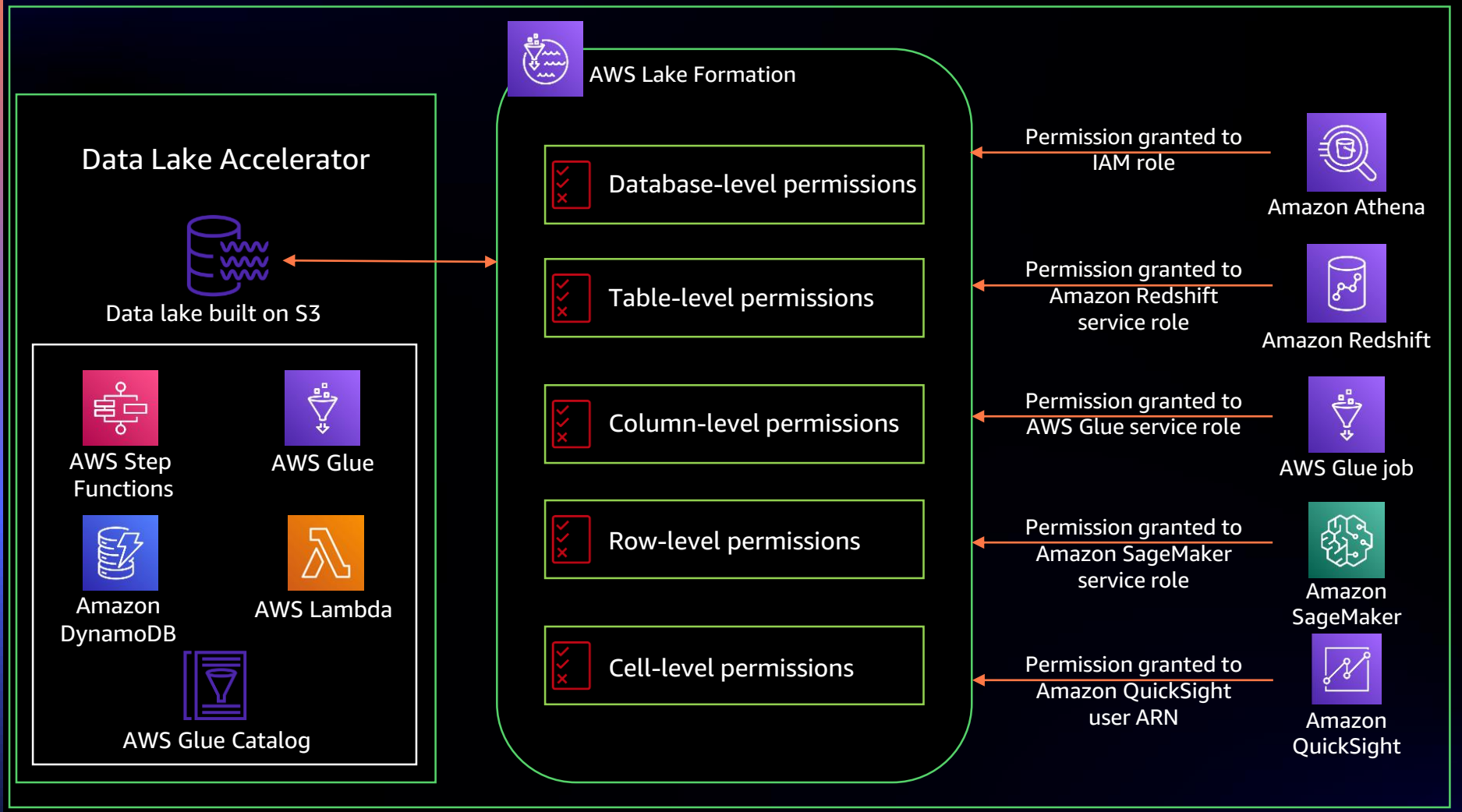
Reference architecture

USE SERVERLESS AND MANAGED SERVICES TO PAY FOR EXACTLY WHAT'S USED



© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AWS Lake Formation integration



- Data Lake Accelerator uses Lake Formation for granular-level data lake access management
- Lake Formation setup is integrated as part of the Accelerator deployment
- Setup will create required roles, register S3 buckets, and grant access to AWS Glue service role to run ETL on data in S3
- All access to data is managed with Lake Formation access control
- Access to data can be managed at a granular level (database, table, column, row, and cell)

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Machine learning

AUTO-ML WORKBENCH

The Data Lake Accelerator uses AWS SageMaker Autopilot to automatically build, train, and tune the best machine learning models based on your data and makes real-time predictions for future data

Data-driven machine learning

Auto-ML Workbench does automatic data transformation and feature engineering to arrive at the best machine learning model after running through 200+ different models

Real-time predictions

The best candidate model that is identified during the process can be deployed into real-time endpoints to make future predictions

Continuous tuning

The autogenerated model is continuously monitored for any drift against the metrics; models will be regenerated/retuned when the model gets stale over time

---



Raw data

Historic data in CSV format with target column

Model autogeneration

Train and tune is done for multiple models after data transformation

Selects best algorithm

Best model is chosen from the ranked-list of recommendations

Real-time predictions

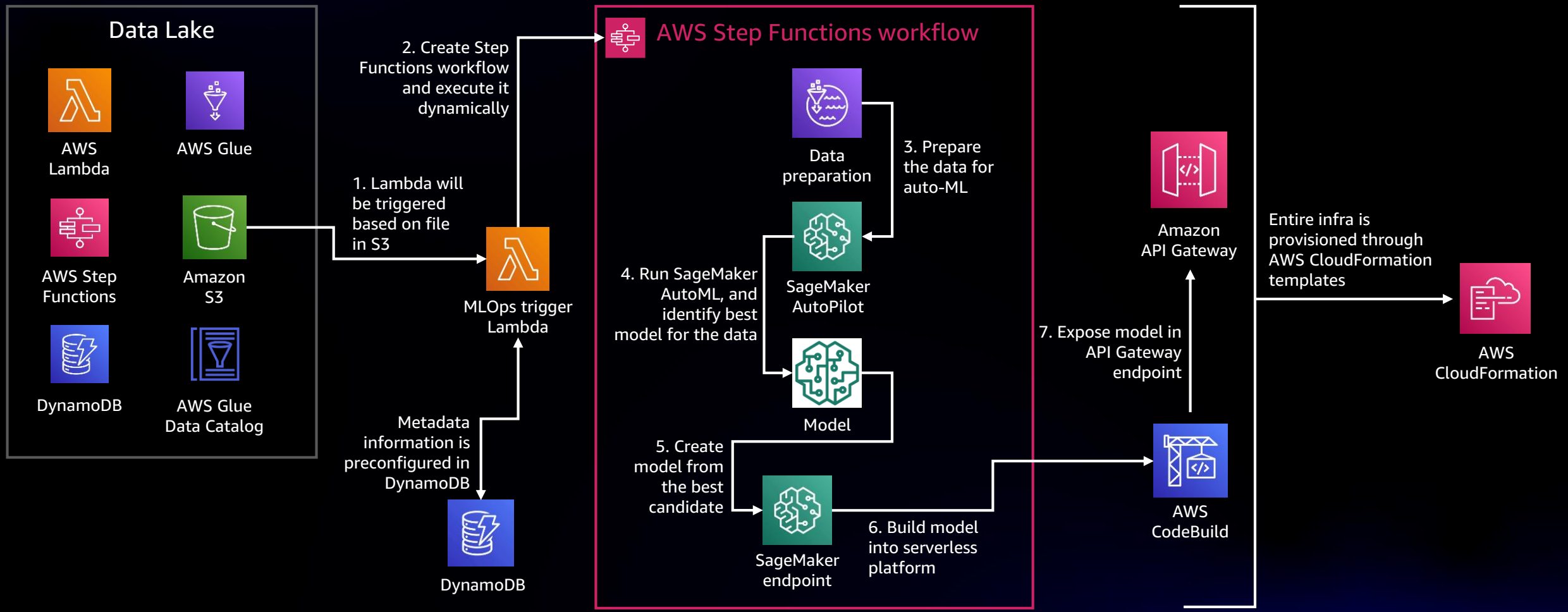
Model is deployed and is available for real-time predictions

Periodic monitoring

Models will be monitored repeatedly to identify and drift toward accurate predictions

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Auto-ML architecture



© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

SAP Data Lake Accelerator

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

SAP Data Lake Accelerator

Delivery accelerator for setting up a data lake in a well-defined zonal and secure architecture, enabling fully automated data movement using AWS-native services



Architecture to bring data together

Rapidly brings together disparate and siloed data through a templatized architecture that runs across SAP and advanced AWS services



SAP data extraction automation

Sets up an automated SAP source system data model analysis that flows into the automatic generation of extraction services and then publishes the configuration for the AWS data lake to consume



Continuous improvement & delivery

Allows for automated deployment and reduced time for lifecycle as new functionality is needed to easily tag, search, share, transform, analyze, and govern specific subsets of data

Outcomes

Create
an enterprise-wide single version of truth for corporate reporting

Accelerate
time and effort required to set up and configure a data lake and reduce deployment costs

Enable
near-real-time analytics to drive better business decision-making

Develop
a foundation for building and deploying ML, IoT, and advanced services

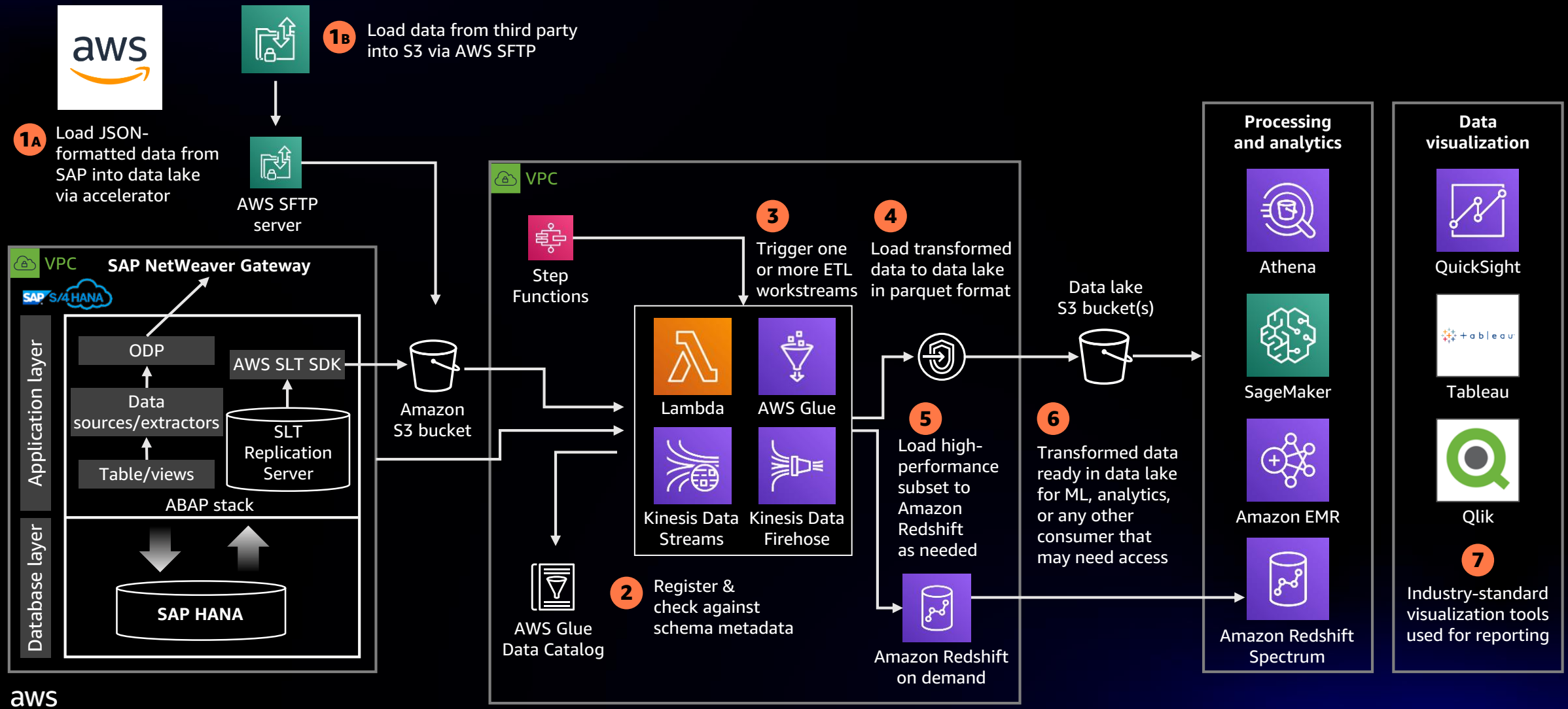
Allow
for cloud-native data agility and AWS system access

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

SAP Data Lake Accelerator

SIMPLIFIED ARCHITECTURE



© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Data Lake Accelerator experts

Global Offering Sponsor

Chris Wegmann

chris.c.wegmann@accenture.com

Global Offering Owner

Jonathan Steinert

Jonathan.r.steinert@accenture.com

Product Owner

Daniel Grube

daniel.grube@accenture.com

Lead Amazon Solutions Architect

Steven Cox

steven.cox@accenture.com

NA Offering Lead

Tim Oliver

timothy.w.oliver@accenture.com

Team

Joshua Anwar & Saroj Lamichhane

saroj.lamichhane@accenture.com

joshua.k.anwar@accenture.com

Data & AI Play Leads

Natalie Timokhina & A.K. Radhakrishnan

natalie.timokhina@accenture.com

a.k.radhakrishnan@accenture.com

SAP on AWS

Trey Taylor

AABG Play Lead

trey.taylor@accenture.com

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

accenture

Thank you!

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.