AWS re:Invent

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

FSI306

CME Group: Enabling data science through event-based design

Rick Fath
Senior Director Platform Engineering
CME Group

Keerti Shah
Senior Solutions Architect
AWS

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

Introductions



Rick Fath
Senior Director Platform Engineering
CME Group



Keerti Shah
Senior Solutions Architect
AWS

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

Who are we?

WHERE THE WORLD MANAGES RISK





CME Group



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

How we do it?

CUSTOMERS MEAN EVERYTHING

CME Group

- CME Group is recognized as the **world's largest most diverse exchange**
- Our CME Globex electronic trading platform extends into **150+ countries**
- In notional terms, more than **$1 quadrillion** changes hands electronically at CME Group each year
- First exchange to operate around the clock; more than 90% of our trading volume is electronic, and investors **access our markets 24/7**



aws

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

CME Group and the cloud by the numbers

Category	Total
Messages processed daily avg	~2 billion
Trading volume daily avg	~20 million
AWS accounts	40+
Storage by size	~20 petabytes
Active Regions	10+
AWS services used	75+

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

Business problems

- How do we address the growing list of data integrations between new acquisitions and vendors?
- How do we consolidate bespoke data copy solutions into something more extensible and reusable?
- How do we reduce the time it takes to onboard new integrations, i.e., weeks to hours?
- How do we expand SaaS data feeds (e.g., Salesforce, Google Analytics)?
- How do we make data accessible and usable for our data scientists?

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

Design requirements

- Onboarding ease was the top priority (config > code)
- Needs to be self-service to reduce bottlenecks
- Requires low operational overhead (serverless and scalable)
- Needs to be cost-effective
- Must complement our real-time streaming pipelines with Amazon S3-based solution
- Extensible to allow growth and deeper integrations over time

aws

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

Data ingestion architecture

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

Architecture deep dive



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

Data collection

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

Data collection – Business problem

DATA COLLECTION – AWS-HOSTED PRODUCER

#	Problem
1	Growing amount of custom “data copy solutions” requiring too many **bespoke solutions** Needed something reusable and extensible
2	**Reduce data exposure** risk of internal data and buckets; keep “private” buckets private
3	Strong desire to complement our streaming data pipelines and data center-based file transfer architecture with an S3-based integration to extend to external entities **Needed an S3-to-S3-based copy solution** to support AWS-hosted customers, vendors, and partners
4	Bucket policy management **growing in complexity** as new integrations are added Simplify bucket policy and reduce deployment dependences and exposure risks with new integrations; evaluate Amazon S3 Access Points

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

Amazon S3 Access Points

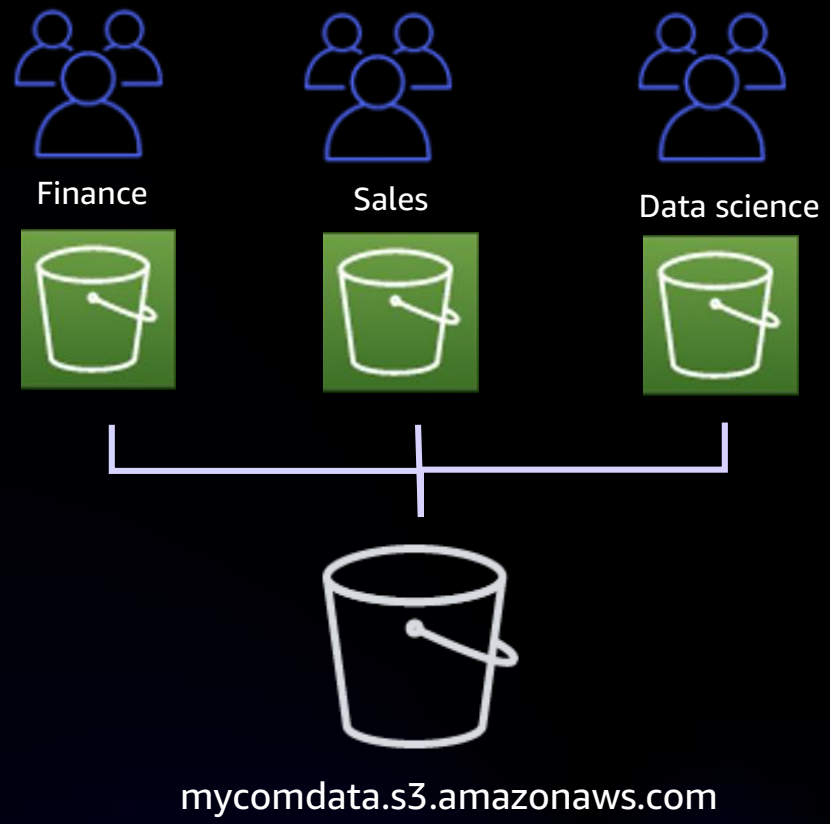
Simplified control for shared buckets accessed by many teams

Use cases

Decentralized teams

Data mesh

Cross-account data exchange



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

Accessing objects in Amazon S3

Users can access objects through dedicated access points in addition to the bucket directly



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

Simplify access control for shared buckets

- Access point allows breaking down policies into smaller chunks
- Smaller blast radius anytime permissions are changed – limited to AP only

Simple bucket policy statement that delegates object access to access point

```
”sid”: “AllowAccessThroughAccessPointToThisBucket”,
“Effect”:”Allow”,
“Principal”:{“*”},
“Action”:[“*”],
“Resource”:[“arn:aws:s3::mycomdata”],
“Condition”:{“ArnLike”:{“s3:DataAccessPointArn”:
”arn:aws:s3:us-west-1:123456789012:accesspoint/*”}}
```

“Finance” AP policy statement allows GET & PUT access to objects with “finance” prefix for the “finance” user

```
”sid”: “FinanceAccessPointPolicy”,
“Effect”:”Allow”,
“Principal”:{“AWS”:”arn:aws:iam::123456789012:user/finance”},
“Action”:[“s3:GetObject”, “s3:PutObject” ],
“Resource”:[“arn:aws:s3:us-west-
1:123456789012:accesspoint/finance/object/finance/*”]
```

“Data-science” AP policy statement allows GET access to all objects with the tag = bi for the “ds” user

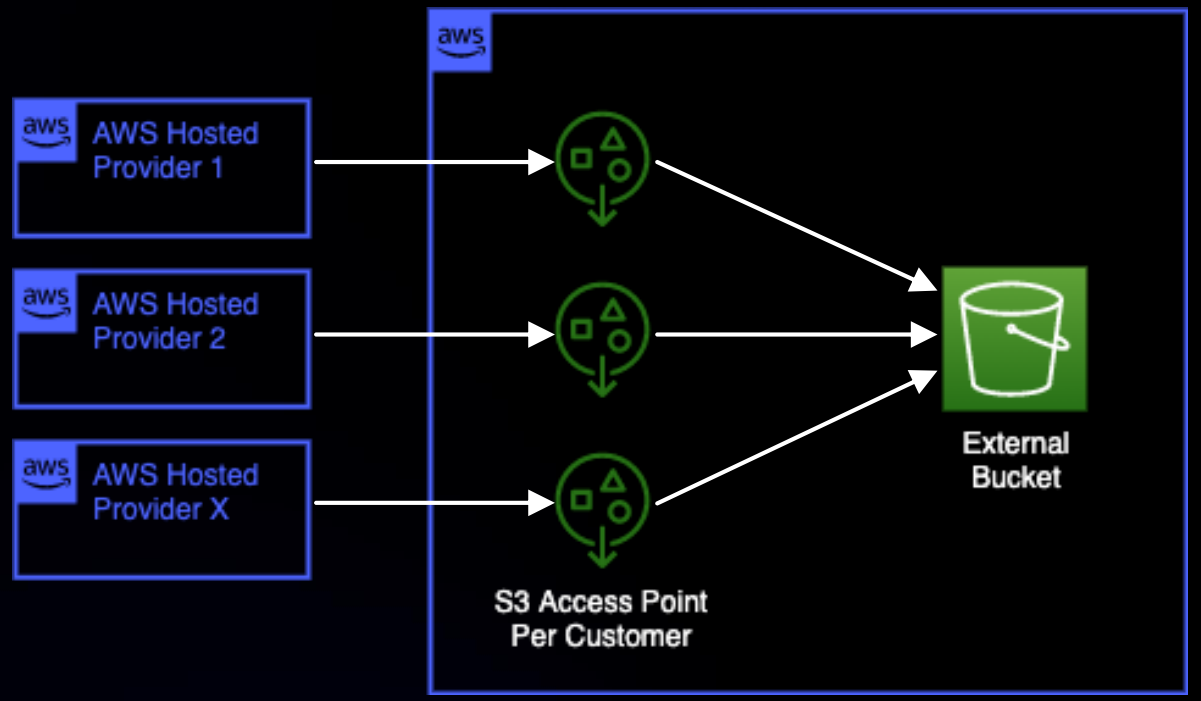
```
”sid”: “Data-Science-AccessPointPolicy”,
“Effect”:”Allow”,
“Principal”:{“AWS”:”arn:aws:iam::123456789012:group/ds”},
“Action”:[“s3:GetObject” ],
“Resource”:[“arn:aws:s3:us-west-1:123456789012:accesspoint/data-science/*”],
“Condition”:{“StringEquals”:{s3:ExistingObjectTag/data”:”bi”}}
```

aws

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

Data collection – AWS-hosted providers

- Centralized “edge” bucket
  - Keep “private” buckets private
  - Introduce more stringent retention policy on inbound bucket
- Migrate to S3 Access Points
  - Simplified & decoupled bucket policy
  - Binding access points to VPC\IPs
  - Access point per producer
  - Object ownership enforcement



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

Data collection – Business problem

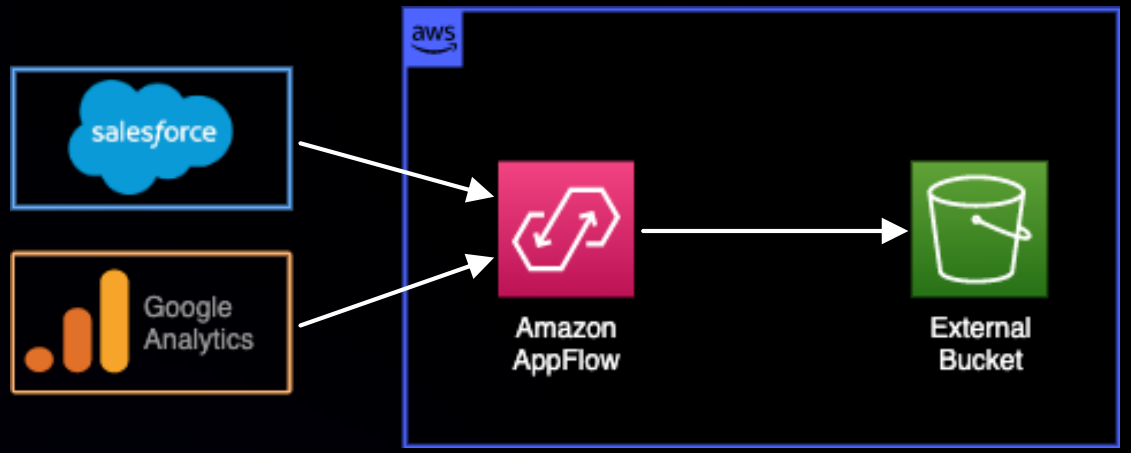
DATA COLLECTION – SAAS PLATFORM COLLECTION

#	Problem
1	Growing business appetite to mine data from SaaS data platforms for data science Needed a solution to ingest this data into the data lake
2	Growing list of target platform data capture, including Salesforce, Pardot, and Google Analytics, to name a few
3	SaaS data ingestion should be simple, secure, and efficient operationally Serverless is ideal Evaluate existing commercial solutions, as well as Amazon AppFlow

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

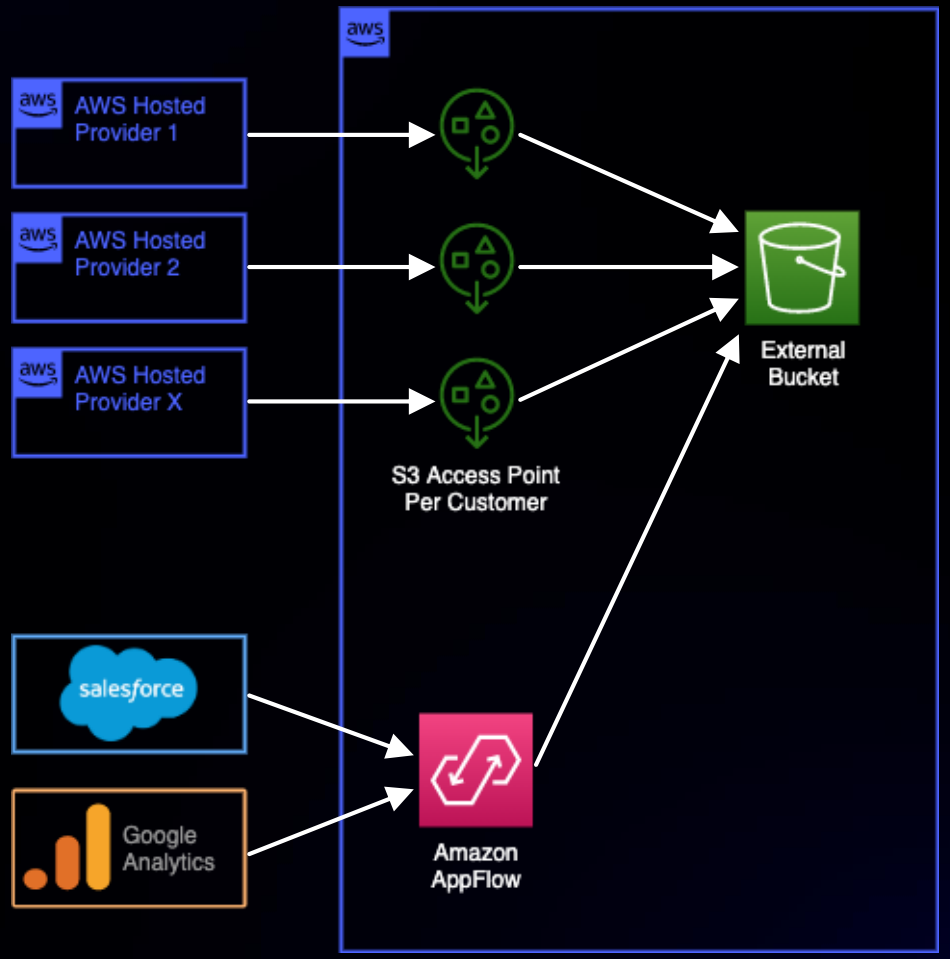
Data collection – SaaS platforms

- Introduce Amazon AppFlow
  - More network efficient vs. data center pulls
  - Simple and cost-effective
  - Features of interest
    - Scheduled, incremental flows
    - Encryption on connections
    - Integrated logging
    - Supports modern data formats



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

Data ingestion architecture



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

Data routing

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

Data routing – Business problem

DATA ROUTING – SIMPLE S3 COPY

#	Problem
1	Event-driven routing is essential In order to support SLAs, data movement needs to be near-real time
2	Support for multiple destination delivery is also critical because a single dataset may have multiple usage and storage strategies (e.g., commercial, archival, compliance, and business intelligence)
5	Routing internals needs to be configuration-based, minimize custom code Require automated onboarding to scale

aws

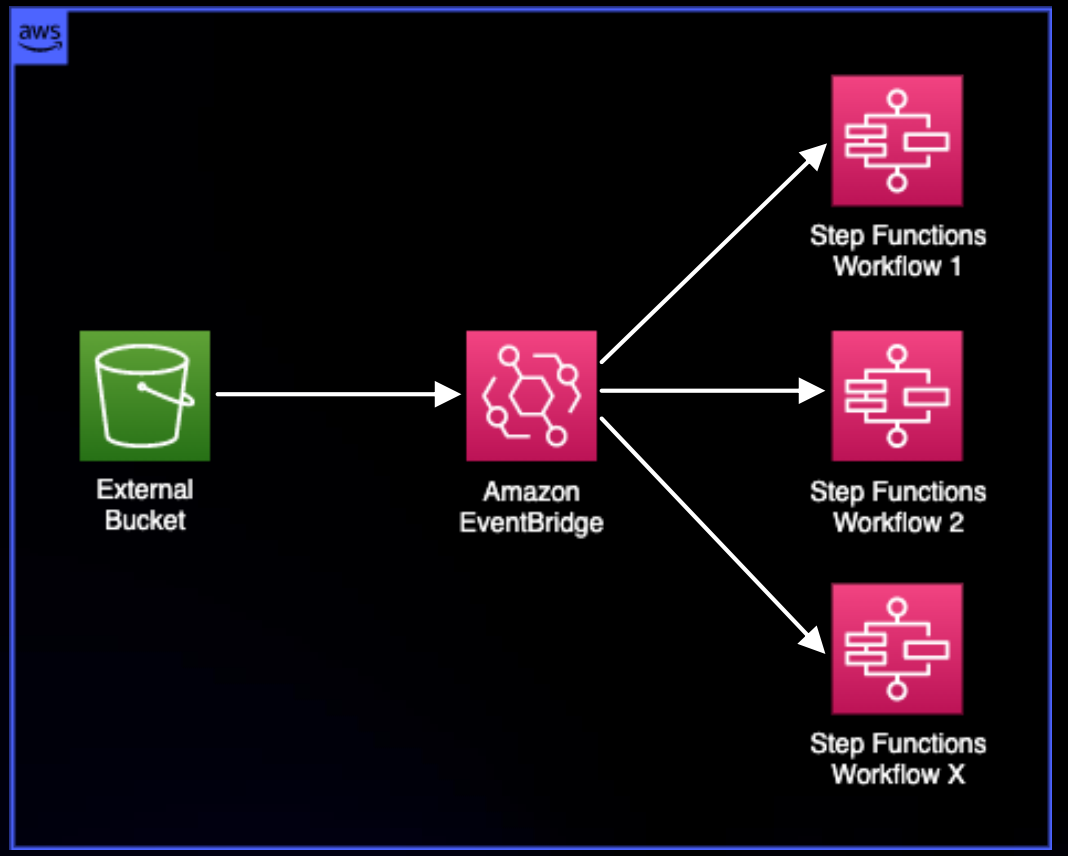
© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Data routing

Amazon EventBridge: A serverless event bus service for AWS services, your own applications, and SaaS providers

Amazon S3 Event Notifications vs. Amazon EventBridge

- Support for wildcard or regular expression matching
- S3 Event Notifications limits the choice of targets – AWS Lambda , Amazon SQS, and Amazon SNS
- EventBridge supports 15+ services as targets, including Lambda, Amazon SQS, Amazon SNS, Amazon Kinesis Data Streams, Amazon Kinesis Data Firehose, AWS Step Functions, and many more
- EventBridge rules can run on schedule as well as when events are received

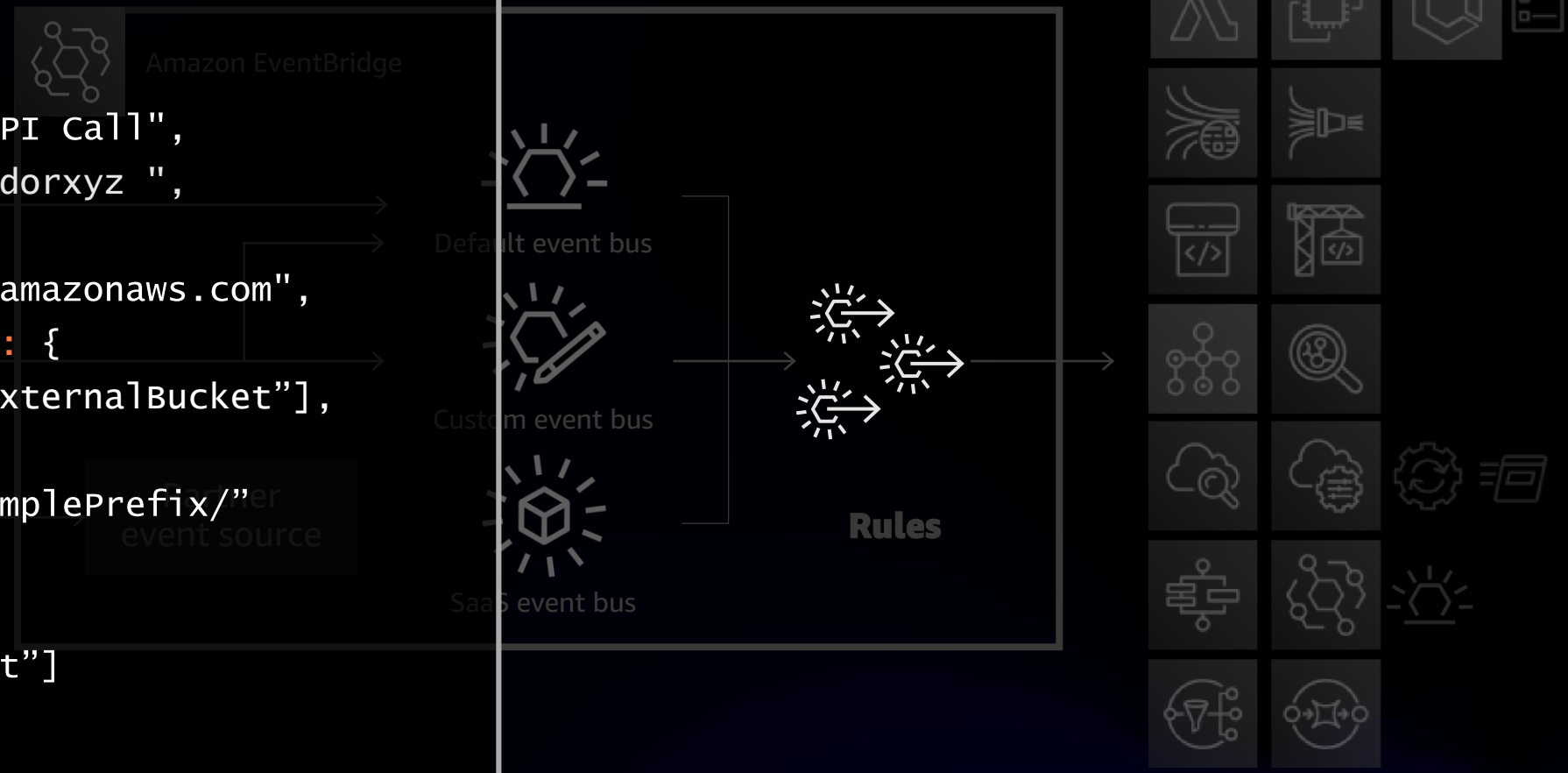


© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon EventBridge

Example event:

```
{
  "detail-type": ”AWS API Call",
  "source": "aws.s3/vendorxyz ",
  "detail": {
    ”eventSource": ”s3.amazonaws.com",
    ”requestParameters": {
      “bucketName”: [“externalBucket”],
      “key” : [{
          ”prefix”: “samplePrefix/”
       }]
     },
“eventName”: [“PutObject”]
}
```

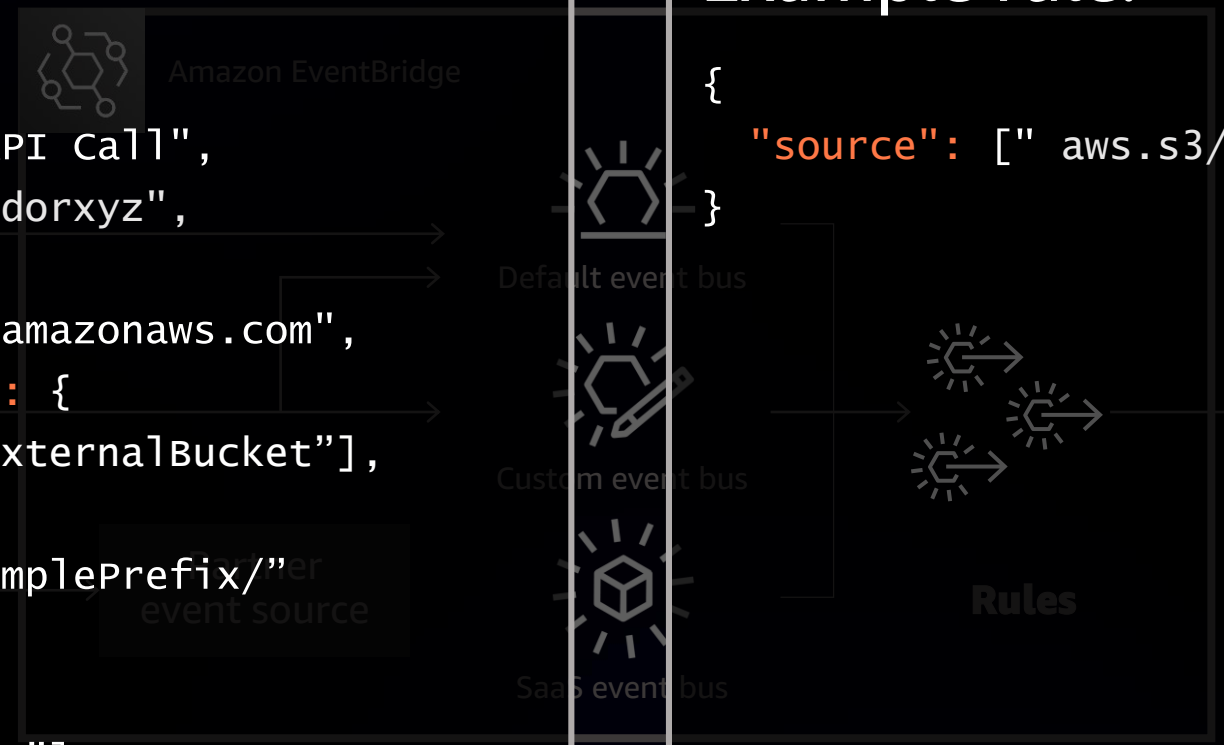


© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon EventBridge

Example event:

```
{
  "detail-type": ”AWS API Call",
  "source": "aws.s3/vendorxyz",
  "detail": {
    ”eventSource": ”s3.amazonaws.com",
   ”requestParameters": {
      “bucketName”: [“externalBucket”],
      “key” : [{
          ”prefix”: “samplePrefix/”
       }]
     },
“eventName”: [“PutObject”]
}
```



Example rule:

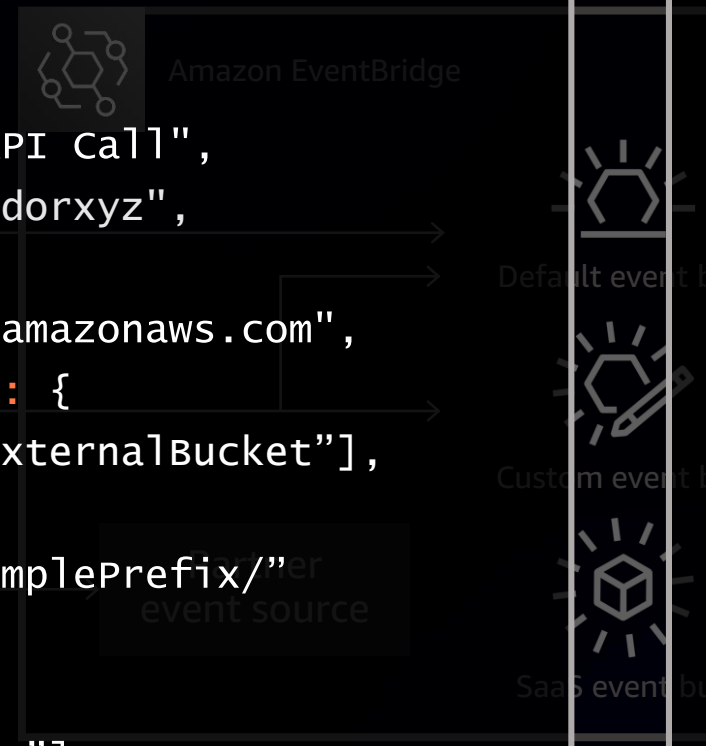
```
{
  "source": [" aws.s3/vendorxyz "]
}
```

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

Amazon EventBridge

Example event:

```
{
  "detail-type": ”AWS API Call",
  "source": "aws.s3/vendorxyz",
  "detail": {
    ”eventSource": ”s3.amazonaws.com",
    ”requestParameters": {
      “bucketName”: [“externalBucket”],
      “key” : [{
          ”prefix”: “samplePrefix/”
       }]
     },
“eventName”: [“PutObject”]
}
```



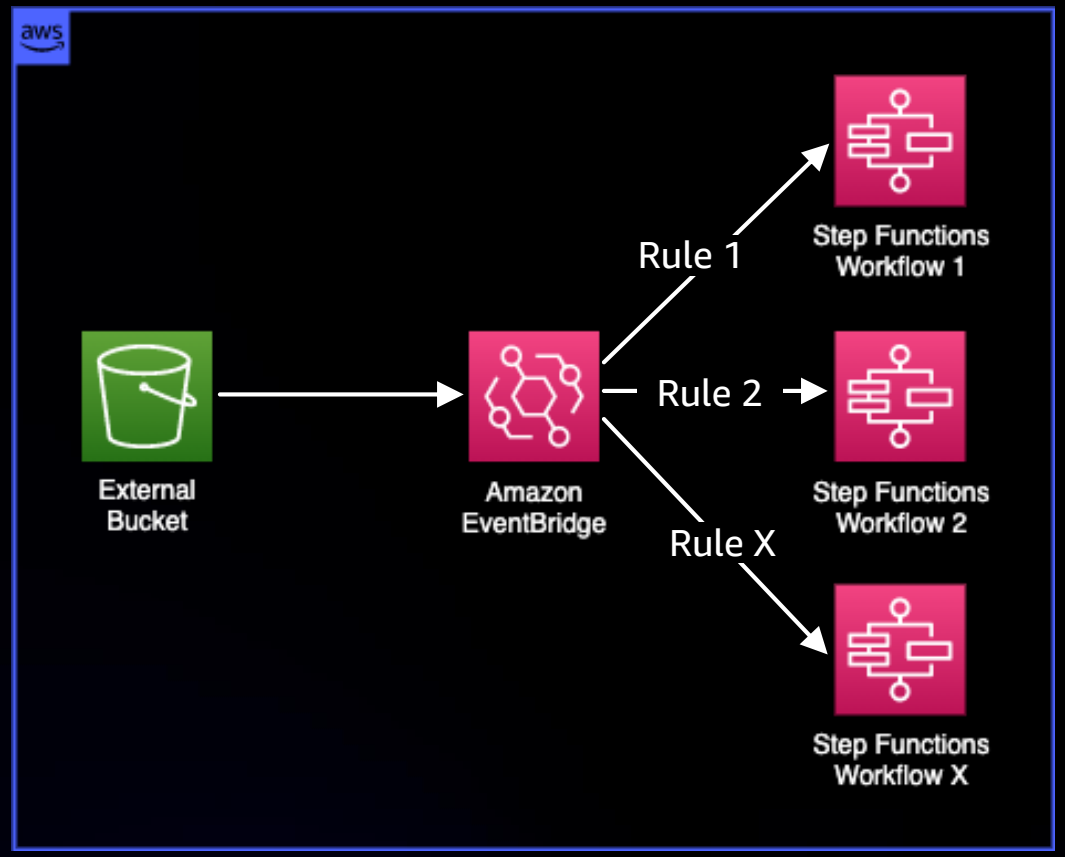
Example rule:

```
  "detail": {
   ”requestParameters": {
      “bucketName”: [“externalBucket”],
      “key” : [{
          ”prefix”: “samplePrefix/”
       }]
    }
   }
```

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

AWS Step Functions

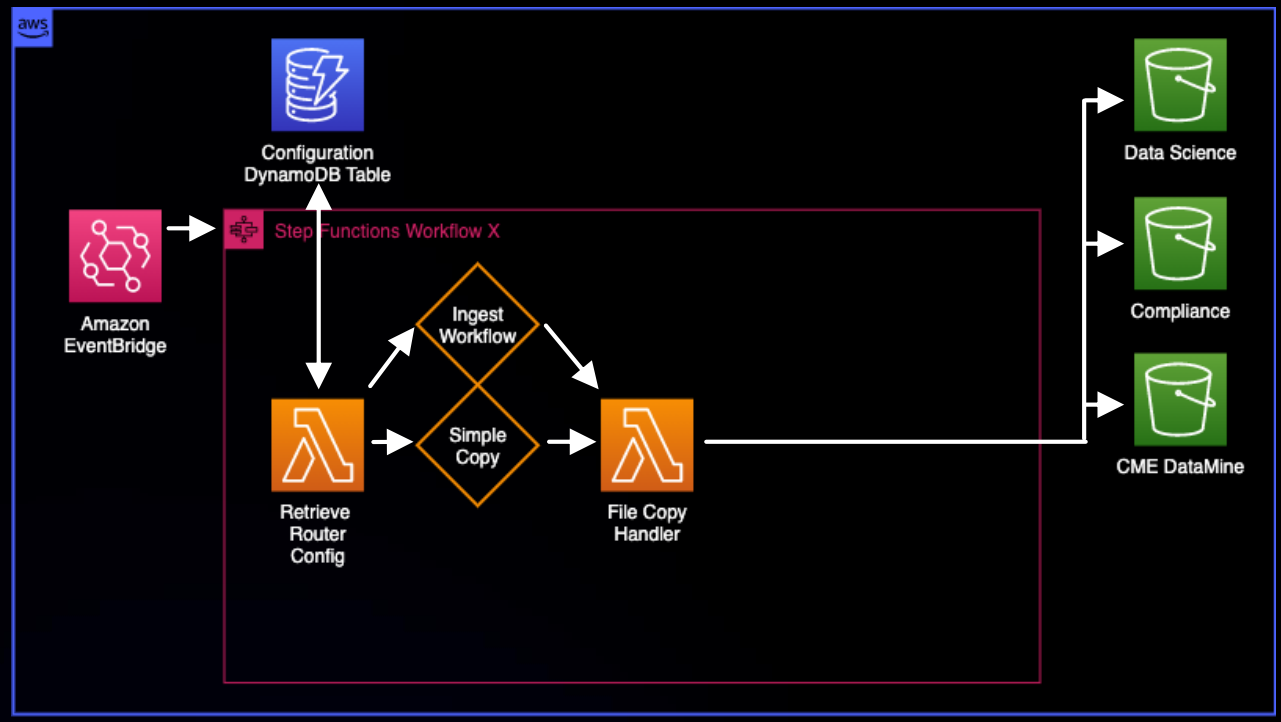
Serverless solution for orchestrating business workflows



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

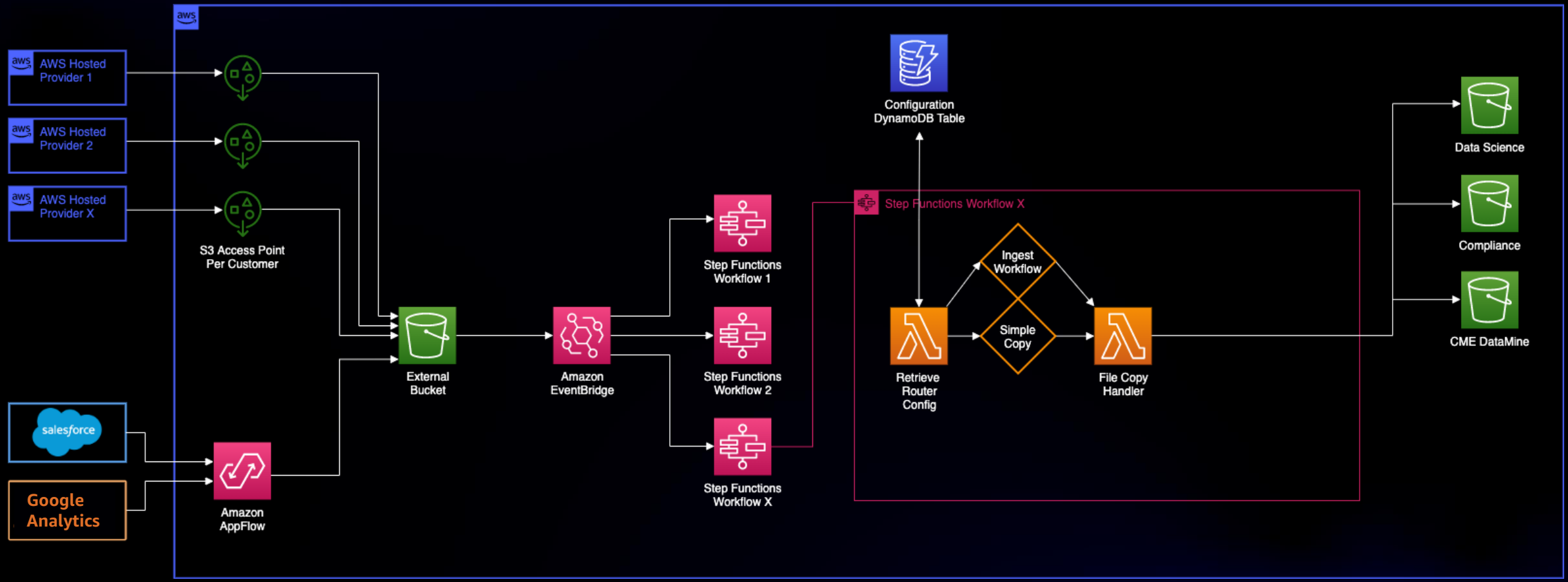
Data routing – Business workflow

- Real-time invocation via Amazon EventBridge
- Multiple S3 destinations based on storage strategy
- Amazon DynamoDB as serverless configuration store
- Easy to extend to new buckets
- Keeps private buckets “private”
- Supports simple file renaming



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

Event-driven architecture



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

Data routing – Business problem

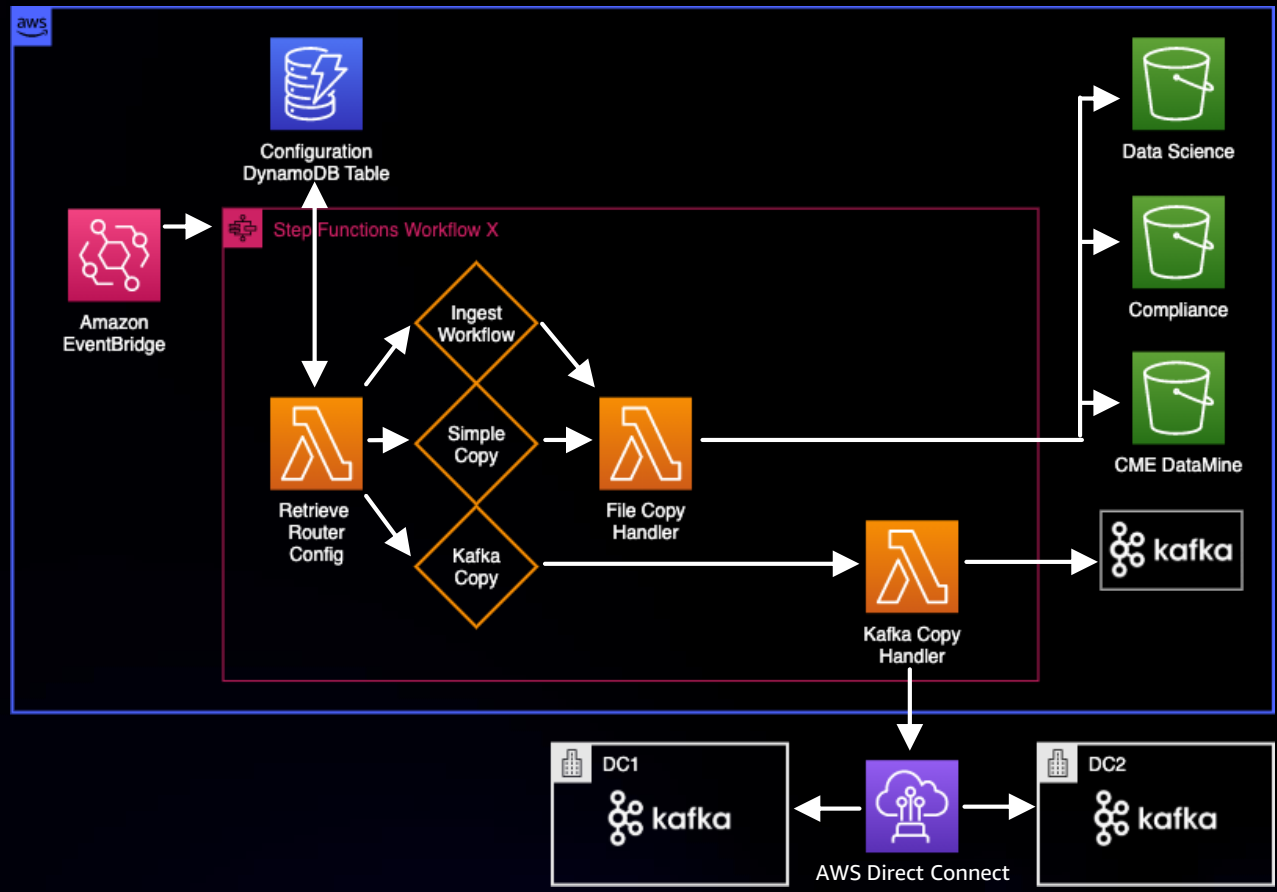
DATA ROUTING – KAFKA STREAMING

#	Problem
1	Needed to publish an S3 object to CME-managed Kafka for time-sensitive data center native workload
2	Data center developers should not be concerned with cloud architecture; need transparent transport

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

Data routing – Kafka publisher

- Real-time Kafka publisher to data center Kafka
- Simplifies data center consumption
- Supports fanout distribution



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

Data transformation

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

Data transformation – Business problem

#	Problem
1	Complex SaaS data formats not user-friendly Very nested and not optimized for query Need to ensure consistent data storage architecture with data lake
2	Data producers are responsible for ensuring data is optimized, which leads to format inconsistency and high learning curves Reduce need for producers to be data engineers
3	Data-format optimizations typically done in batch at end of day, resulting in delayed access to users Needed more immediate, intraday data access and responsive query with tools like Amazon Athena

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

Data transformation



- Data ingestion service is born
- A lightweight ETL transformation service that converts and optimizes data to Apache Hudi
- Reduces need for end-of-day batch conversion
- Data produces focus on sending data, not optimizing
- Allows for repartitioning of data

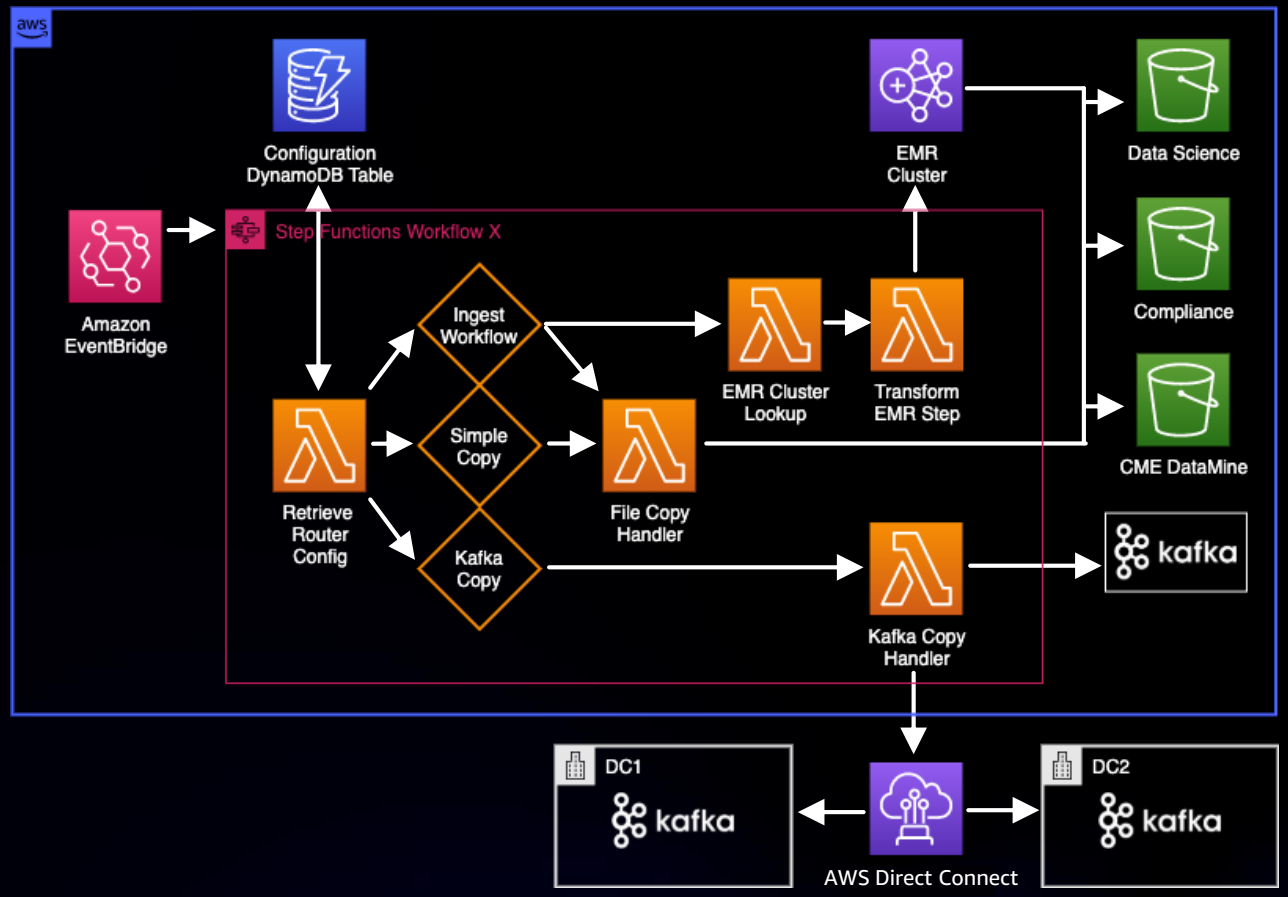
Added bonus: extended service to even internal datasets as well

aws

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

Data transformation

- Cost-effective due to Amazon EMR auto scaling strategies
- AWS integration with Step Functions for EMR steps
- Recent Hudi support for Amazon Athena



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

Data access

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

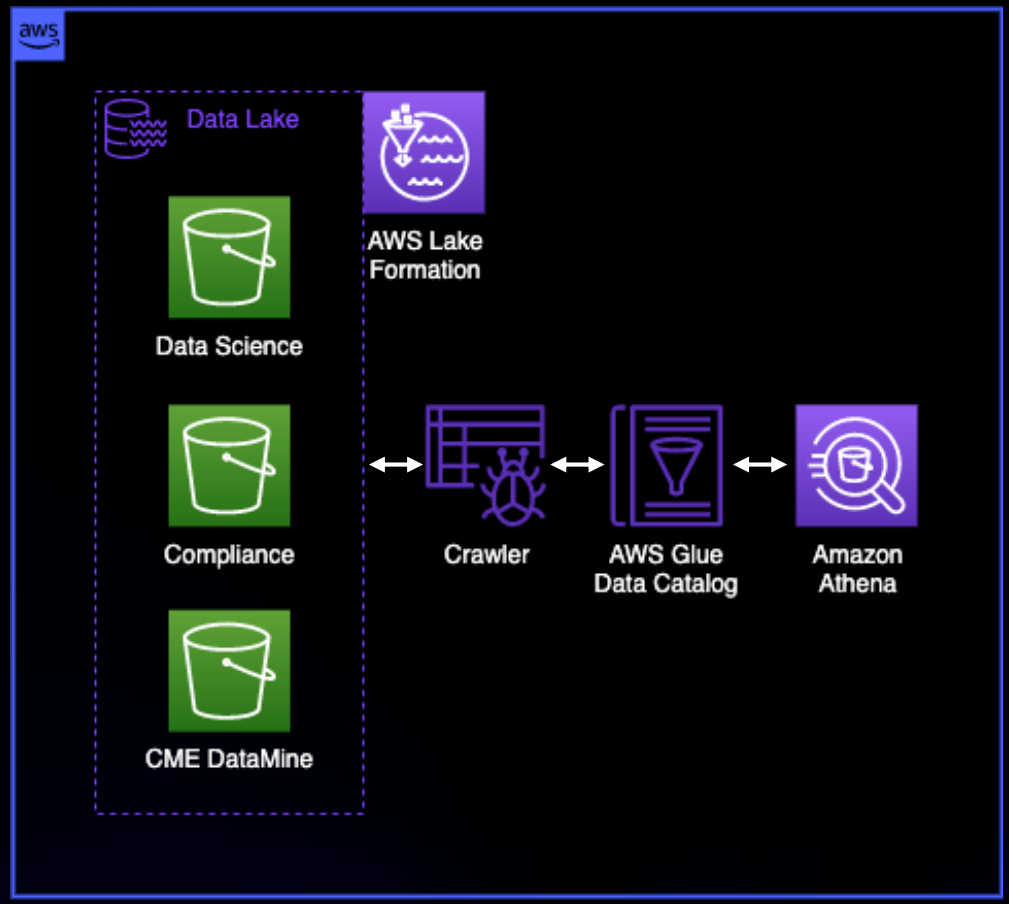
Data access – Business problem

#	Problem
1	Need to ensure data catalog is self-service for simplicity in onboarding
2	Need to ensure consistency around naming and tagging in catalog
3	Need to make partition management easy and automatic for tables

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

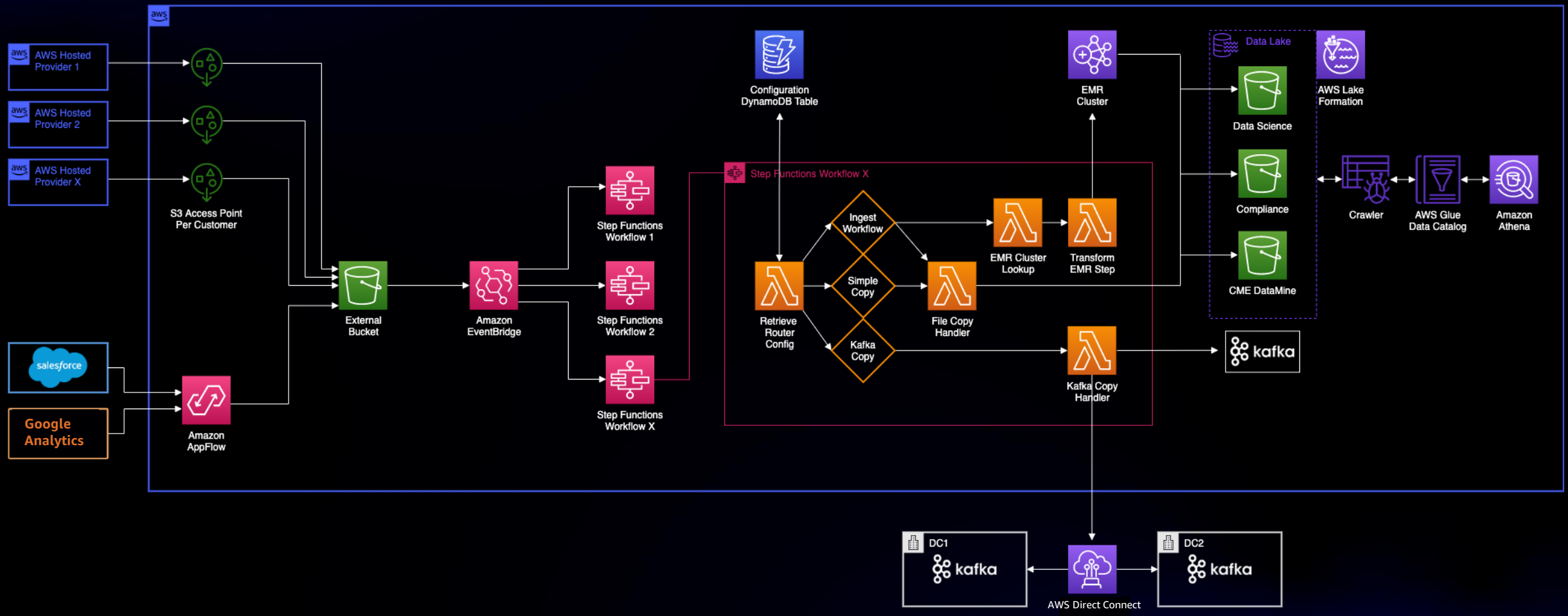
Data catalog onboarding

- Data catalog management (AWS Glue)
  - Self-service onboarding (pull request)
  - Open Policy Agent (OPA) enforcement per catalog standards
  - Crawler enablement for partition management
- Entitlements management (AWS Lake Formation)
  - Cross account catalog sharing
  - Role-based permissions
  - Environment table alias for cross-account access



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

Event-driven architecture



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

Data science

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

Data science – Business problem

#	Problem
1	Needed to provide flexibility in both software packages and compute services needed for data science
2	Need to ensure environment is secure and isolated without direct connectivity to internet
3	Need to provide access to CME production datasets, including restricted data
4	Need to enable data science collaboration based on department membership

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

Data science at CME Group

DATA SCIENCE USE CASE EXAMPLES

- Identifying sales leads
- Understanding market performance around global events
- Liquidity studies
- Regulatory investigations like fraud investigations
- Designing new pricing calculations
- Market volatility analysis

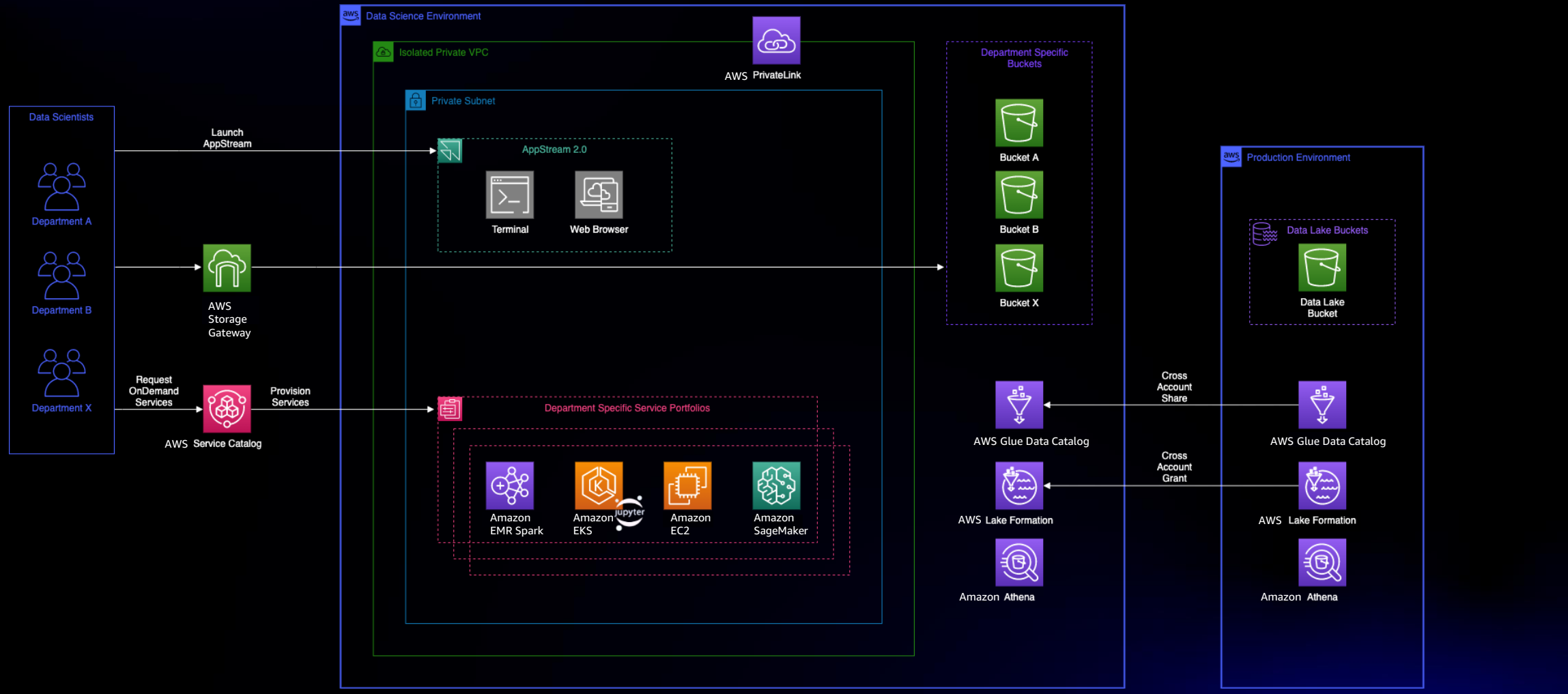


aws

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

Data collection → Data routing → Data transformation → Data access → Data science

- Departmentally segmented **cloud-native environment**
- Self-service access to **on-demand elastic compute**
- Unlimited **secure storage** isolated by department
- Self-service access to **production datasets**
- Software packages **freedom**
- Enable **collaboration**



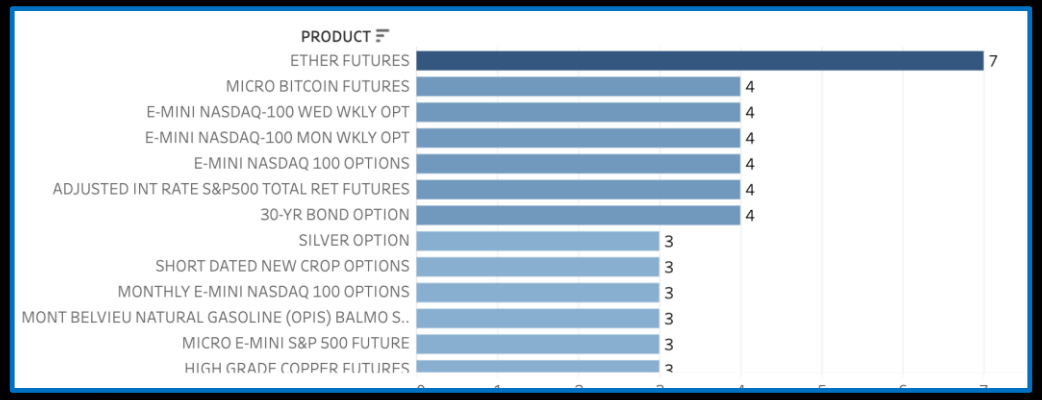
© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

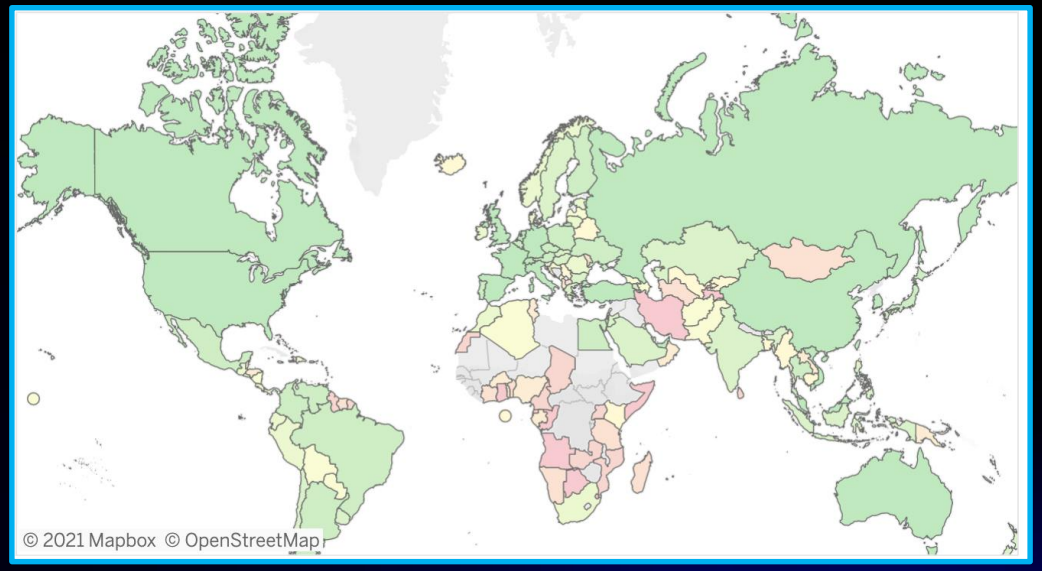
Data science spotlight: Marketing analysis

KNOW THY CUSTOMER

- Client Journey Dashboard – Ability to reconstruct a timeline of customer-relevant interactions with CME digital assets (e.g., visiting web pages, email interactions)

- Actionable Insights Alerts – Ability to configure a customizable alert on potential sales leads based on relevant behavior across groups of client firms or individuals





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

Lessons learned

- Start small with big picture in mind
- Plan for failures and observability
- Automated testing key to keep innovation pace
- Self-service is best path to get adoption
- Not all services work perfect together – work with AWS on product feedback

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

Next steps

- Feature roadmap
  - Inline file scanning
  - Introduce Salesforce as a destination via Amazon AppFlow
  - Add SFTP – AWS Transfer Family data collection endpoint
  - Global access points
  - AWS Glue crawler S3 event notifications

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

Thank you!

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