
AWS
re:Invent

ARC414-S

Accelerated analytics: Building the next-gen data platform for Hertz

Lucy Meyo
VP Architecture and Integration
Hertz

Rohit Sinha
Senior Manager
Analytics and Cognitive
Deloitte Consulting LLP

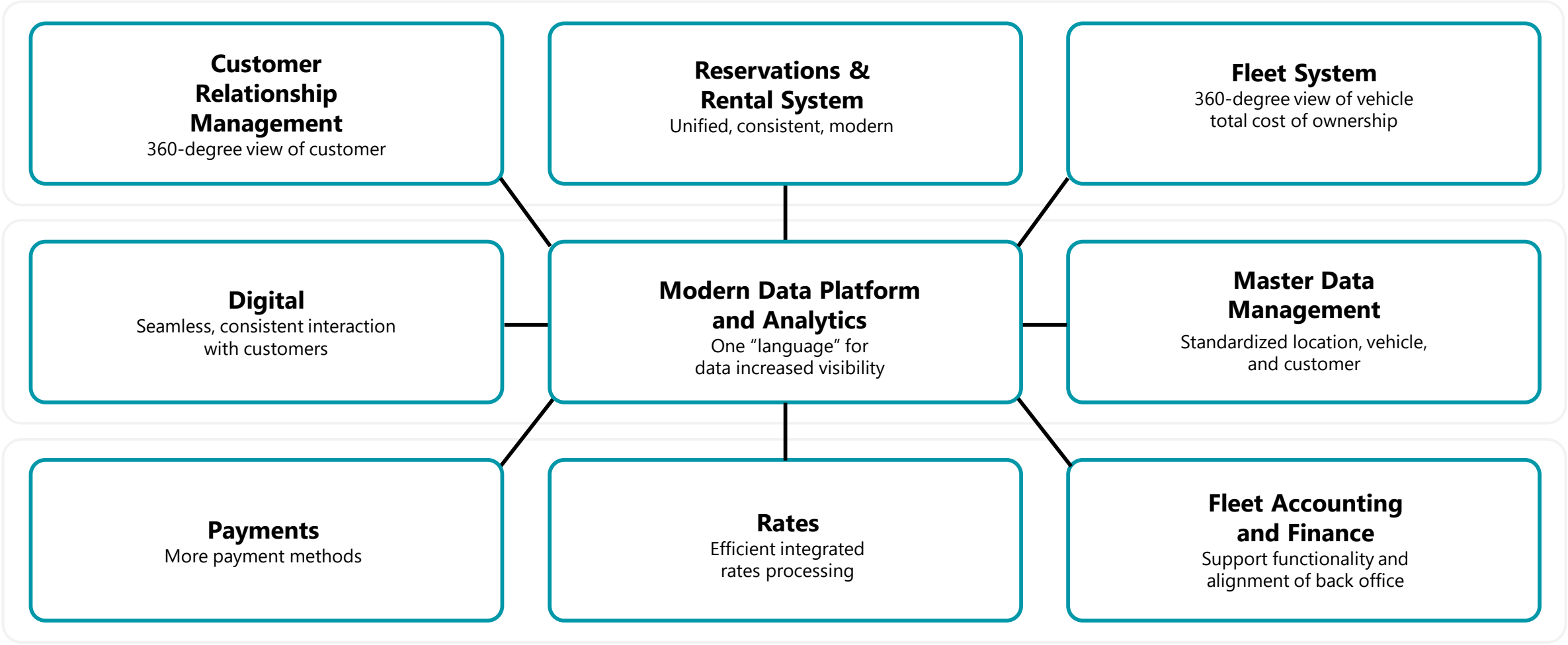
Gowtham Ramu
Specialist Leader
Cloud Engineering
Deloitte Consulting LLP

Agenda

1. **The opportunity**
2. **Accelerated journey to the cloud**
3. **Role of next-generation data platform**
4. **Cloud-native architecture**
5. **Key takeaways**

Digital Transformation

The goal of the Digital Transformation program at Hertz is to deploy an integrated global solution that provides a better view of our customer and greater visibility into our assets



The opportunity

A ***modern data and analytics platform*** was critical to the success of the overall ***digital transformation at Hertz*** to extract value and insights from large amount of data being produced from new systems and create a common platform for cross-functional insights and answering new business questions using the advanced analytics and data science models.

Sharply reduce TCO of data management and analytics with the move to AWS

Leverage scalable elastic compute of cloud infrastructure for modern analytics and data science use cases

Centralize data assets in AWS data lake

Improve regulatory and security controls with the process of centralizing data assets in AWS data lake

Use cloud infrastructure scalability and elasticity as well as spot instances to optimize cost of compute-intensive operations and answer new business questions

With pattern-based ingestion and component-based architecture, enable quicker time to implement new use cases

Enable digitization through new business models around data and analytics

Hertz journey of how Deloitte helped build the MVP for the data platform in less than 10 months

Data platform journey so far

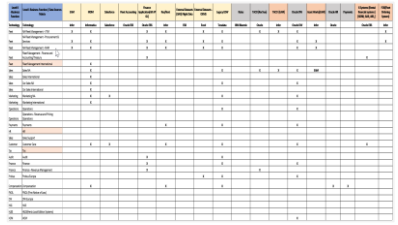
1 Data and reporting requirements

Defined data and reporting requirements across all functions at Hertz and programs within digital transformation to support unified reporting, self-service BI, and analytics

- Conducted workshops with 30+ functional groups/sub-groups to finalize Day1/Day2 reporting and analytics requirements
- Rationalized ~1,800+ reports down to ~280 prioritized Day1 reports/dashboards

2 Data source discovery

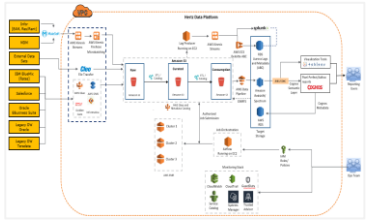
Leveraged data and reporting requirements to do source system discovery; mapped new source system integrations as well as legacy systems and catalog integration gaps



- Identified 30+ data sources outside of new systems to identify integration gaps based on Day 1/Day 2 critical report requirements and associated data needs; catalogued total of 200+ integrations coming in and out of data platform from different systems

3 Architecture alignment and infrastructure standup

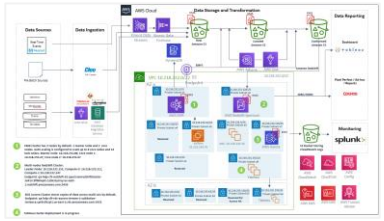
Achieve alignment across a scalable, forward-looking data platform architecture, and stand up infrastructure in AWS to support integration, reporting, analytics, and data science use cases



- Finalized data platform architecture blueprint and technical bill of materials
- Successfully established the data platform architecture on AWS to support in-time integration with Hertz digital transformation program

4 Infrastructure automation

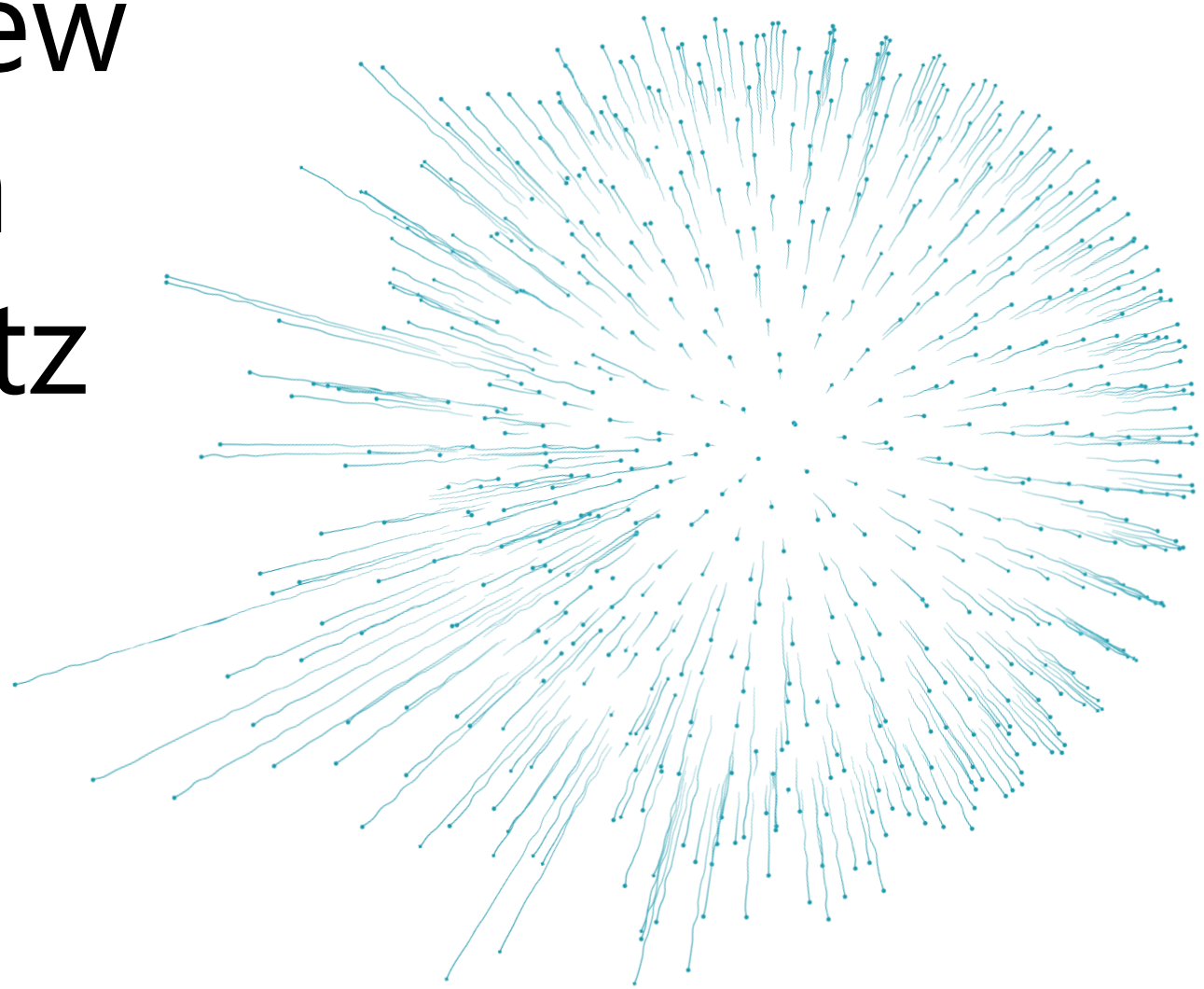
Created automated infrastructure provisioning scripts to standardize infrastructure components across the data and analytics stack



- Built all environments (Dev, QA, UAT, and prod) using CFTs and Ansible scripts for infrastructure automation

Solution overview and role of data platform at Hertz

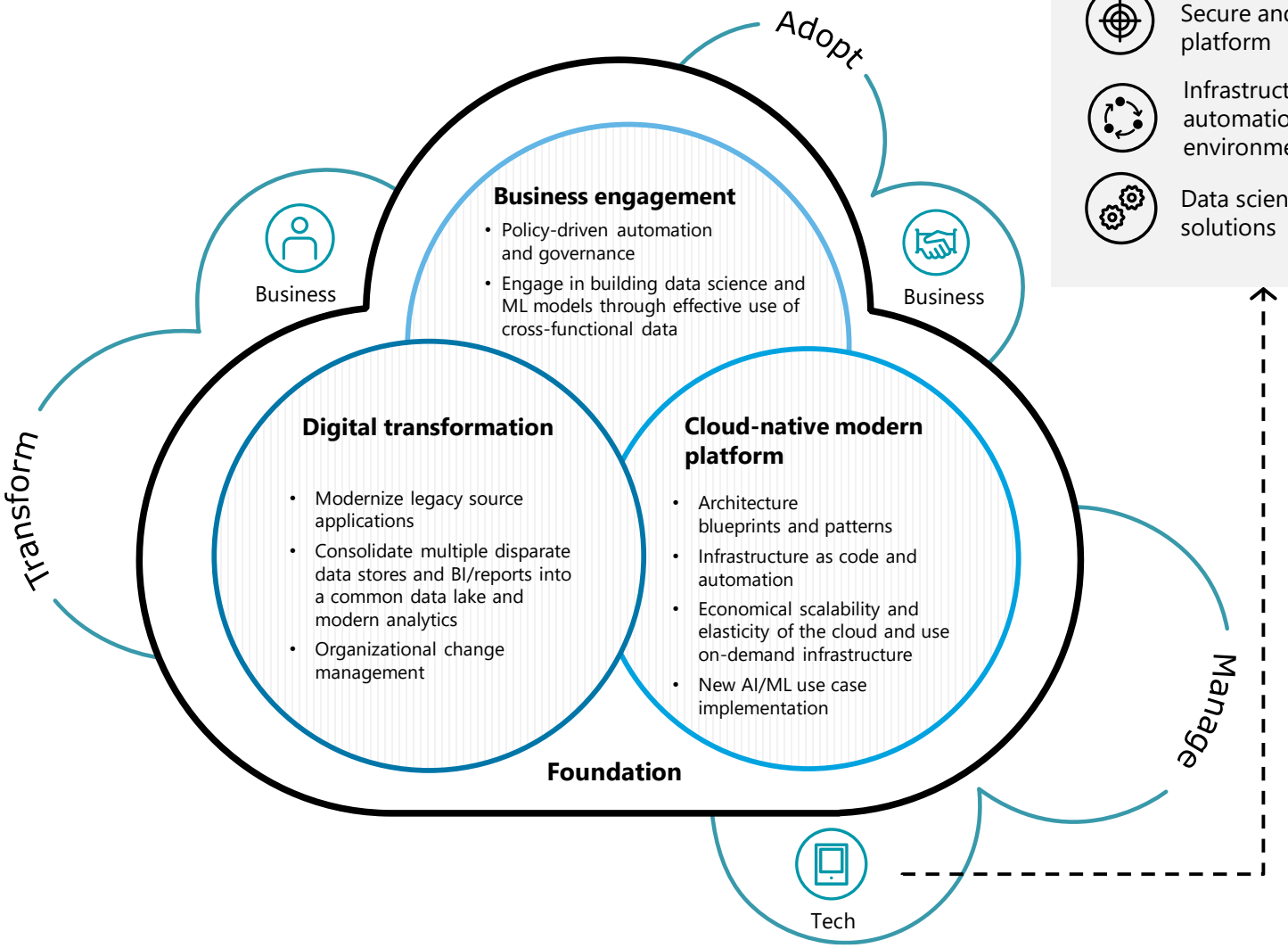
Key purpose and underlying architecture



Solution overview

Cloud-native data platform

Enabling digital transformation journey and providing consolidated data across multiple systems for accelerated insights generation

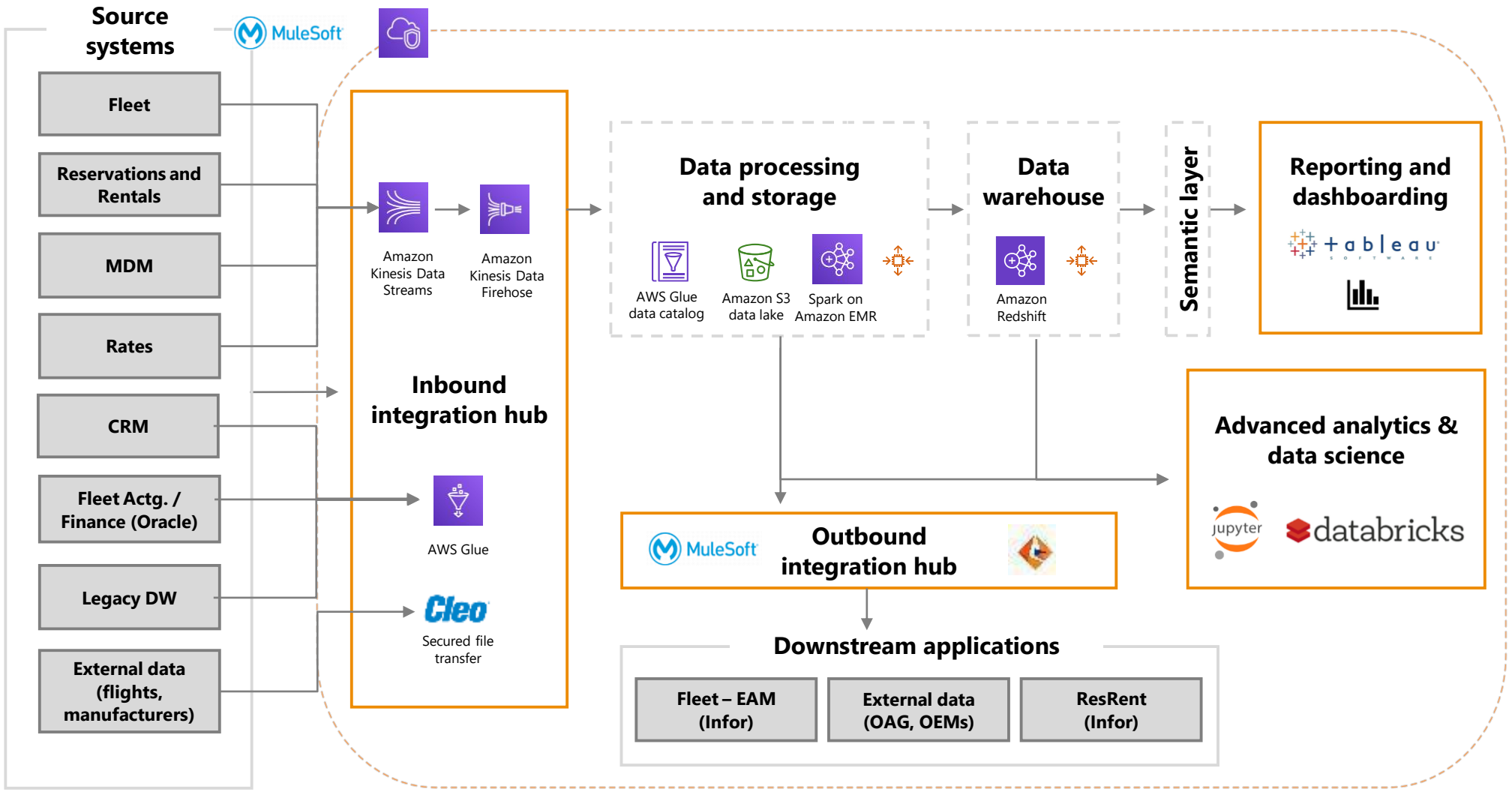


Cloud-native solution architectures

- Batch and streaming ingestion patterns
- Secure and scalable platform
- Infrastructure automation for all environments
- Data science/ML solutions

The role of the next-generation data platform at Hertz

High-level architecture illustration of the data platform at Hertz and the three key purposes of the modern platform



Data platform serves three key purposes

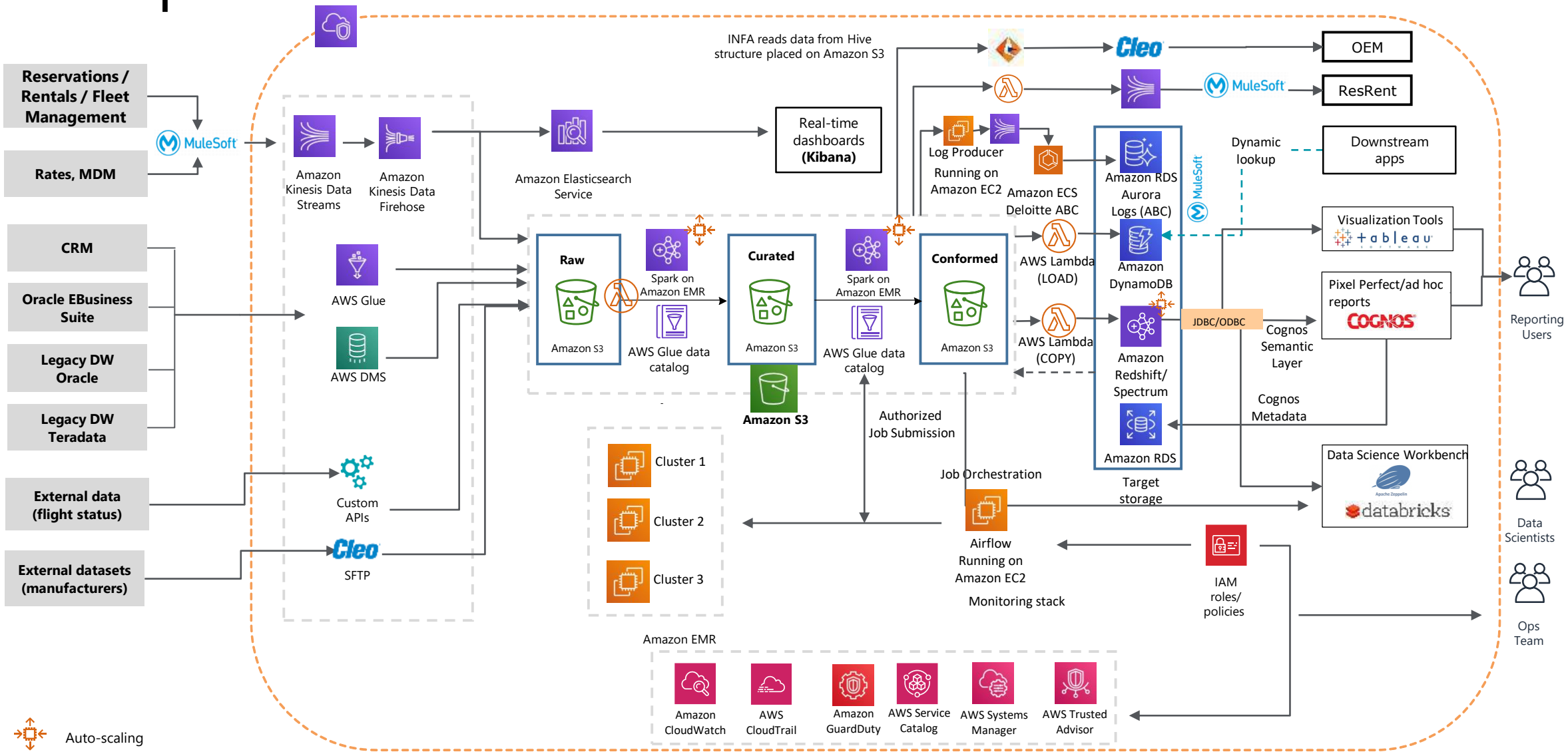
- ***Integration hub*** for getting data from and to other applications
- ***Unified reporting layer*** to serve cross-functional reporting
- ***Advanced analytics platform*** for building data science models to harness deep insights from data

Data platform key components

Cloud-native architecture

Built with all native AWS services to stand up a modern data platform

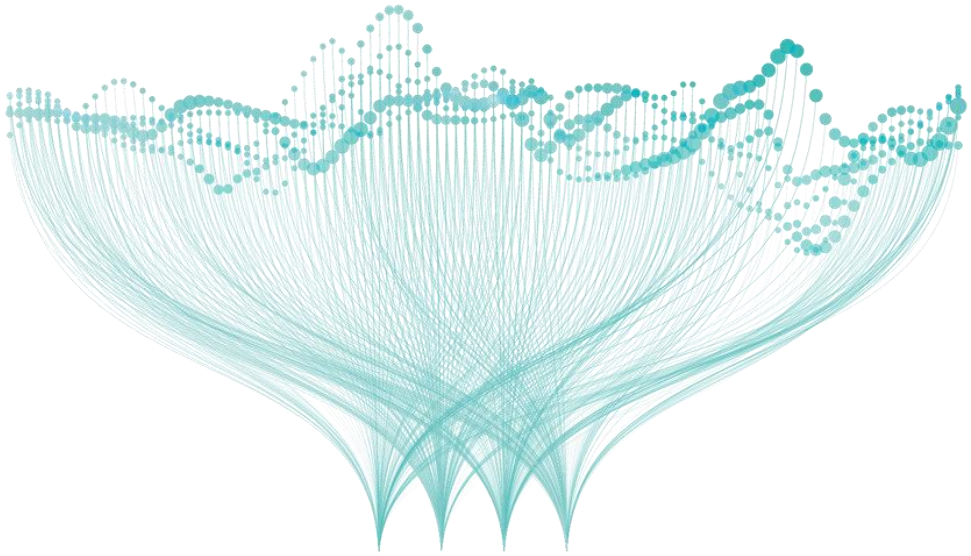
Data platform architecture



Repeatable patterns in architecture

The patterns below illustrate high-level reusable components within the data platform at Hertz that was deployed using native AWS services

- Real-time stream ingestion in Amazon S3 with Amazon Kinesis Data Streams, Amazon Kinesis Data Firehose, and Amazon S3 for all event-driven ingestion in the data platform
- Real time to batch use cases where data-generating system can only produce events and consuming systems can only take batches (orchestrated using Kinesis Data Streams, Kinesis Data Firehose, and AWS Glue or Informatica reading from Hive)
- Real-time API call with external system (Flight Aware) to tie reservation data for optimizing fleet management
- Low-latency fast lookups by transactional applications to look up in real time performed using lookup tables hosted in Amazon DynamoDB with defined read and write capacities

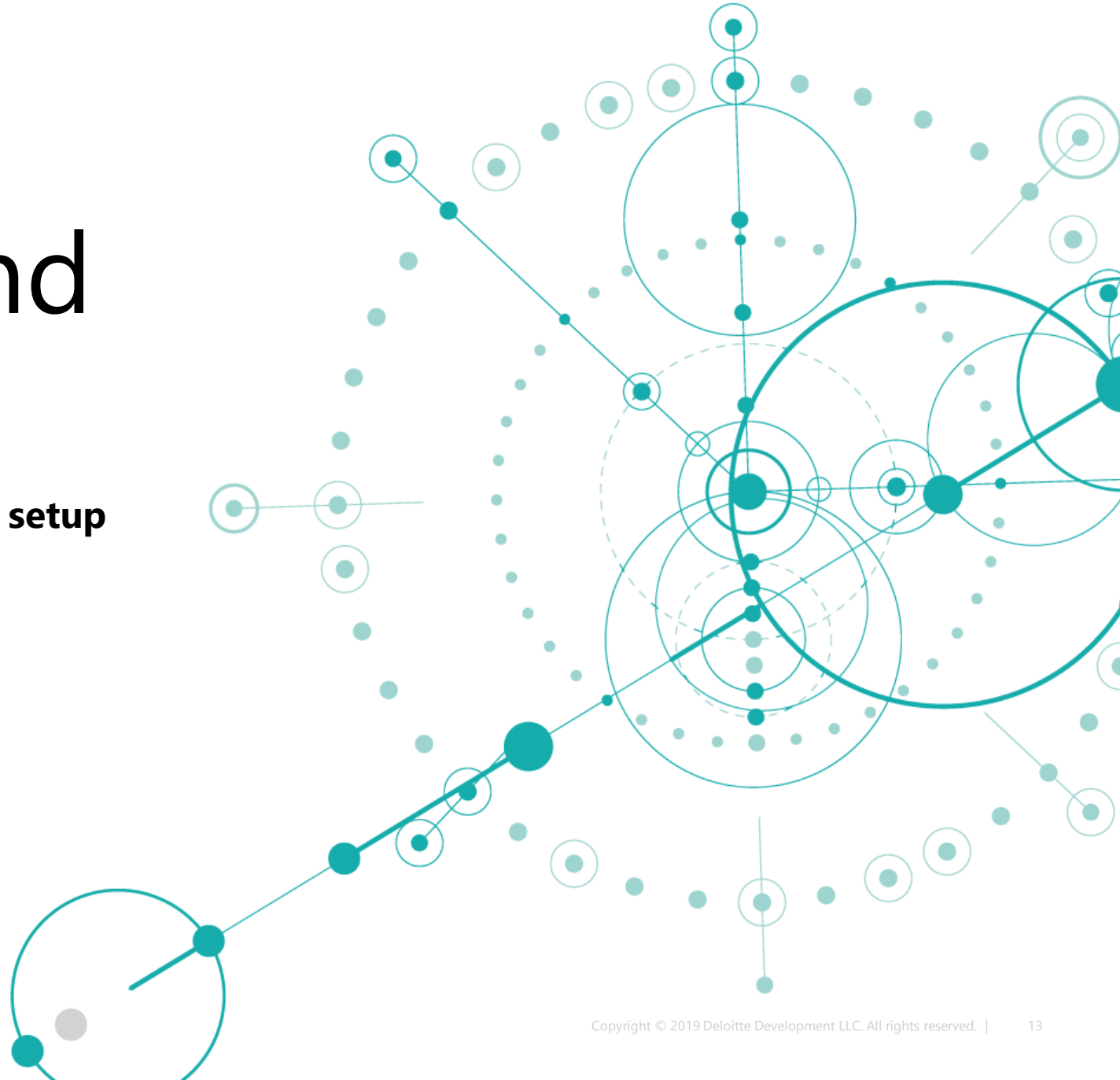


Impact

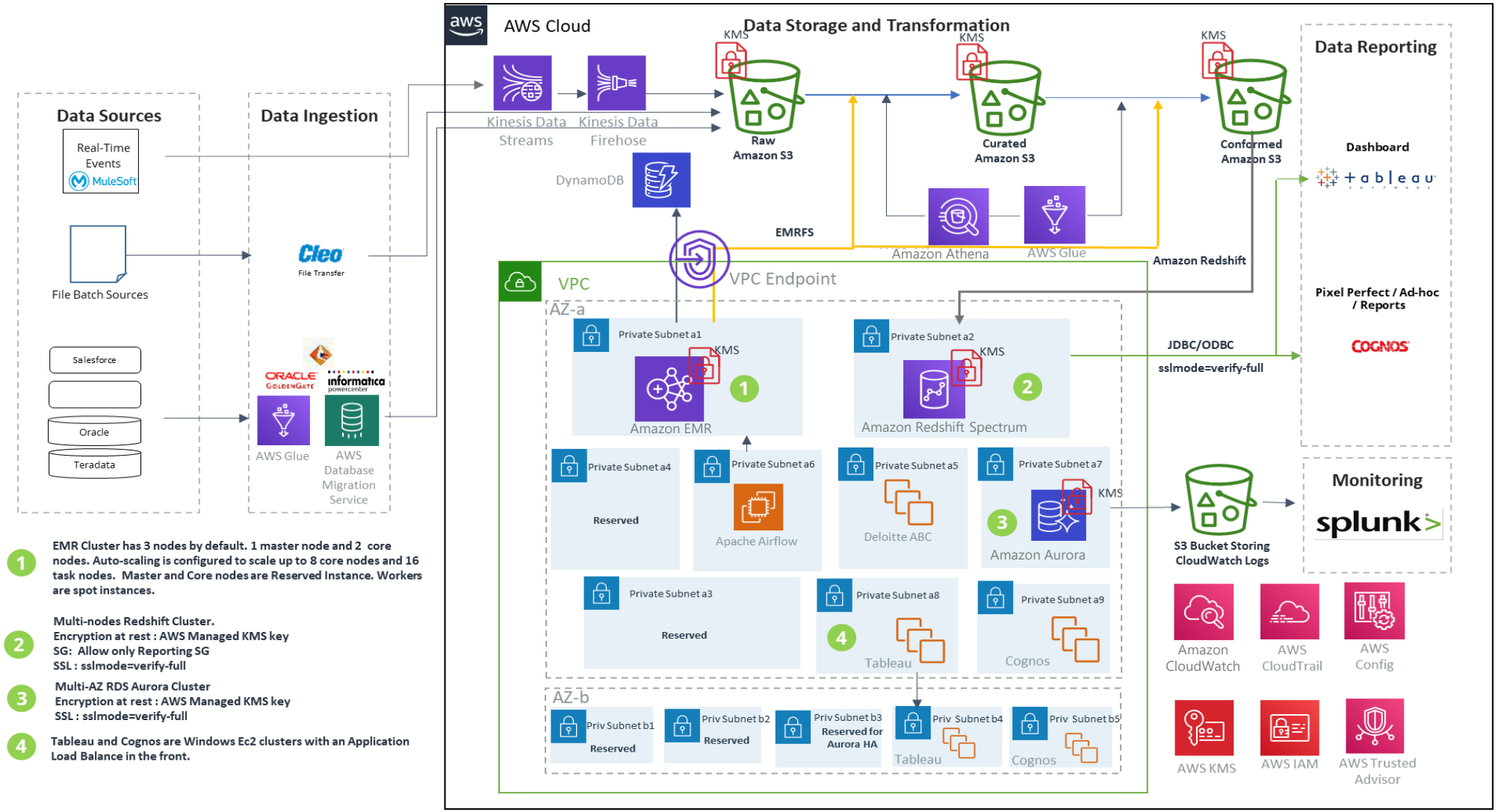
Patterns simplify the ingestion and consumption out of the data platform and avoid point-to-point integration with different components and designs, making it modular code

Infrastructure provisioning and automation

Under the hood infrastructure design and setup



Infrastructure blueprint



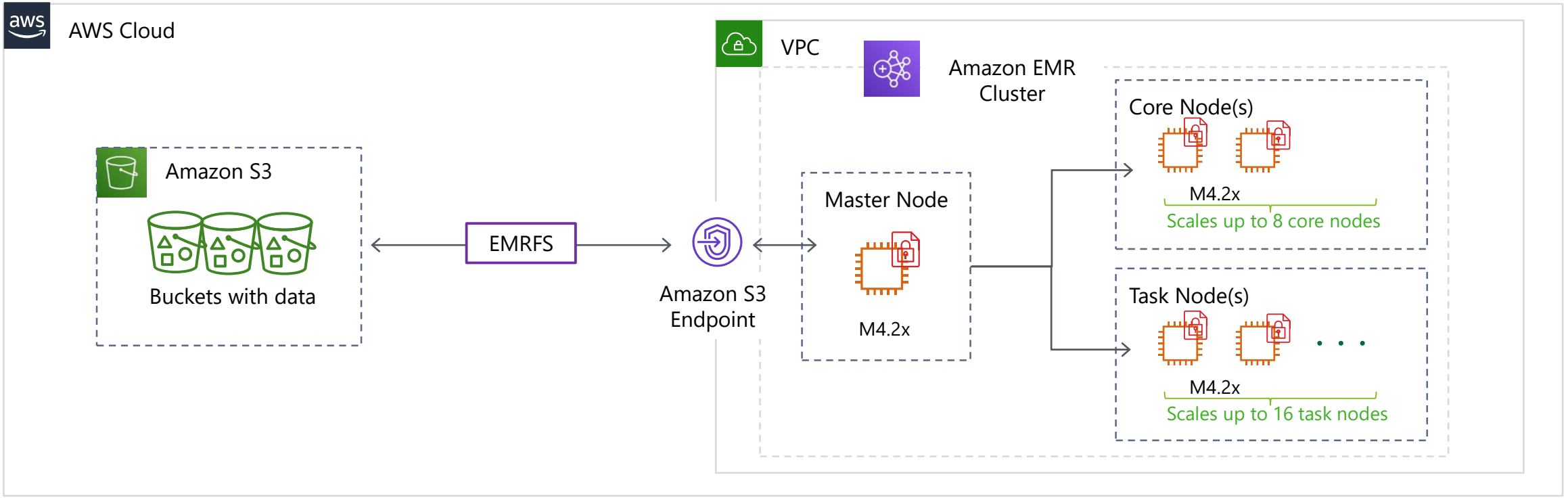
1. EMR Cluster has 3 nodes by default. 1 master node and 2 core nodes. Auto-scaling is configured to scale up to 8 core nodes and 16 task nodes. Master and Core nodes are Reserved Instance. Workers are spot instances.

2. Multi-nodes Redshift Cluster.
Encryption at rest : AWS Managed KMS key
SG: Allow only Reporting SG
SSL : sslmode=verify-full

3. Multi-AZ RDS Aurora Cluster
Encryption at rest : AWS Managed KMS key
SSL : sslmode=verify-full

4. Tableau and Cognos are Windows Ec2 clusters with an Application Load Balance in the front.

Amazon EMR stack diagram

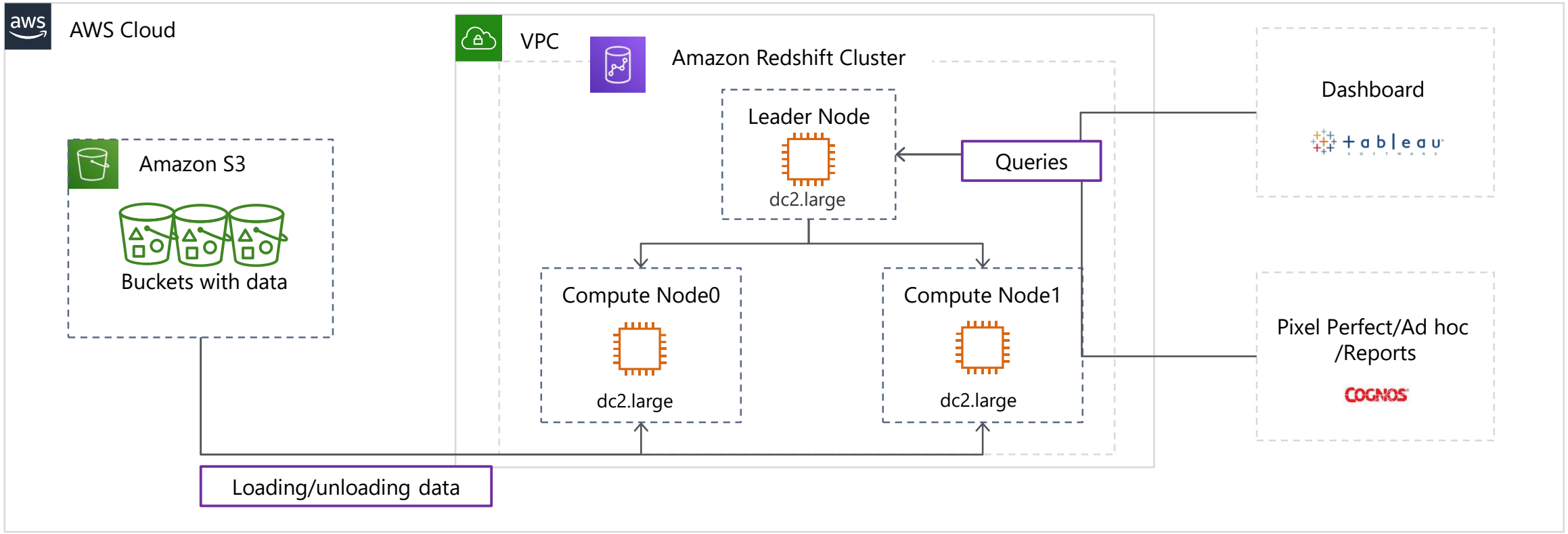


Auto-scaling is configured to scale up to 8 core nodes and 16 task nodes

- **Scale-Out** "Add **1** Instance if **ContainerPendingRatio** is greater than **0.7** for **1** five-minute period with a cooldown of **300** seconds"
- **Scale-In** "Terminate **1** Instance if **ContainerPendingRatio** is less than **0.2** for **3** five-minute periods with a cooldown of **300** seconds"
- **Scale-Out** "Add **1** Instance if **HDFSUtilization** is greater than **90** for **1** five-minute period with a cooldown of **300** seconds"
- **Scale-In** "Terminate **1** Instance if **HDFSUtilization** is less than **70** for **3** five-minute periods with a cooldown of **300** seconds"

The setup includes YARN Resource Manager, Spark History Server, and JupyterHub notebooks

Amazon Redshift stack diagram

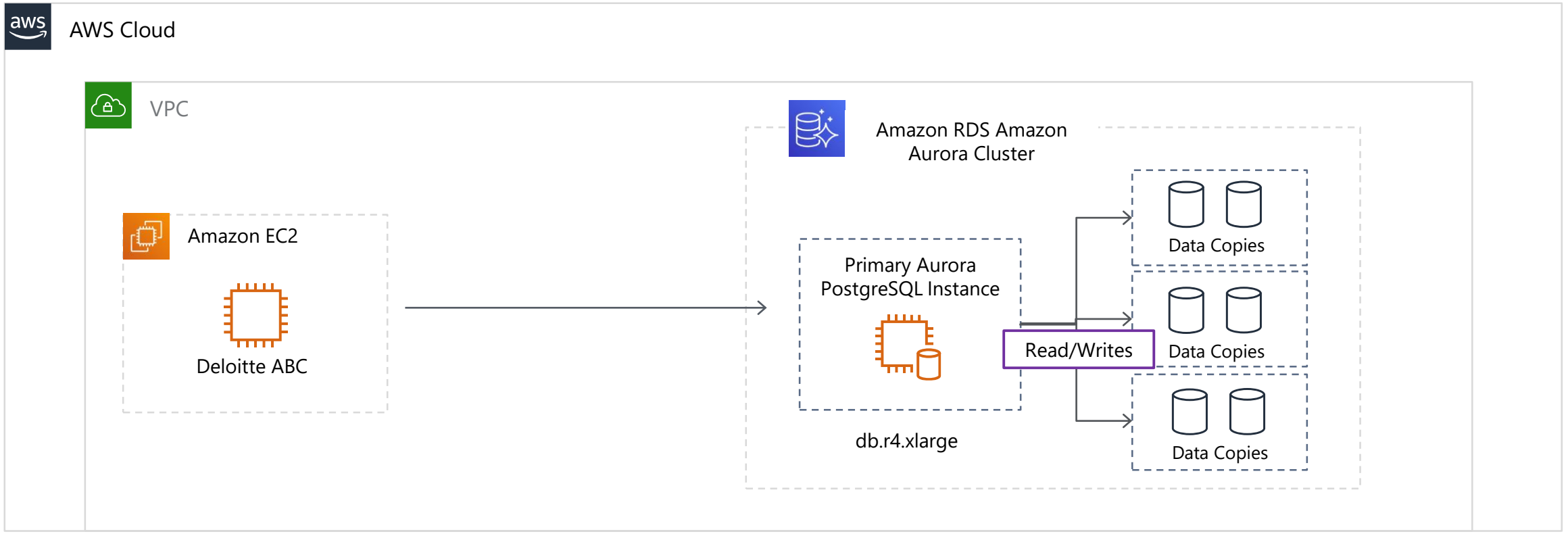


spectrumNodeType : dc2.large

spectrumClusterType: multi-node

Publicly accessible: No

Amazon RDS Amazon Aurora stack diagram



Deloitte's Audit-Balance-Control (ABC) Framework uses Amazon Aurora Postgres to write errors and logs entries for every data movement with the platform in an asynchronous manner

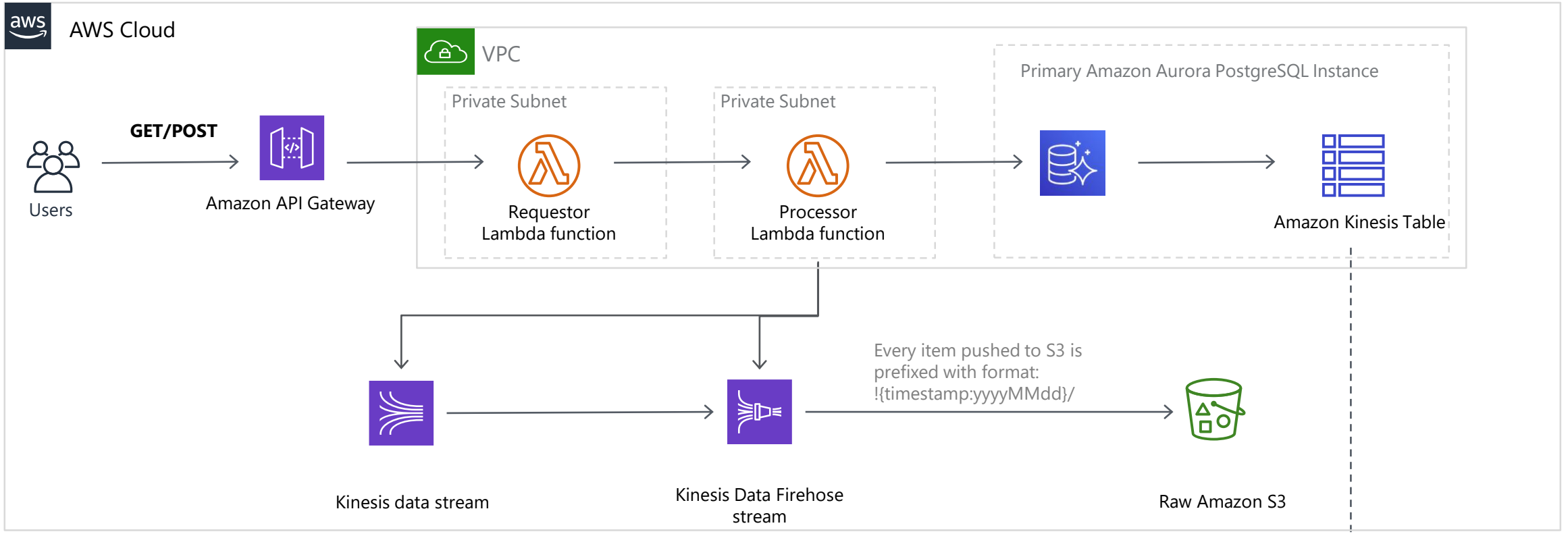
DBInstanceClass: db.r4.xlarge

DbEngineVersion: 9.6.9

MultiAZ: True

RDSDBParameterGroup: Customer CFT

Amazon Kinesis API stack diagram



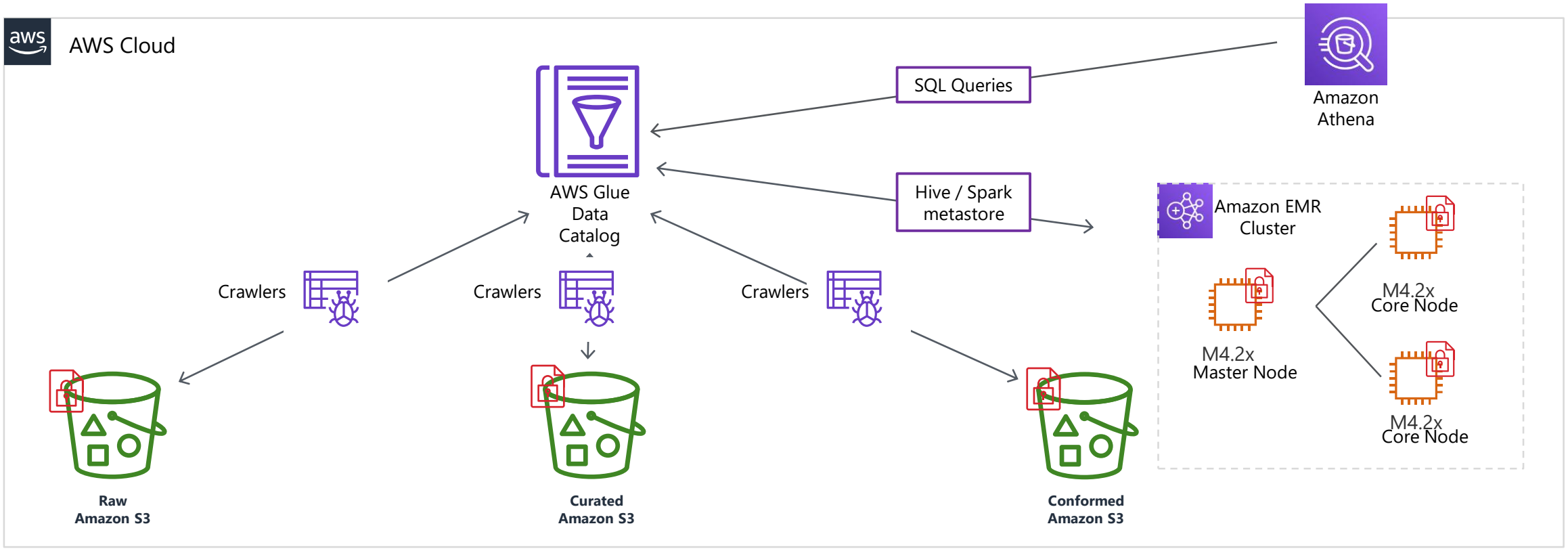
POST or GET calls to API Gateway invoke Requestor Lambda

Requestor Lambda invokes the Processing Lambda.
This Lambda creates the appropriate Kinesis data stream, delivery stream

The information of the streams being created is updated in the PostgreSQL Amazon RDS DB

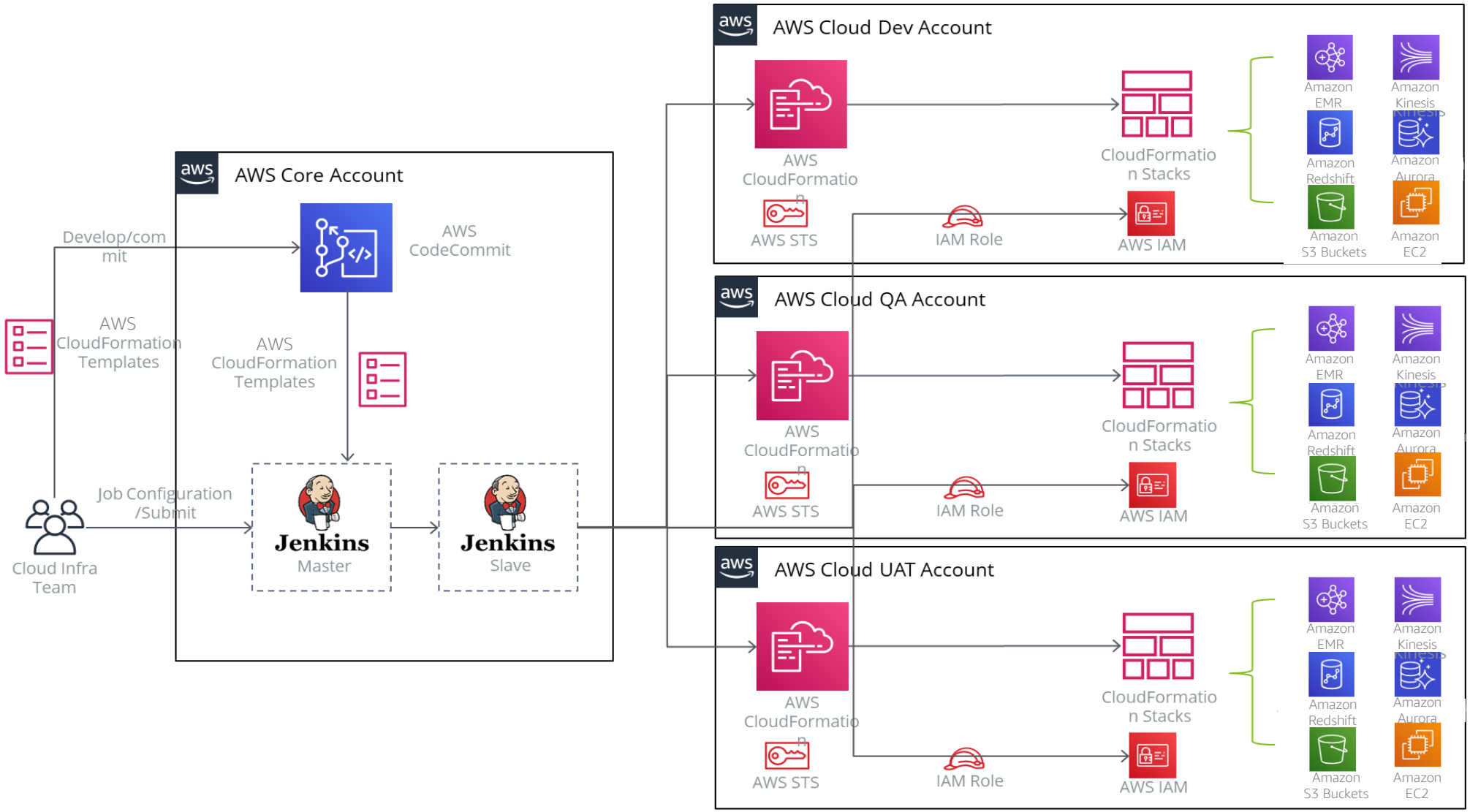
kinesis_stream_name	kinesis_firehose_stream_name	s3_bucket_name	status	timestamp
logs_dataplatform_abc	logs_dataplatform_abc			
algo_dataplatform_xyz	algo_dataplatform_xyz			

AWS Glue stack diagram



S3 bucket naming structure:
environment-platform-appID-type-clientname-ID-region
S3 encryption: AES-256 (SSE-S3)
Amazon EMR cluster encryption: KMS key
AWS Glue database naming structure: environment_platform_appID_SeqNum
Amazon Athena will use the AWS Glue catalog for querying

Infrastructure automation



Demo

Key learnings

Leverage AWS native services to enable a more seamless deployment (AWS pipeline and deploy)

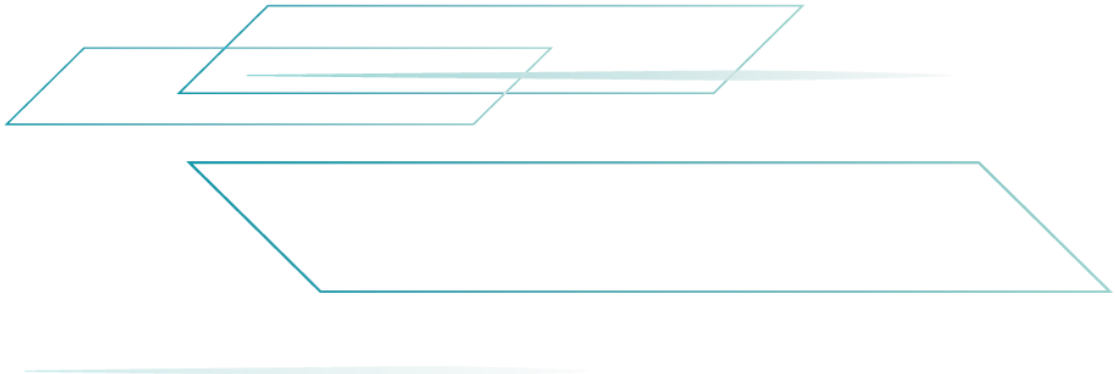
Amazon RDS and Amazon Redshift don’t enforce encryption in transit by default, and effort for encryption needs to be factored in

Streamline instance type to leverage reserved instance (RI) discount for better cost optimization

Don’t hesitate to deep dive and ask questions or seek help; e.g., Customize AWS Secrets Manager password spec

Key takeaways

Benefits realized



Improvements realized

Stood up a modern data and analytics platform on AWS to create a consolidated data lake for value and insights generation from a large amount of data being produced from Hertz digital transformation and new systems

Enabled enhanced reporting and visualization of cross-functional data to support multiple business functions (sales, marketing, ops, customer care, etc.)

Created a platform for cross-application integration and cross-functional insights, and delivered on Hertz's analytical needs

Onboarded selected high-value data science use cases (e.g., fleet optimization, capacity substitution model, net price, etc.) on the data platform that needed scalable compute

Thank you!

AWS re:Invent

aws

Deloitte.

Please complete the session survey in the mobile app.