AWS re:Invent

ANT327

Best Practices to Secure Data Lake on AWS

Varun Rao Bhamidimarri
Solution Architect
AWS

Tony Nguyen
Senior Big Data Consultant
AWS

AWS re:Invent

aws

What to expect

1

Understand

the value proposition for the data lake and how Amazon Web Services (AWS) can help

2

Get a sense

of some best practices for secure data lake implementation

3

Dive deep

into role/scenario based approaches to data lake security

Assumptions and Housekeeping

- Targeted towards anyone wanting to build a secure data lake on AWS
- Assumes:
  - Foundational AWS knowledge
  - High level knowledge of Data Lake and AWS Analytics service portfolio
  - Knowledge of security concepts such as SSL / TLS, encryption, authentication / authorization
- This session slides and recording will be shared online
- Please don’t forget to submit your feedback!

What is a data lake?

- Collect and store all data, at any scale, and low cost
- Helps locate, curate, and secure your data
- Provide democratized access to data within your organization
- Quickly and easily perform new types of data analysis

Primary components of a data lake

1. Automated and reliable data ingestion
2. Preservation of original source data
3. Lifecycle management and cold storage
4. Metadata capture
5. Managing governance, security, privacy
6. Self-service discovery, search, access
7. Managing data quality
8. Preparing for analytics
9. Orchestration and job scheduling
10. Capturing data change

Attributes of a modern data architecture

API & UI

Entitlements

Catalogue & search

Storage & streams

Key pillars of a data lake

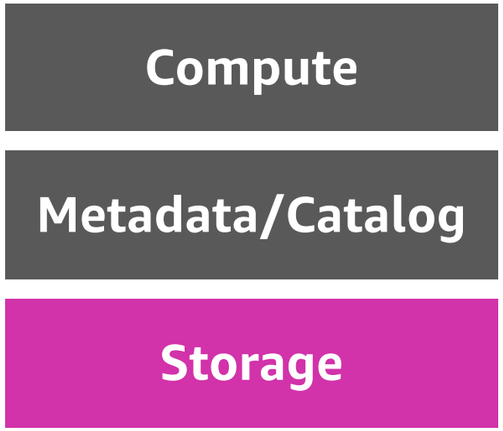
Compute

Metadata/Catalog

Storage

Architectural layers of a Data lake (without security)

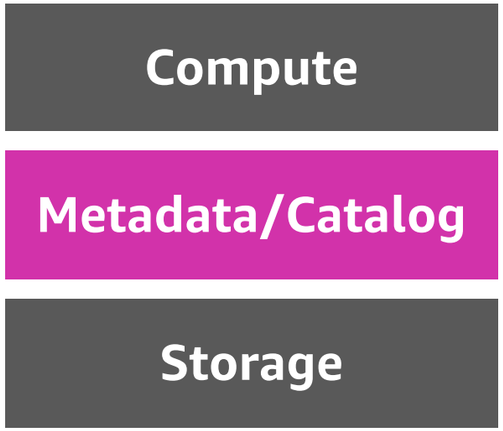
Primary components of a data lake



Architectural layers of a data lake (without security)

- **Object** storage – Amazon S3/Amazon Glacier
- **Block** storage – Amazon Elastic Block Store (Amazon EBS)
- **File** storage – Amazon Elastic File System (Amazon EFS)
- **Attached** instance store
  - Amazon EC2 instance
  - Amazon Redshift clusters
- Also need to consider perhaps not as obvious services such as Amazon Kinesis and Amazon DynamoDB

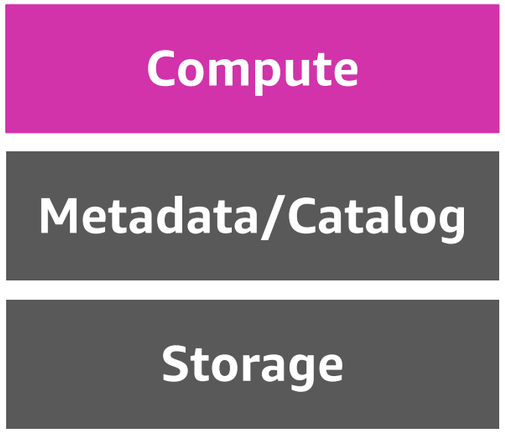
Primary components of a data lake



Architectural layers of a data lake (without security)

- Automatically **index** data
- Easy **search** with tags/business domain
- **Curate** and assign relevancy score
- Easily **commission** and **decommission** data sets
- Capture data **lineage**

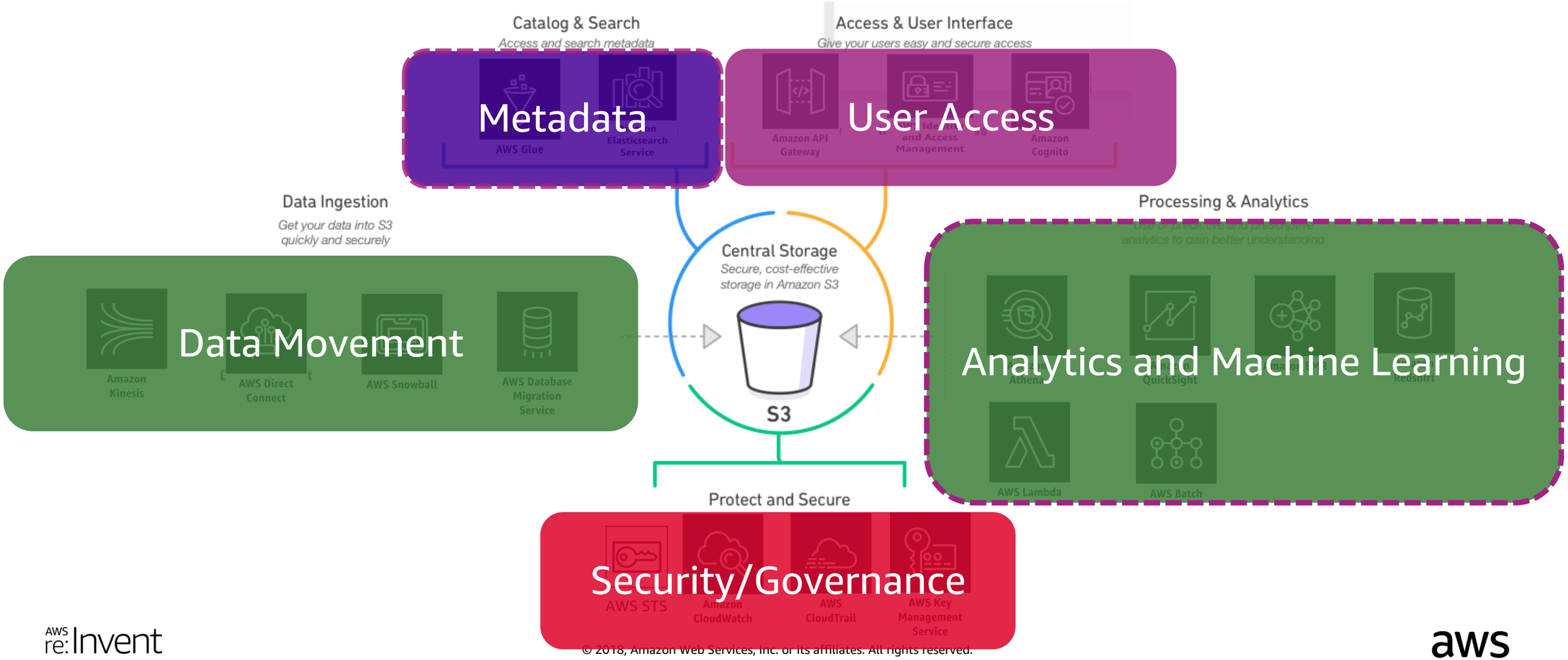
Primary components of a data lake



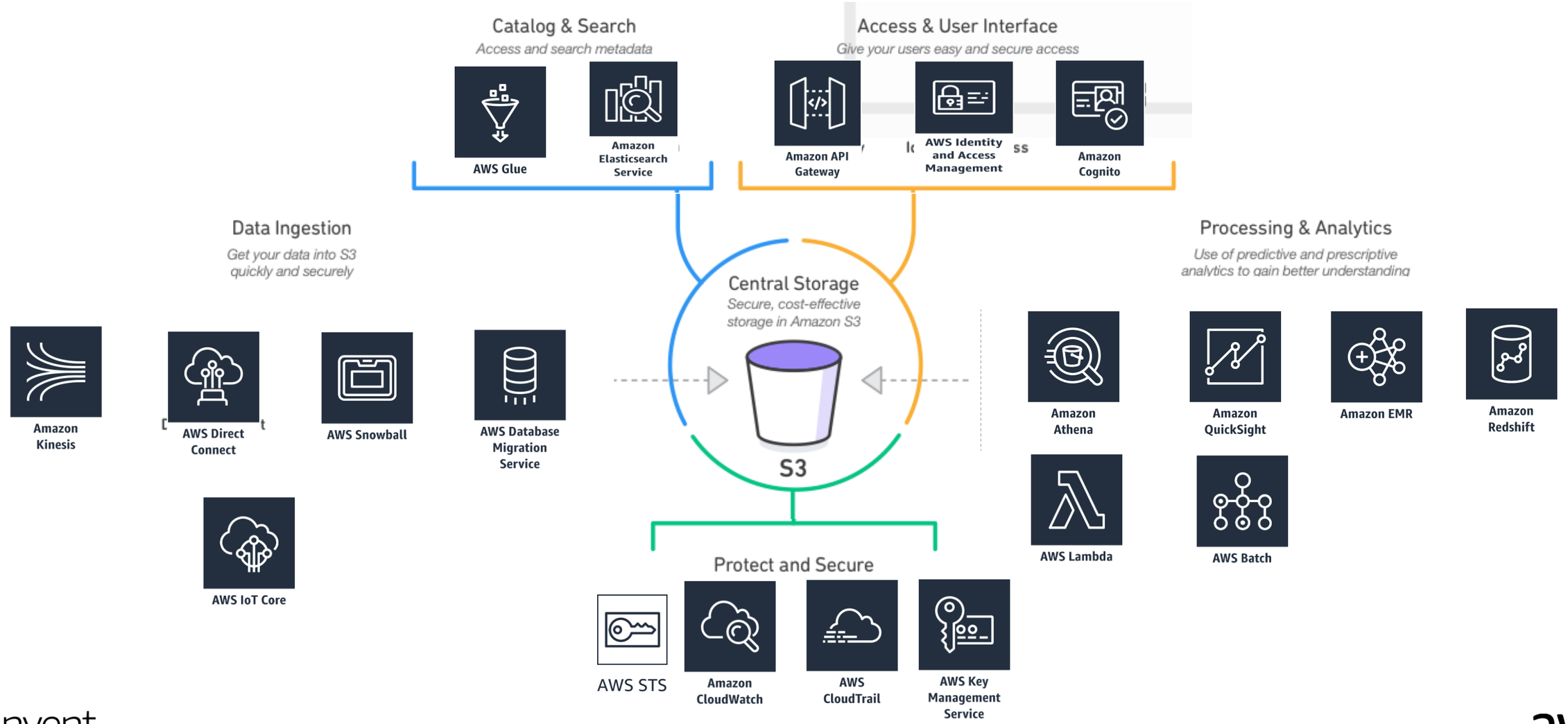
Architectural layers of a data lake (without security)

- **Server-based** compute
  - More than just standalone Amazon Elastic Compute Cloud (Amazon EC2), also includes Amazon EMR, Amazon Redshift
- **Serverless** compute (AWS Lambda, Amazon Athena, Amazon API Gateway, and others)
- **Hybrid**
  - Amazon Redshift Spectrum

Building a data lake on AWS



Building a data lake on AWS



Securing all of these tools is challenging

- Having such a **diverse set of tools** from the ecosystem allows you to choose the best tool for the job…
- …but also makes a **single unified solution** for security challenging!
- How do you secure each layer, while still **satisfying your specific security and compliance requirements**?

What's required for a secure data lake?

Security challenges with data lakes

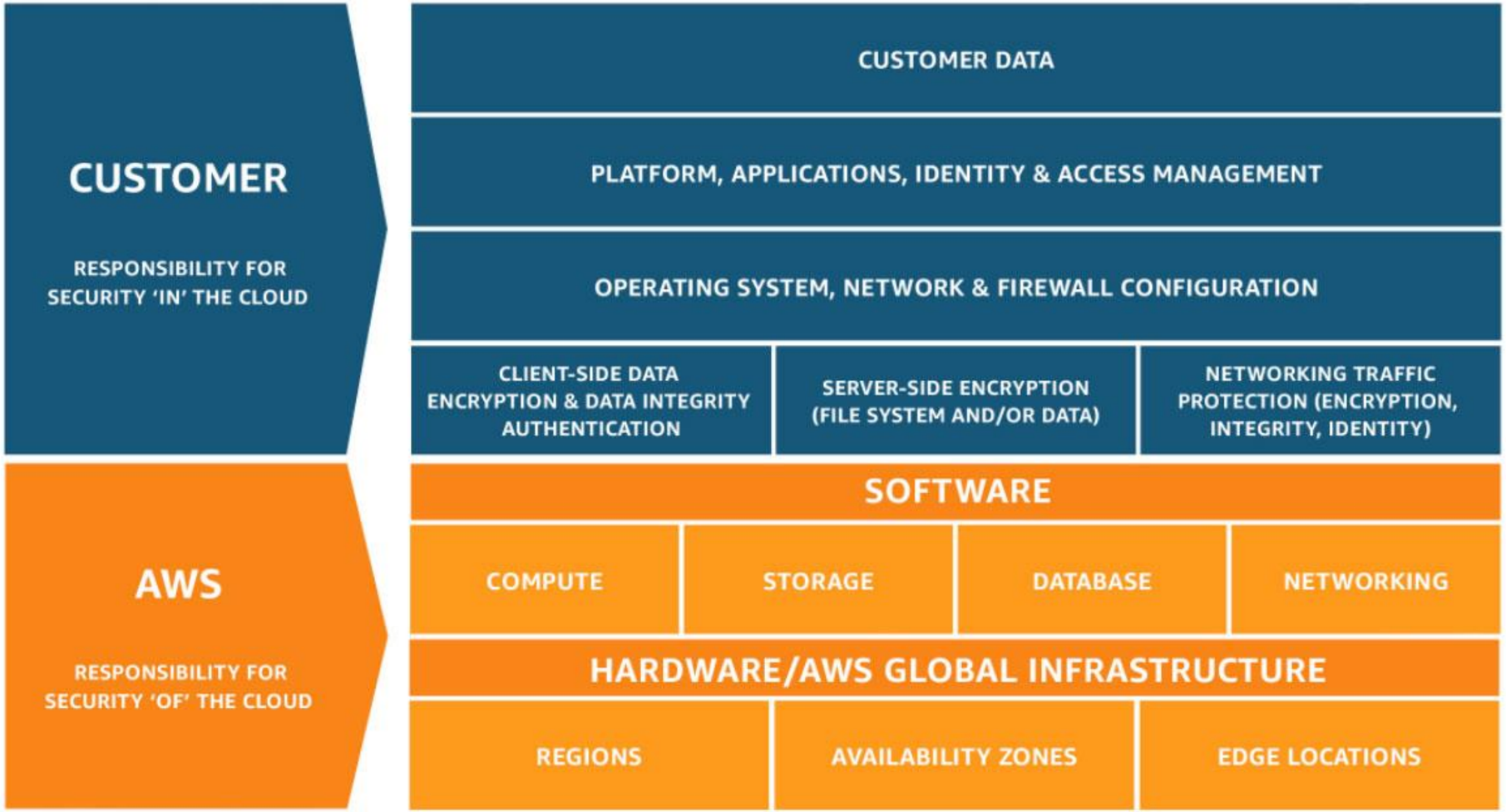
Data challenges

- Controlling access to data
  - Data masking, row / column / cell level encryption, key management
- Data loss / exfiltration
- Loss of data integrity
- Data provenance
- Compliance requirements (GDPR and others)

Management challenges

- Central administration
- Federated authentication, typically with Active Directory
- Role-based access control (RBAC)
- Centralized audit
- End-to-end data protection (at-rest and in-transit)

Shared responsibility model



Shared responsibility model – service types

- Infrastructure Services (EC2, EBS)
  - Rich control, similar to on-premises (this control might be via API). Most customer responsibility
  - Separation of control plane and data plane
- Managed Services (EMR, RDS, Redshift)
  - Services that are deployed for you on top of EC2
  - Control plane and data plane are separate, but there is joint control (and therefore joint responsibility)
- Serverless Services (S3, DynamoDB, Athena, Glue)
  - Services that are network endpoints that respond to commands, generally a unified control and data plane
  - Least customer responsibility – typically controlled only by IAM

Shared responsibility model – comparison

Amazon EMR

- Amazon EC2 infrastructure needs to be managed
- Root-level access via SSH
- Patching of instances
- Some level of Amazon CloudWatch / Amazon CloudTrail logging is done for customer, but not exhaustive
- Instance profile role, Amazon EMR Service role need to be configured by customer
- Local disk encryption, Amazon S3 encryption, etc. needs to be configured by customer…

Amazon Athena

- No infrastructure to manage
- Service access is governed via IAM policy documents
- Amazon S3 access is via bucket policy / IAM policy
- Encryption is managed

Let’s start at the foundation

AWS re:Invent

aws

AWS helps you secure

Customers need to have multiple levels of security, identity and access management, encryption, and compliance to secure their data lake

Security

Amazon GuardDuty

AWS Shield

AWS WAF

Amazon Macie

Amazon Virtual Private Cloud (Amazon VPC)

Identity

AWS Identity and Access Management (IAM)

AWS Single Sign-On

Amazon Cloud Directory

AWS Directory Service

AWS Organizations

Encryption

AWS Certificate Manager

AWS Key Management Service (AWS KMS)

Encryption at rest

Encryption in transit

Bring your own keys, HSM support

Compliance

AWS Artifact

Amazon Inspector

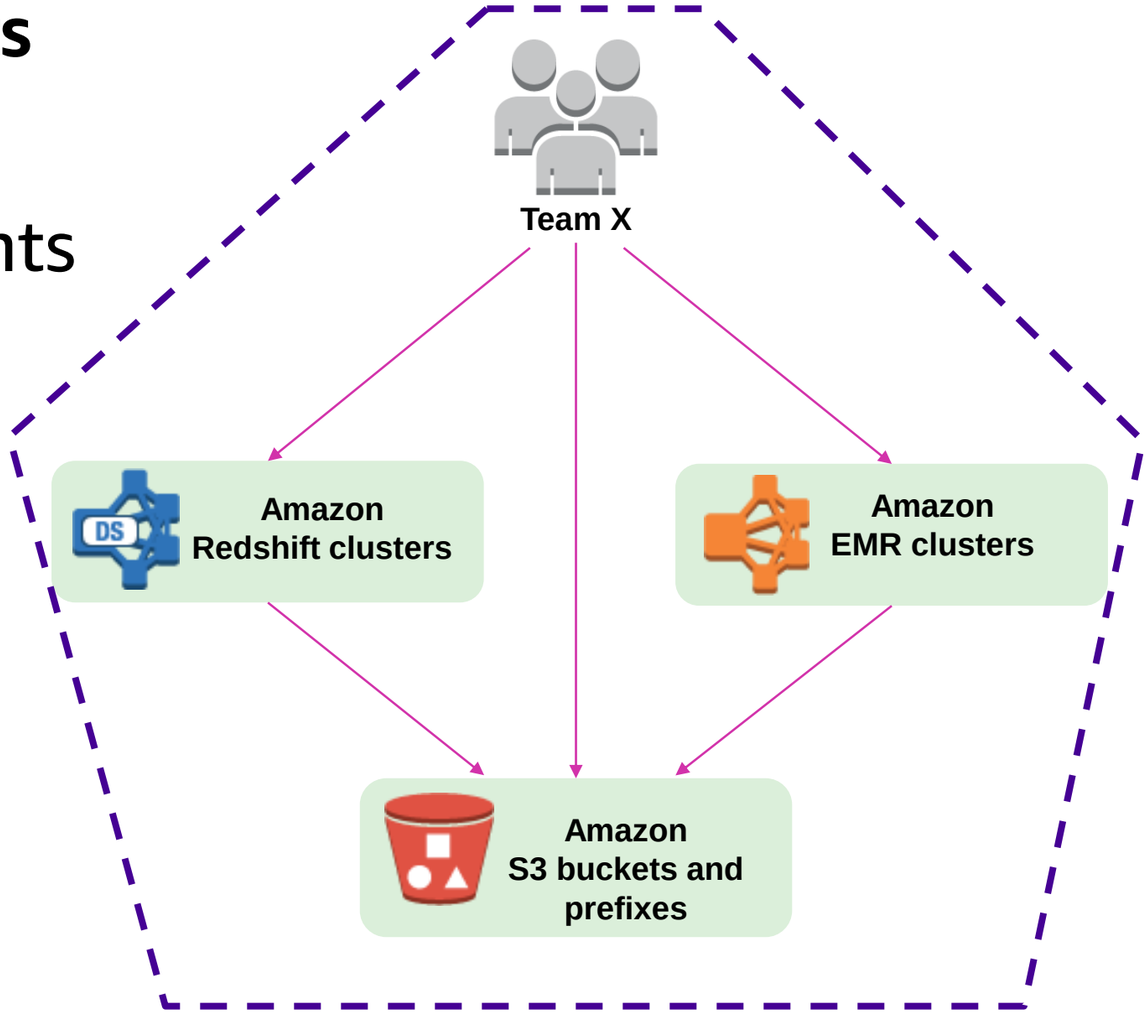
AWS CloudHSM

Amazon Cognito

AWS CloudTrail

Prefer “coarse-grained” ownership

- Teams own **entire Amazon S3 buckets and clusters**
- Ownership segregated by AWS accounts
- Access control easier to setup and maintain
- Suitable for **autonomous teams**



“Coarse-grained” ownership

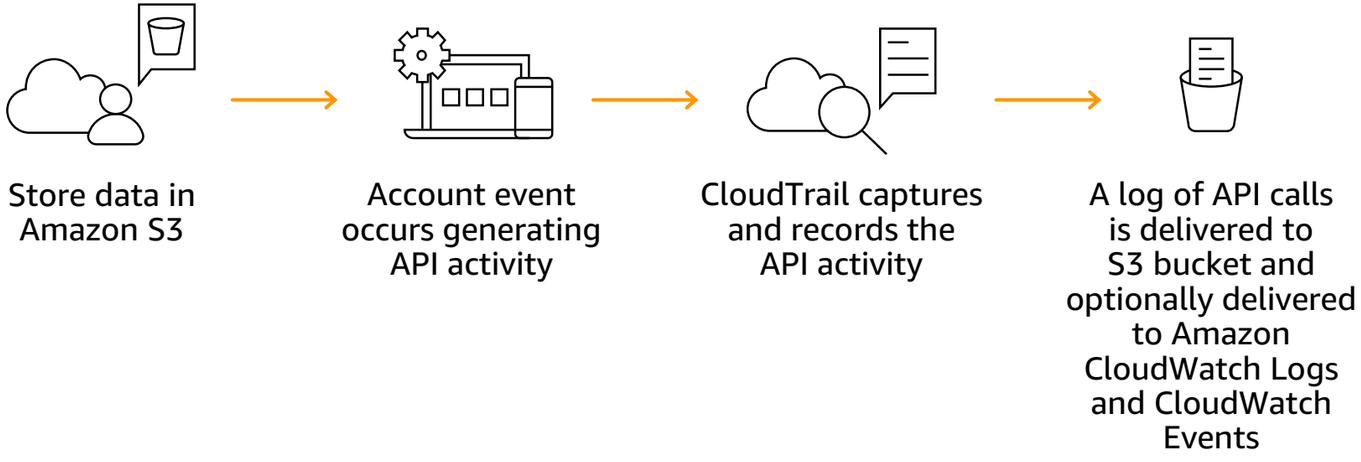
Encrypt data at rest

Pick encryption mode for Amazon S3 objects

Encryption point	AWS KMS	Custom KMS	S3 built-in
Client side	CSE - KMS	CSE - C	
Server side	SSE - KMS	SSE - C	SSE - S3

Key Management overhead

Compliance: Log and audit all AWS activity



- Log and continuously monitor every account activity and API calls with Amazon CloudTrail
- Increase visibility into your user and resource activity
- Log management and data events into separate trails
- Centralize logs into separate security account
- Disable S3 delete using IAM

Security in the cloud - basics

- Account
  - Federate accounts with Active Directory / Identity Provider
  - Setup multi-factor authentication (MFA)
  - Avoid using root account credentials
  - IAM access should be least privilege
- Network
  - Private VPC Subnets
  - VPC endpoint/Interface endpoints
  - Least privilege for Security groups
- Storage
  - Encrypt using KMS

Data workflow

Different types of roles

Security Admin

Data Curator

Data Lake on AWS

Analyst

Data Engineer

Data scientist

Data workflow

- All roles have actions and responsibilities that correspond to each phase in the overall data workflow
- Think of the data lake in terms of producers and consumers

What data do I have?

"Through 2018, 80% of data lakes will not include effective metadata management capabilities, making them inefficient."

-Gartner

Onboarding new data

Data Owner

Developer / Data Engineer

Developer / Data Engineer

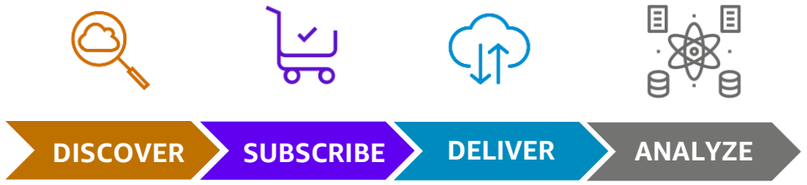
Data Curator

Data Catalog

Amazon S3

AWS Glue

Searching and accessing data



Data Scientist / Business User

Data Scientist / Analyst

Data Owner/Security Admin

Data Scientist / Business User

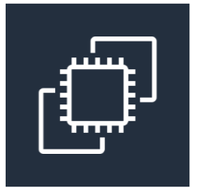
Data Catalog

Amazon S3

Amazon QuickSight

Amazon Athena

Amazon Redshift



Amazon EC2

Amazon SageMaker

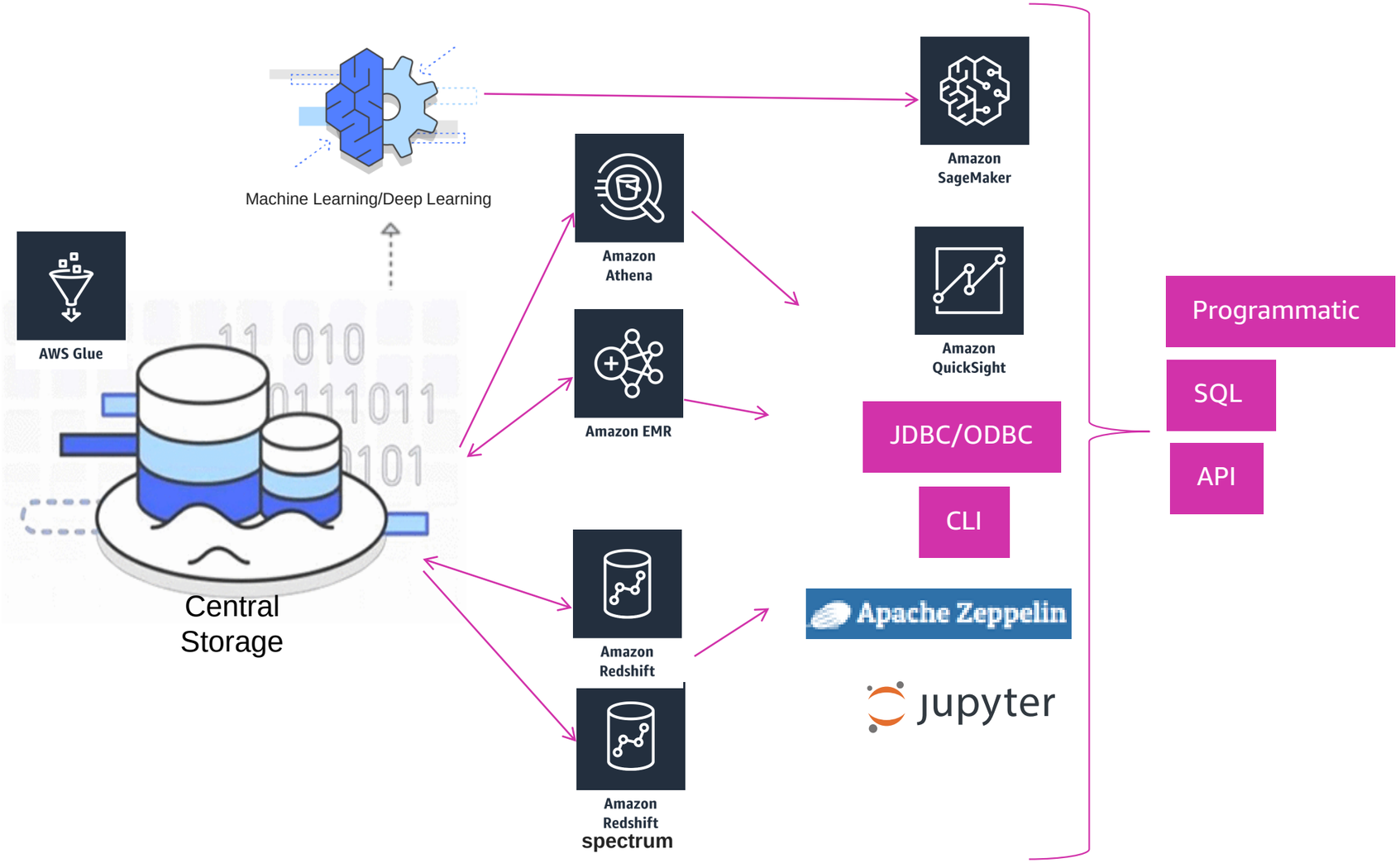
AWS Deep Learning AMIs

AWS re:Invent

aws

Security Admin

Data analytics tools and access patterns



Security admin tasks

- Setup security guardrails
  - Preemptive and detective controls
- Provide data access across teams/environments
  - Validate security requirements based on data classification
  - Verify Data owner/producer has authorized access
- Run regular audits

Amazon S3 – preemptive controls

- Create buckets based on business domains
- Assign bucket policies
  - Restrict by VPC, HTTPS, IP filters, KMS keys
- Restrict using Tags/Conditions
  - "Condition": {"StringEquals": {"S3:ResourceTag/HIPAA":"True"}
  - "Condition": {"StringEquals": {"aws:UserAgent": "AWS Redshift/Spectrum"}
- Enable encryption/Enable versioning
- MFA delete
- Enable backups – across accounts/regions
- IAM permission boundary
- S3 public access setting NEW!

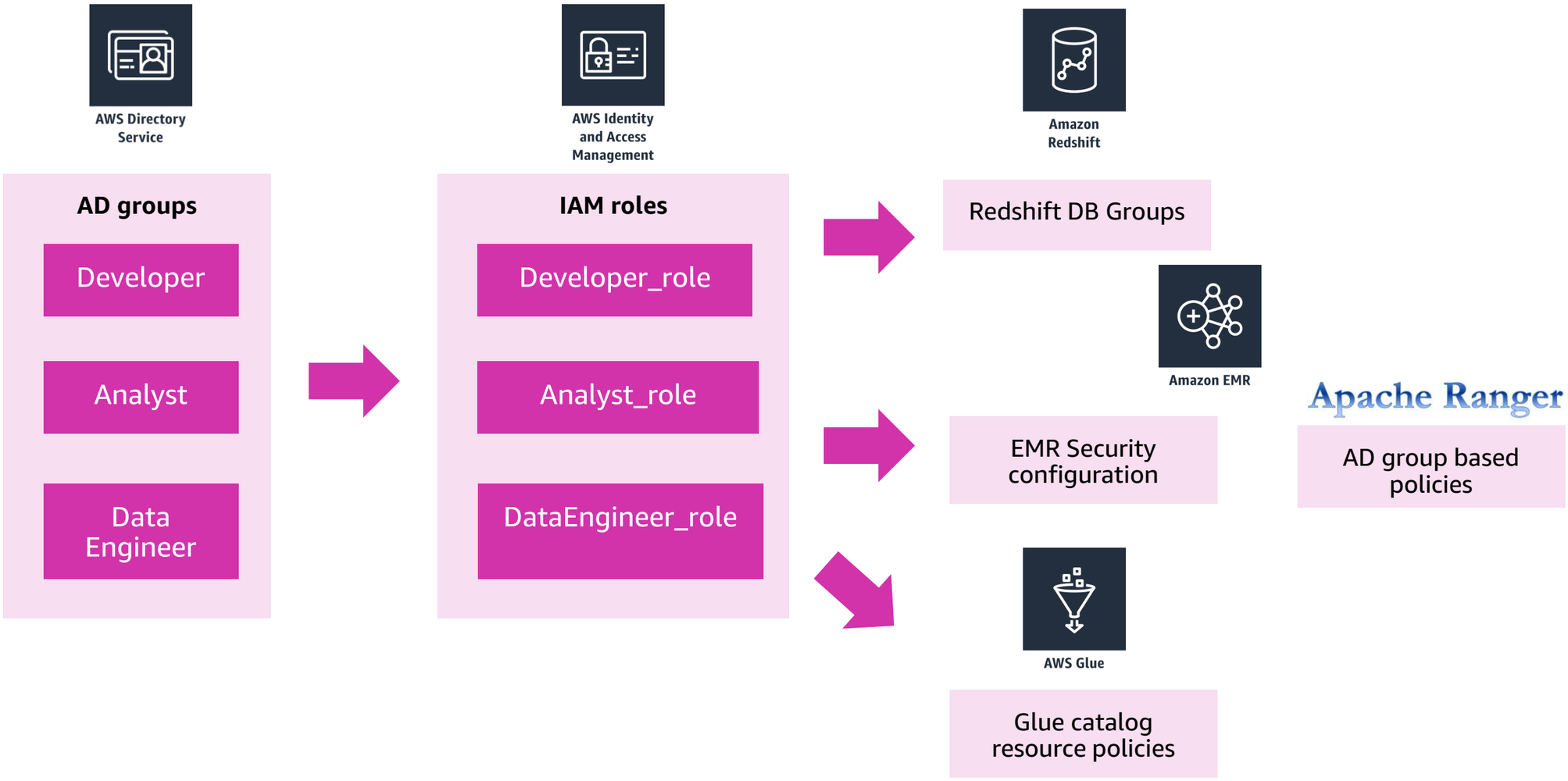
Amazon S3 data – detective controls

- Enable AWS Config to detect S3 bucket level changes
  - **s3-bucket-public-read-prohibited, s3-bucket-public-write-prohibited**, **s3-bucket-ssl-requests-only**
- S3 data access audit using CloudTrail – Log to separate CloudWatch logs
  - Kerberos enabled EMR clusters allows you to track AD user
- Use Amazon GuardDuty to detect unauthorized and unexpected activity
- Enable Amazon Macie to classify sensitive data

Encrypt data in transit

Point "A"	Point "B"	Data flow protection
Enterprise data sources	Amazon S3	**Encrypted with SSL/TLS**; S3 requests signed with **AWS Sigv4**
Amazon S3	Amazon EMR	**Encrypted with SSL/TLS**
Amazon S3	Amazon Redshift	**Encrypted with SSL/TLS**
Amazon EMR	Clients	**Encrypted with SSL/TLS;** varies with Hadoop application client
Amazon Redshift	Clients	**Supports SSL/TLS;** Requires configuration
Apache Hadoop on Amazon EMR		• **Hadoop RPC encryption** • **HDFS Block data transfer encryption** • **KMS over HTTPS is not enabled by default with Hadoop KMS** • May vary with EMR release (such as Tez and Spark in release 5.0.0+)

Security authorization mapping



Map database ACL's to db grant/glue policy

catalog.**user_table**

☑ select ☐ insert

AD group: developer

Database grants

grant group developer select on catalog.**user_table**

Glue catalog

```
Action: ['glue:GetTable*', 'glue:GetPartiton*']
Principal: ["arn:aws:iam::<account>:role/developer_role"]
Resource:["arn:aws:glue:<region>:<account>:table/gluecatalog/user_table",
"arn:aws:glue:<region>:<account-id>:table/gluecatalog/user_table/*"]
```

Map storage ACL's to Amazon S3 policy

s3://bucket/path/

☑ read ☑ list ☐ write

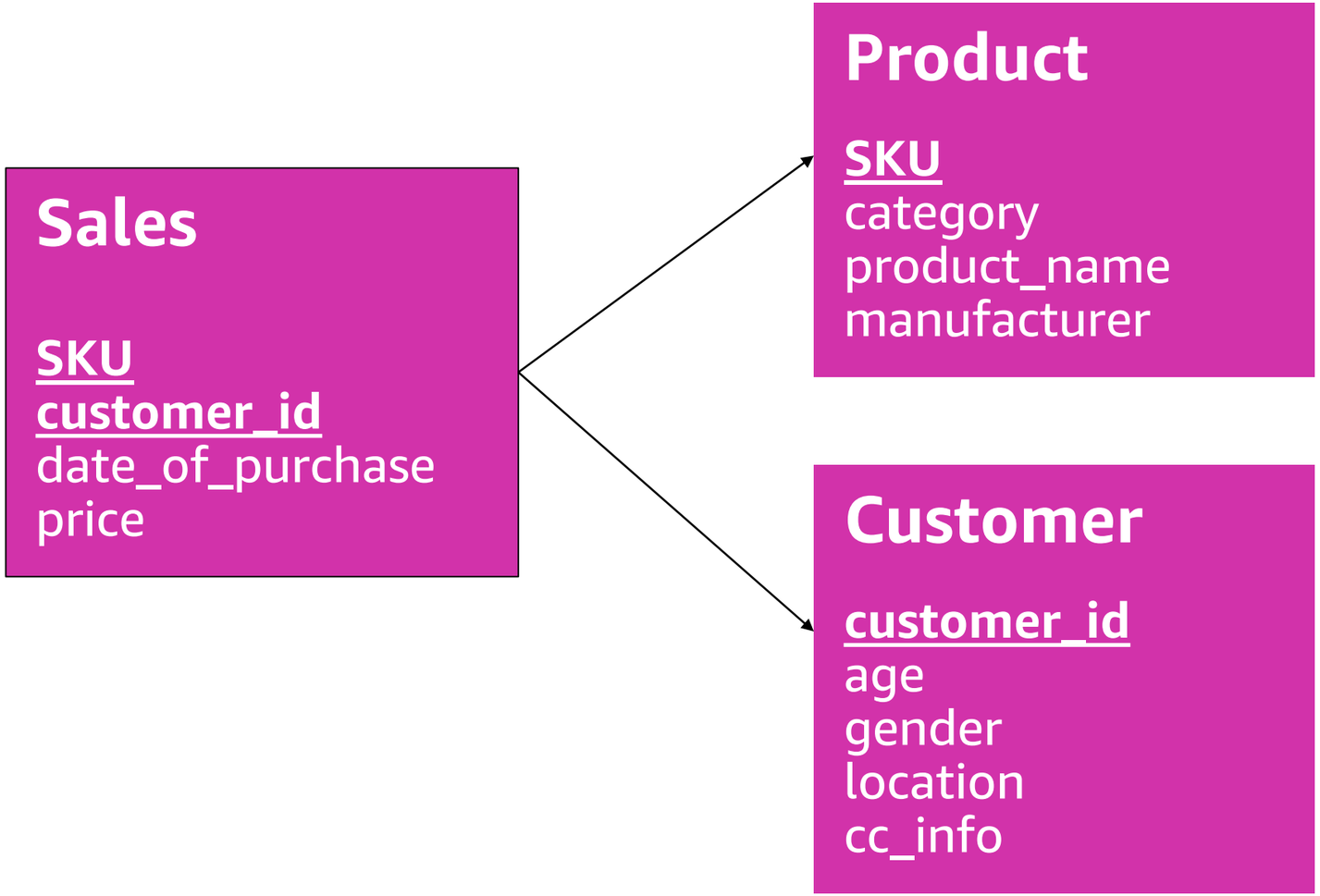
AD group: developer

S3 bucket policy

```
Effect: Allow
Action: [s3:ListBucket', "s3:GetObject"]
Principal: ["arn:aws:iam::<account>:role/developer_role"]
Resource: ["s3://bucket/path", "s3://bucket/path/*"]
```

Let’s take a customer scenario

Scenario – retail company X



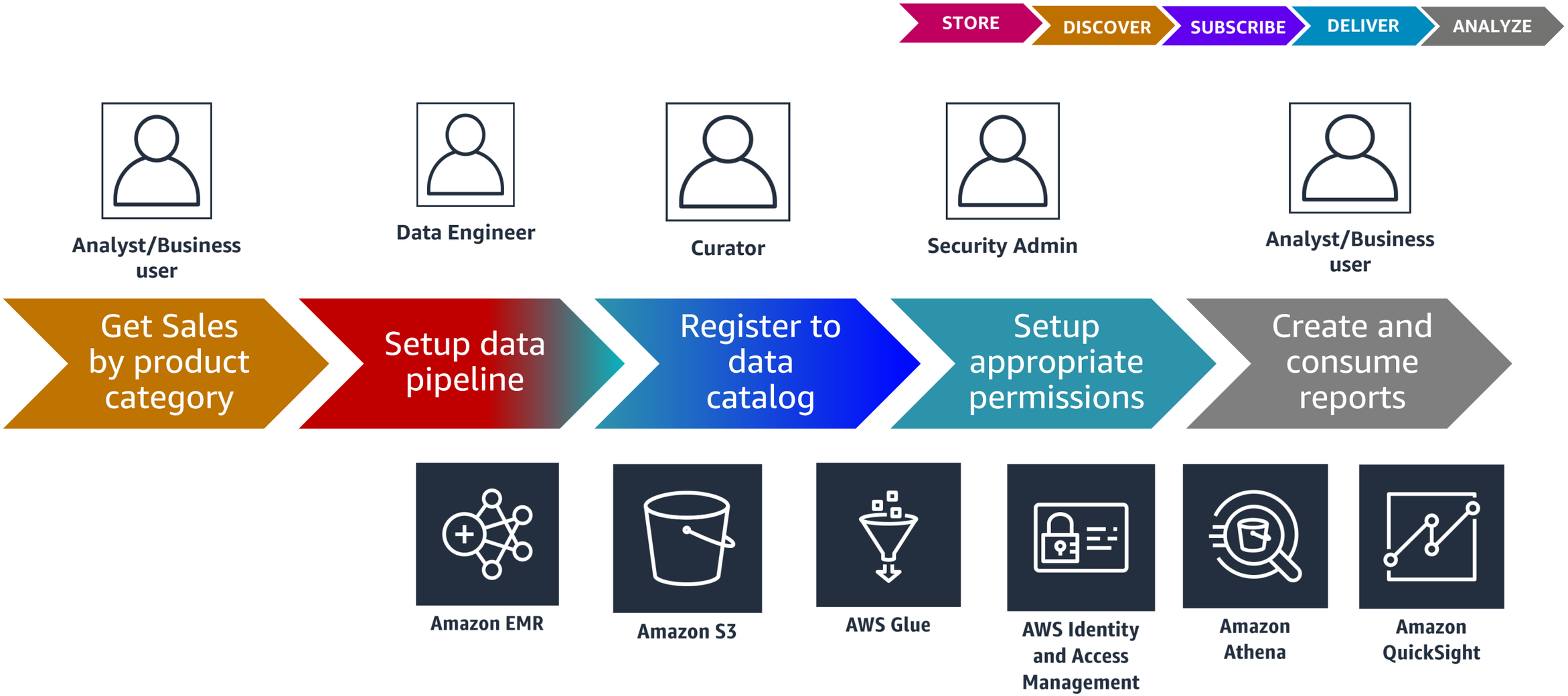
Gather insights from the data

- **Business user (External vendor – belongs to a manufacturer)**
  - Sales by product ***category*** (cannot see other manufacture's data)
  - Sales by ***location***
  - Get sales ***forecast*** by **product**

- **Analyst (Employee –** may belong to a Product line/Business unit**)**
  - Sales by product ***category***
  - Sales by ***location***

- **Data scientist (Employee –** may not belong to a Business Unit**)**
  - **Forecast** the sales of a specific product, based on age group, location and time of the year

Workflow – onboarding new data



Role based tasks

Curator

Setup staging catalog

Enable access to Data Engineer

Verifies and commissions dataset to production catalog

Data Engineer

Setup Amazon EMR cluster

Setup process to move data from source into Amazon S3

Orchestrate and schedule the job

Security Admin

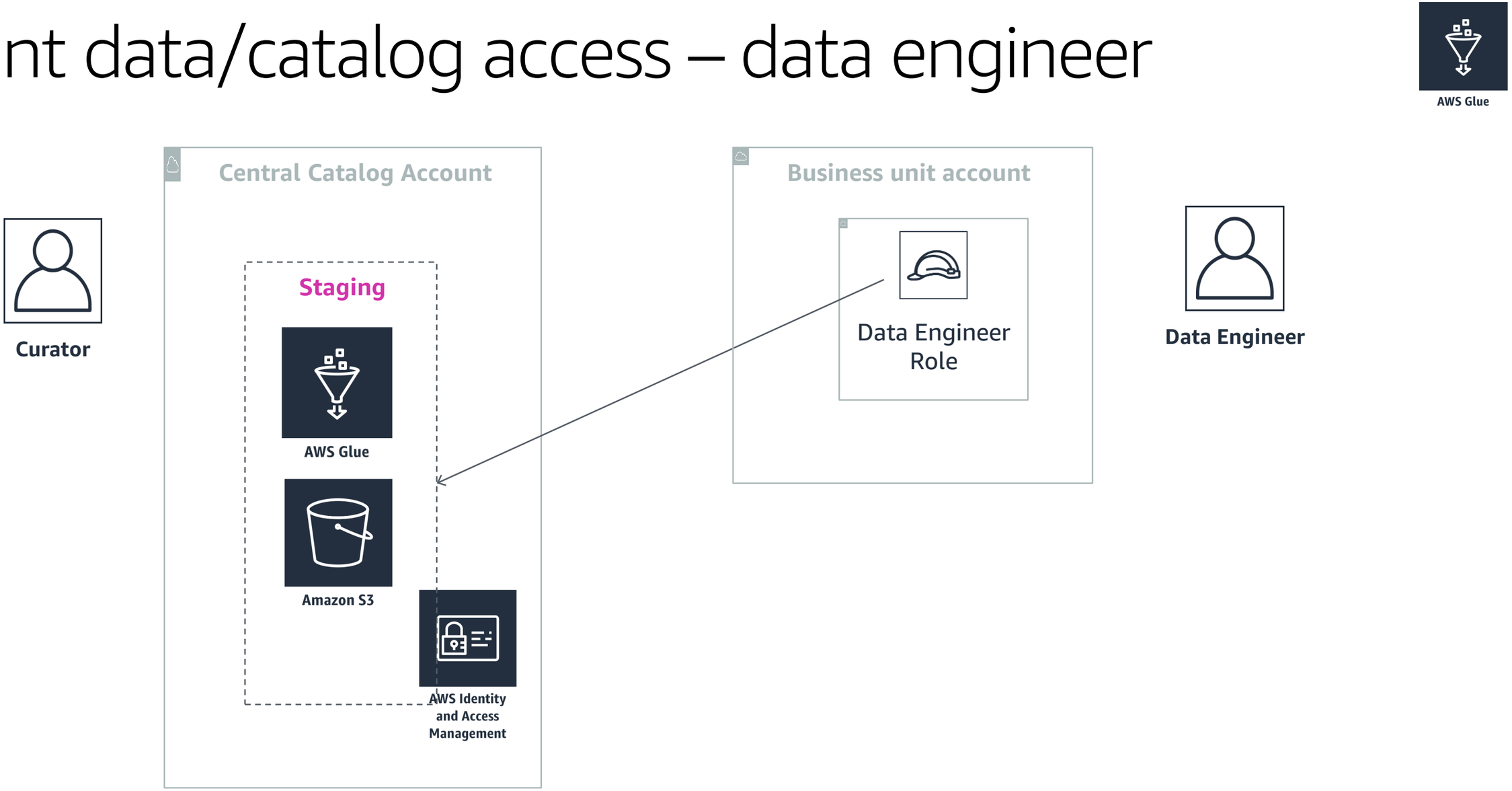
Enable access to Analyst

Setup Row-level security for business users

Analyst

Create and publish dashboard

Grant data/catalog access – data engineer



Onboarding new data – security/configuration

AWS Glue

Catalog policy

```
Effect: Allow
Action: ['glue:*Database*’, ‘glue:*Table*’,’glue:*Partition*’]
```

Storage grants

```
Effect: Allow
Action: ['s3:PutObject', 's3:GetObject', 's3:DeleteObject']
```

Amazon EMR Configuration

```
  "Classification": "spark-hive-site", "Properties":
  {
   "hive.metastore.client.factory.class":
"com.amazonaws.glue.catalog.metastore.AWSGlueDataCatalogHiveClientFactory",
   "hive.metastore.glue.catalogid": "acct-id"
  }
```

AWS Glue catalog - resource policies

AWS Glue

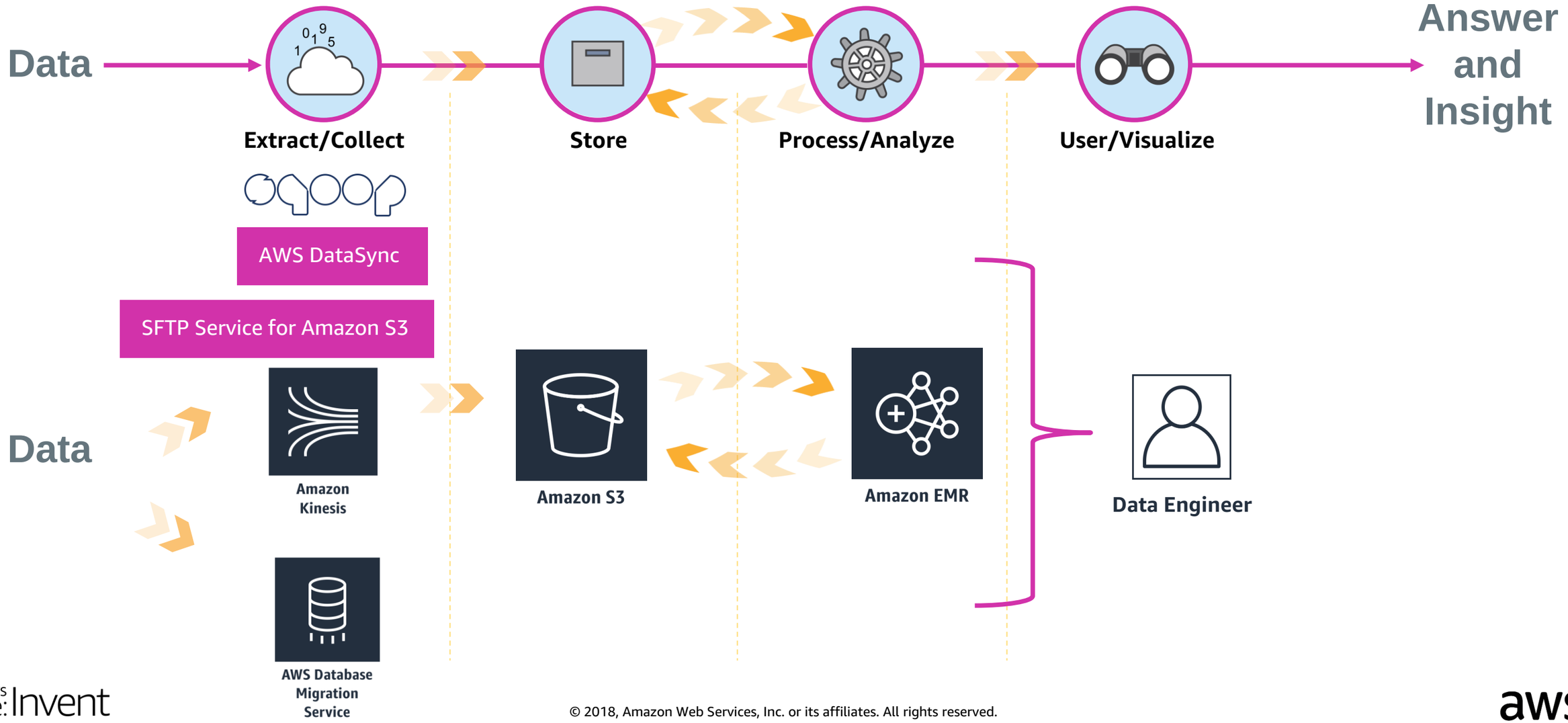
- Fine-grained access control to Catalog using IAM policies
- Restrict what they can view and query

```
"Action": [
    "glue:*Database*",
    "glue:*Table*",
    "glue:*Partition*"
],
"Resource": [
    "arn:aws:glue:us-east-1:[redacted]:table/blog_dev/*",
    "arn:aws:glue:us-east-1:[redacted]:database/blog_dev",
    "arn:aws:glue:us-east-1:[redacted]:catalog",
    "arn:aws:glue:us-east-1:[redacted]:userDefinedFunction/blog_dev/*"
],
```

```
"Action": [
    "glue:GetTable*",
    "glue:GetPartition*"
],
"Resource": [
    "arn:aws:glue:us-east-1:[redacted]:table/blog_prod/prod_*",
    "arn:aws:glue:us-east-1:[redacted]:database/*",
    "arn:aws:glue:us-east-1:[redacted]:catalog"
],
```

Build the data pipeline

Build the data pipeline – Amazon EMR

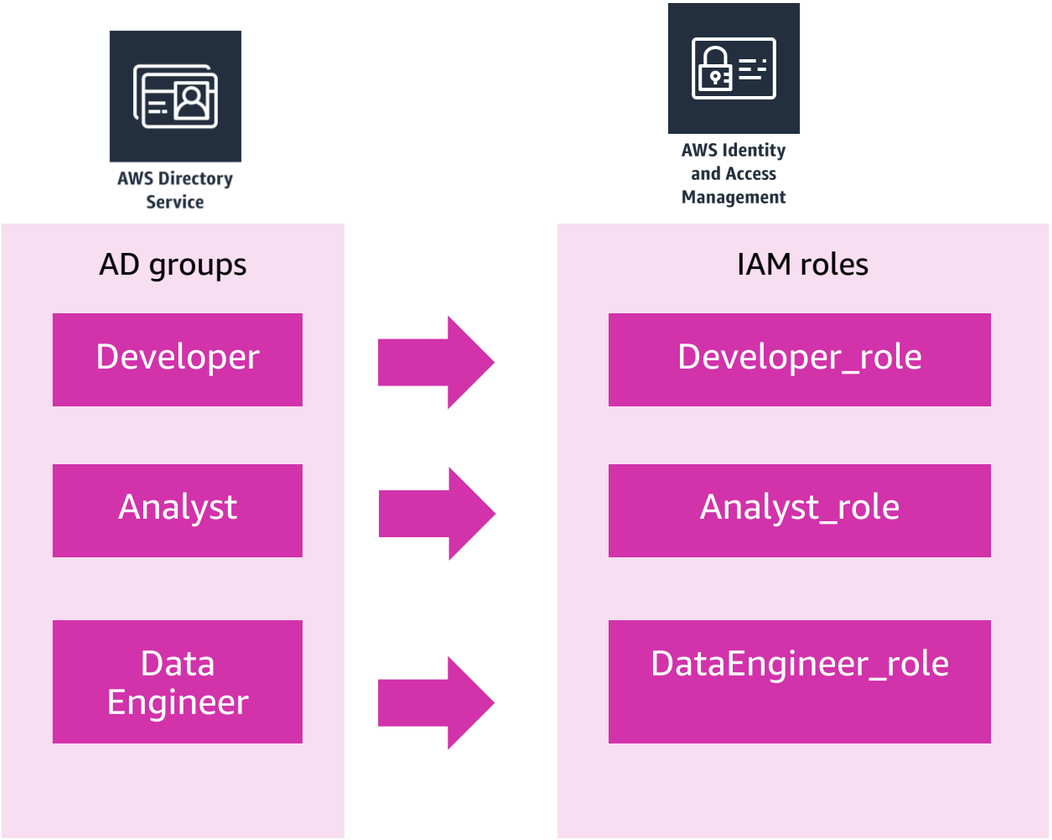


Amazon EMR - authentication

- Configure Kerberos for cluster authentication
- LDAP for HiveServer2, Hue, Presto, Zeppelin
- Perimeter security using Apache Knox
  - Simplify authentication of various Hadoop services and UI's
  - Mask service specific URL's/Ports by acting as a Proxy
  - Enable SSL/TLS termination at the perimeter
  - Ease management of published endpoints across multiple clusters
  - Supports federation

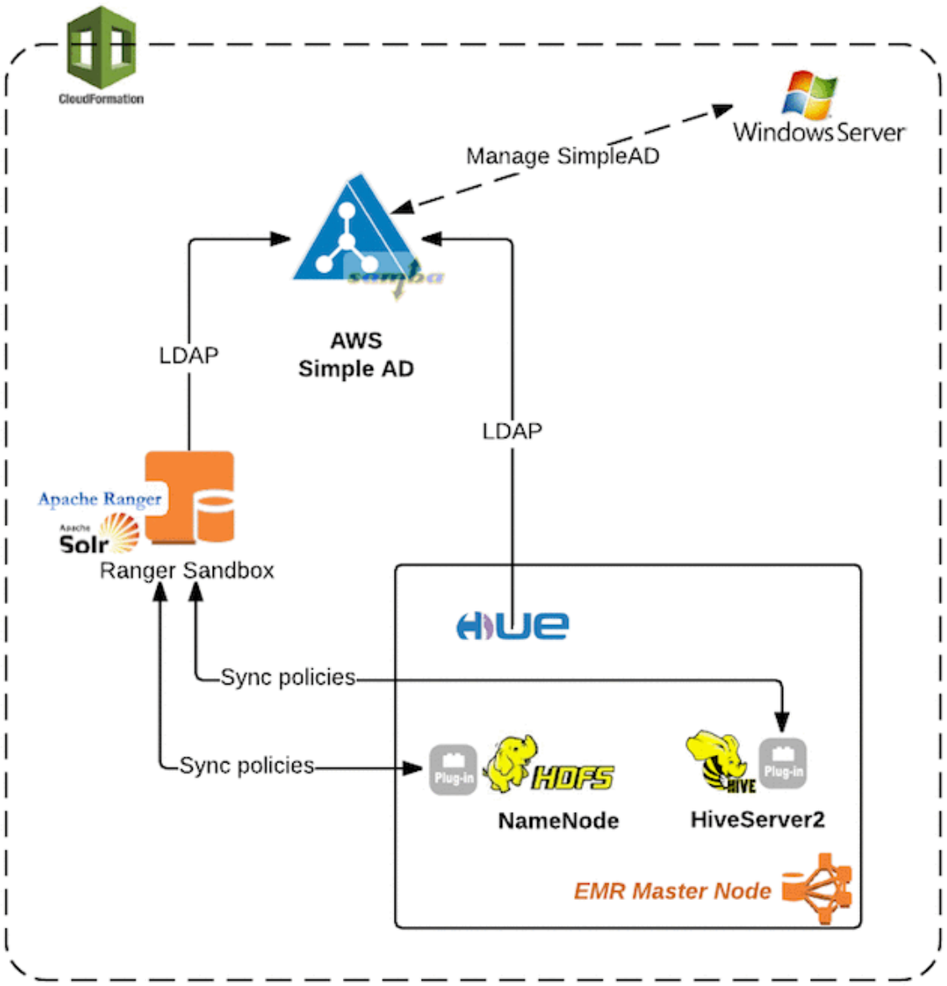
Amazon EMR – storage authorization

- Control access to Amazon S3 based on user's AD groups
- Use different IAM roles for EMRFS requests to Amazon S3
- These IAM roles can be mapped to users, groups or the location of data in Amazon S3.



Amazon EMR – service authorization

- **Apache Ranger** provides authorization of Hadoop cluster services
  - Eg: Hive tables, HDFS files, HBase etc
- Also provides Audits
- Column masking and Row filtering for Hive



Best practices - Amazon EMR security

Amazon EMR

Authentication

- Kerberos
- Knox, Shiro
- LDAP / AD integration

Authorization

- EMRFS storage AuthZ
- Apache Ranger
- Table and SQL-level authorization for Hive using HiveServer2
- Role-based Authorization with AD
- IAM

Audit

- Amazon EMR logs to Amazon S3
- Amazon S3 Access Logs
- Apache Ranger Audit
- Amazon CloudTrail – Amazon EMR API's/EMRFS calls

Data protection at rest

- SSE-S3 , SSE-KMS, Amazon S3 Client Encryption
- Disk encryption using AWS KMS
- SELinux using EMR BA
- Custom AMI

Data protection at motion

- SSL/TLS in transit using Security configurations
- SSL/TLS for calls to S3 (default)

Compliance Programs

- SOC1,2,3
- ISO
- PCI DSS
- FedRAMP
- HIPAA BAA
- DoD SRG IL2/IL4

Data ready - what next?

AWS re:Invent

aws

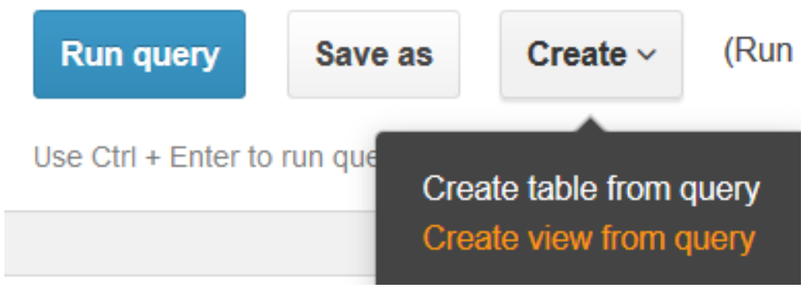
Data ready - what next?

- Curator
  - Verifies data registered with staging catalog
  - Runs sanity checks
  - Commissions the dataset into the production catalog
  - Creates a View to filter data by Product category
    - select * from sales join product where sales.sku = product.sku and category = ‘Electronics’
- Security Admin
  - Enable access to Analyst
  - Setup Row-level security
- Analyst
  - Create and publish dashboard

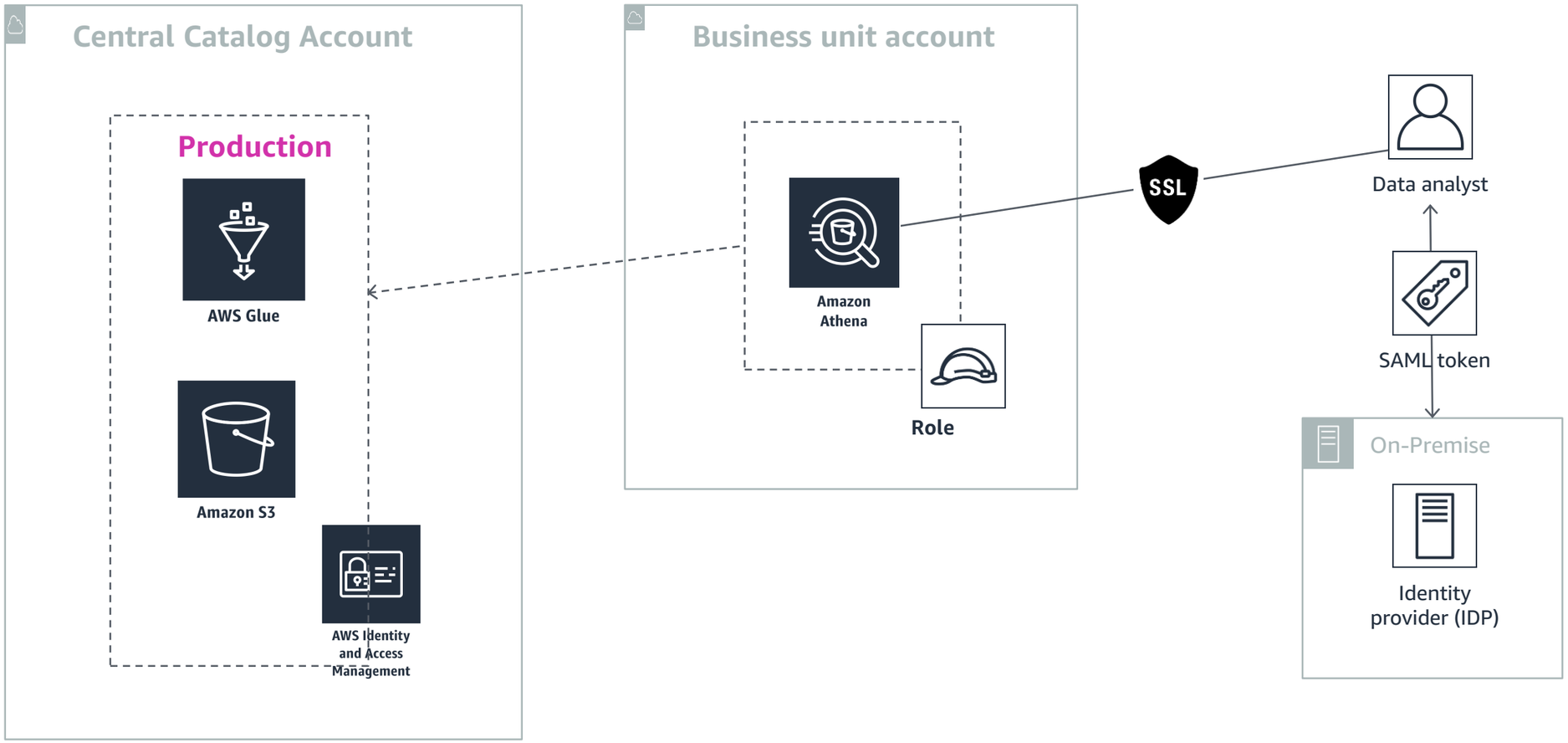
Amazon Athena - create view

Amazon Athena

CREATE VIEW sales_electronics AS
SELECT sum(price) FROM sales, product
WHERE sales.sku = product.sku and
product.category = ‘Electronics’



Amazon Athena – secure data flow

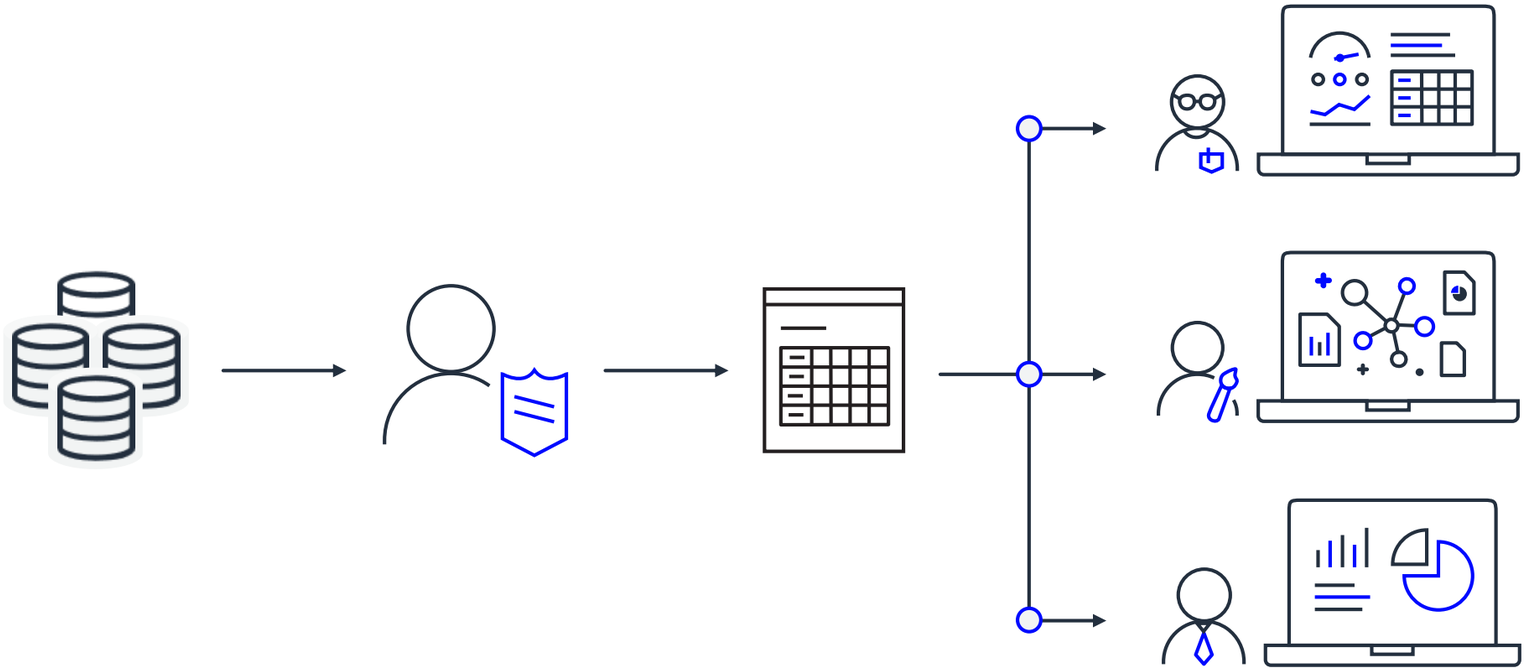


Amazon QuickSight - data governance

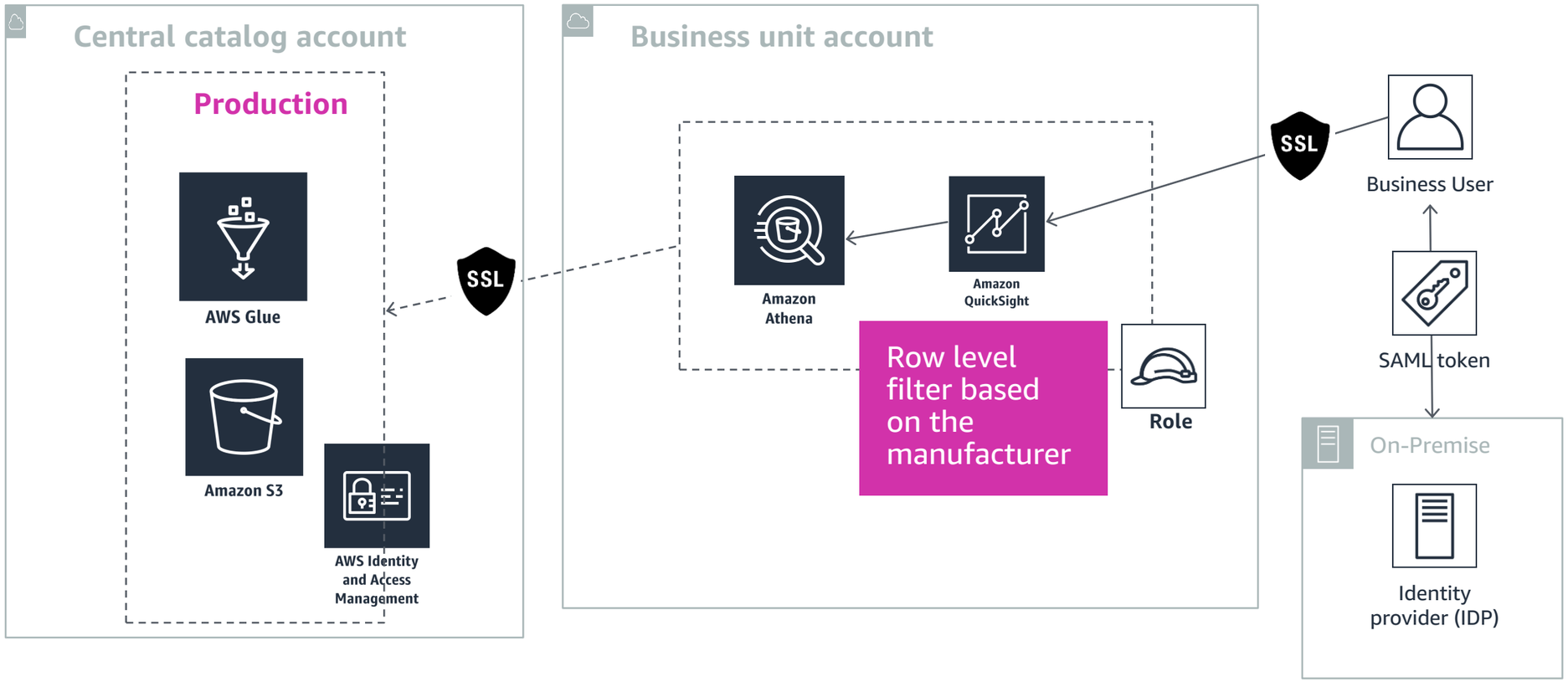
Create managed datasets that give power users and authors the flexibility to perform self-serve analytics on data that you control.

Create datasets that:

- Can be shared with any user
- Automatically refresh
- Have row level security
- Users cannot modify
- Dynamically update with changes



Amazon QuickSight – secure data flow



Amazon Athena - security Controls

Amazon Athena

Authentication

- IAM federation
- Cross-account
- EC2 instance profile

Authorization

- IAM policies mapped to Roles – these polices are passed all the way to storage layer
- Views with Glue Catalog resource policies

Audit

- All API calls are logged to CloudTrail
- S3 Access Logs can provide data access information

Data protection at rest

- CSE-KMS
- SSE-KMS
- SSE-S3
- Use separate KMS keys for source and destination buckets

Data protection in motion

- JDBC connections use TLS/SSL by default
- Data transfer between S3 and Athena is encrypted by TLS

Compliance Programs

- SOC 1,2,3
- HIPAA BAA

Amazon QuickSight - security controls

Amazon QuickSight

Authentication

- IAM federation
- QuickSight-only users
- Cross-account via Amazon S3
- MFA
- Differences between Standard and Enterprise

Authorization

- IAM policies
- Row-level Security

Audit

- Amazon CloudTrail
- Amazon S3 Access Logs

Data protection at rest

- Encrypt your source datasets and Amazon S3
- QuickSight Enterprise edition: data at rest in SPICE is also encrypted

Data protection in motion

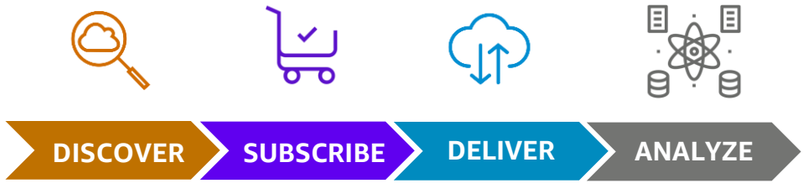
- SSL/TLS
- Interface Endpoints to VPCs and Direct Connect

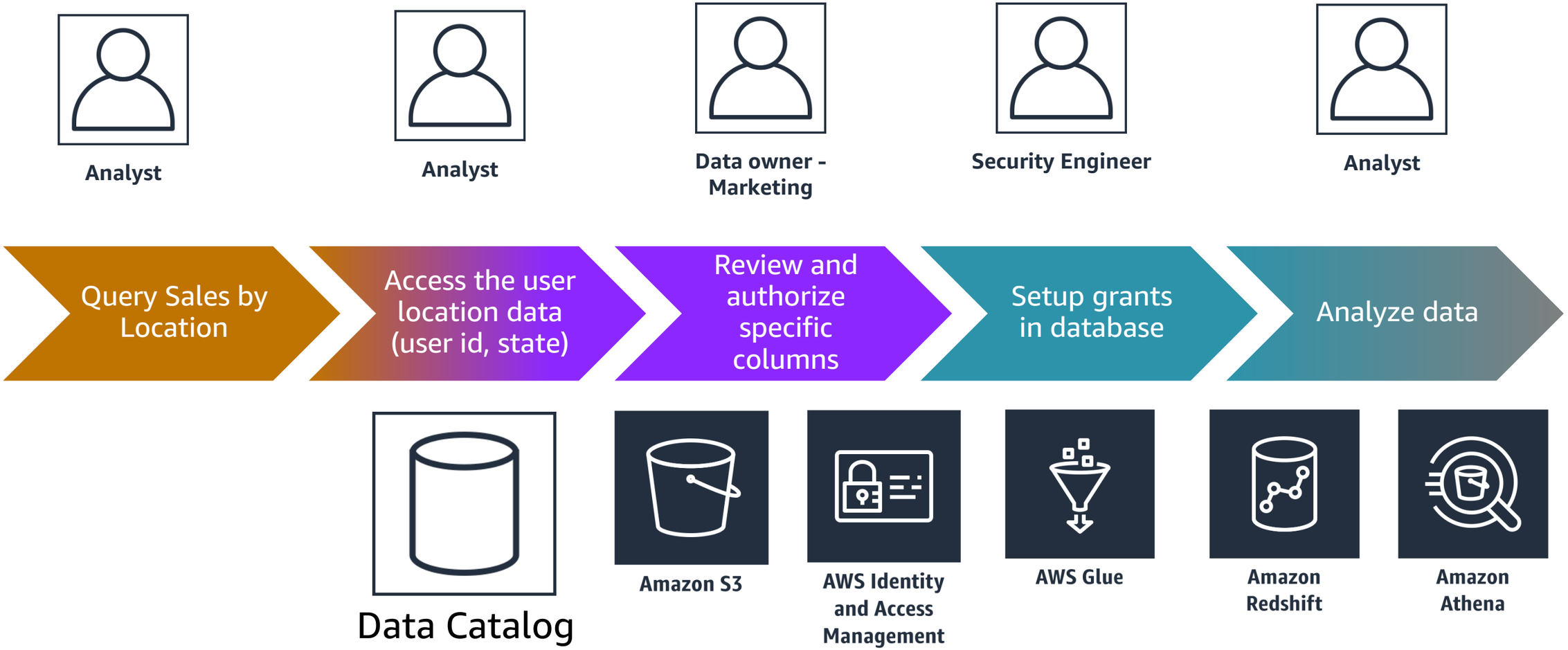
Compliance Programs

- HIPAA
- SOC2
- PCI-DSS
- ISO
- FedRAMP

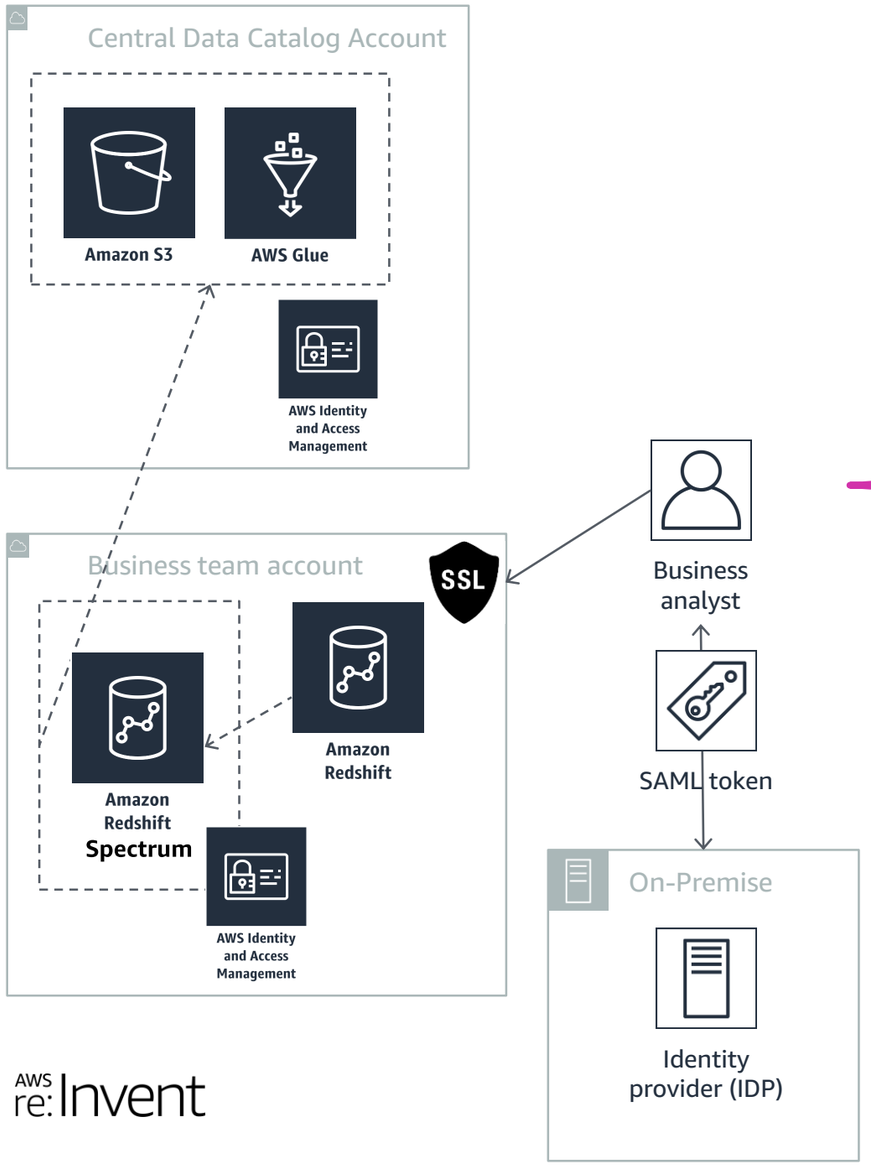
Access existing registered data

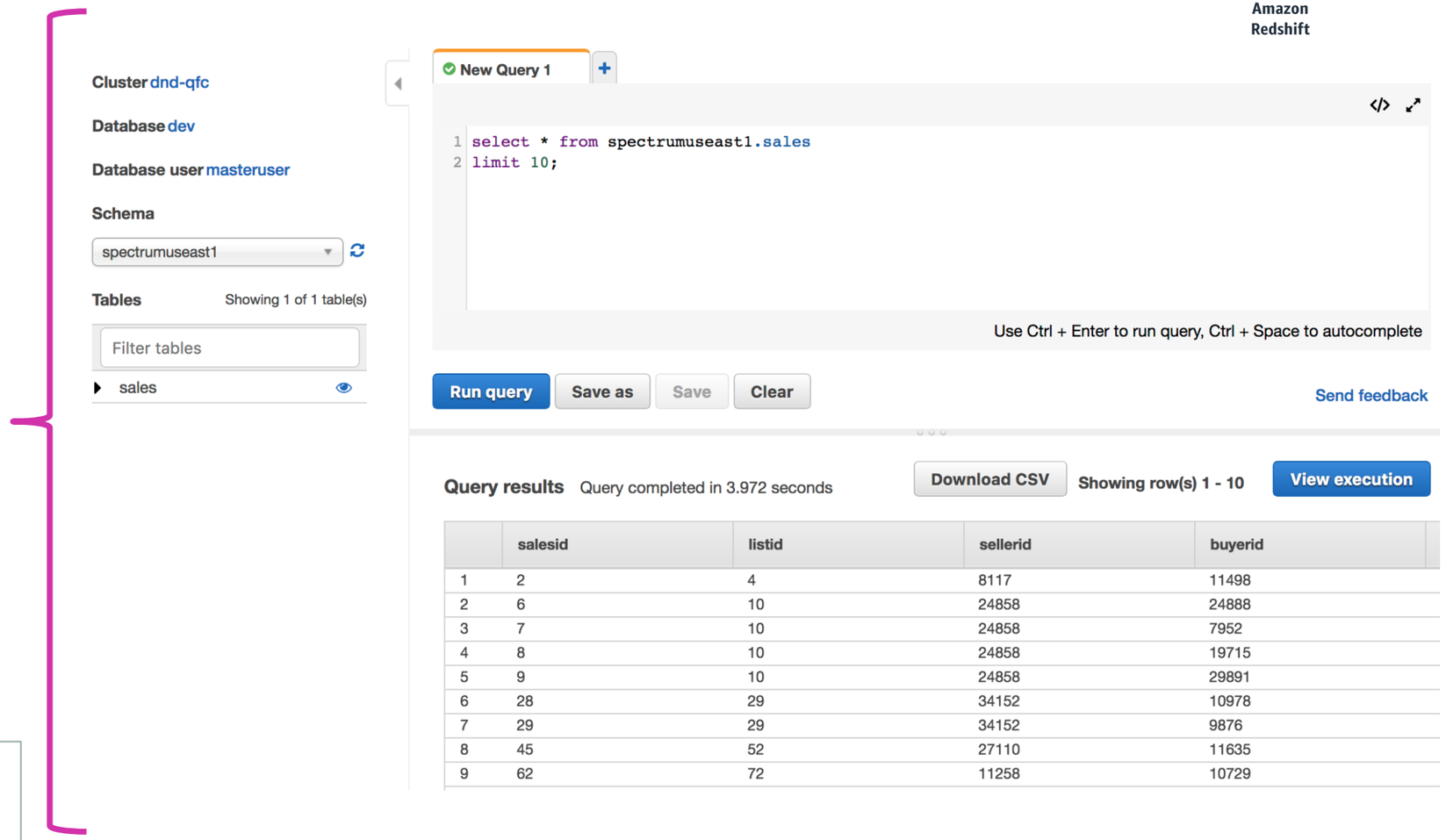
Workflow – access existing data





Amazon Redshift – secure data flow



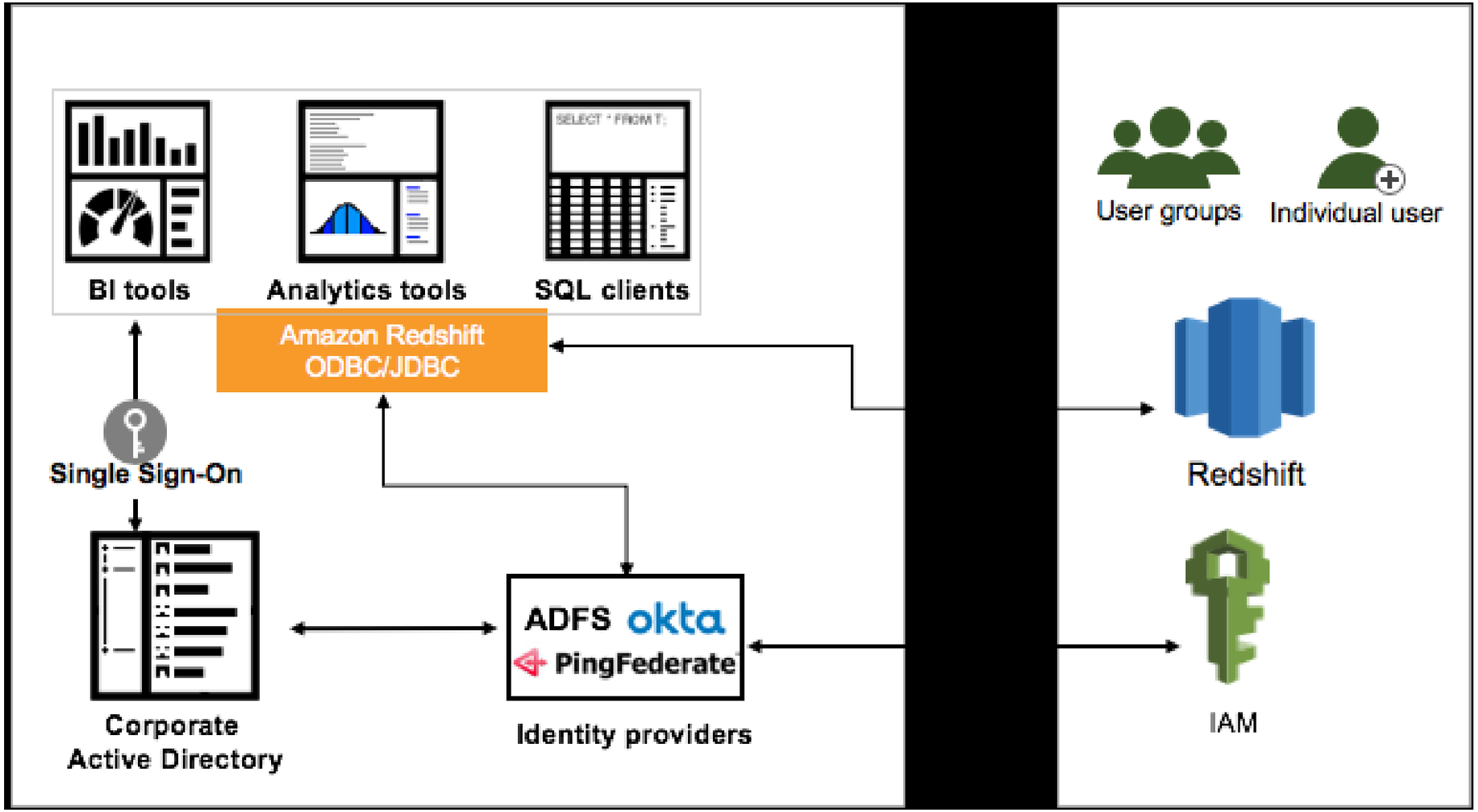


	salesid	listid	sellerid	buyerid
1	2	4	8117	11498
2	6	10	24858	24888
3	7	10	24858	7952
4	8	10	24858	19715
5	9	10	24858	29891
6	28	29	34152	10978
7	29	29	34152	9876
8	45	52	27110	11635
9	62	72	11258	10729

AWS re:Invent

aws

Redshift federated authentication



Amazon Redshift federated authorization

Amazon Redshift

- Setup Amazon Redshift DBGroups
  - For example - CREATE Group ‘XXX’

- Use Grants to setup authorization access
  - GRANT SELECT on table ‘YYYY’ to group ‘group1’

- Configure SAML assertion for your IDP

```
<Attribute Name="https://redshift.amazon.com/SAML/Attributes/DbGroups">
  <AttributeValue>group1</AttributeValue>
  <AttributeValue>group2</AttributeValue>
  <AttributeValue>group3</AttributeValue>
</Attribute>
```

Amazon Redshift audit logging

Cloudtrail

- Amazon Redshift API calls
- KMS API calls
- S3 calls

Amazon S3

- Connection logs
- User logs
- User activity logs

Best practices - Amazon Redshift security

Amazon Redshift

Authentication

- IAM federation
- DB username and password

Authorization

- DB groups with grants
- Restrict access by IAM policy
- Use condition keys “ResourceTag” and “RequestTag”

Audit

- API logs to Amazon Cloudtrail
- Logs to Amazon S3
    - Connection logs
    - User logs
    - User activity logs

Data protection at rest

- KMS
- HSM – AWS CloudHSM Classic
- Key rotation – CMK, DEK

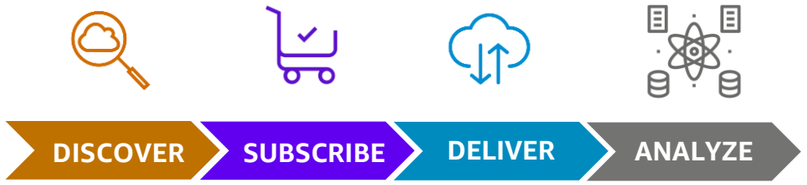
Data protection at motion

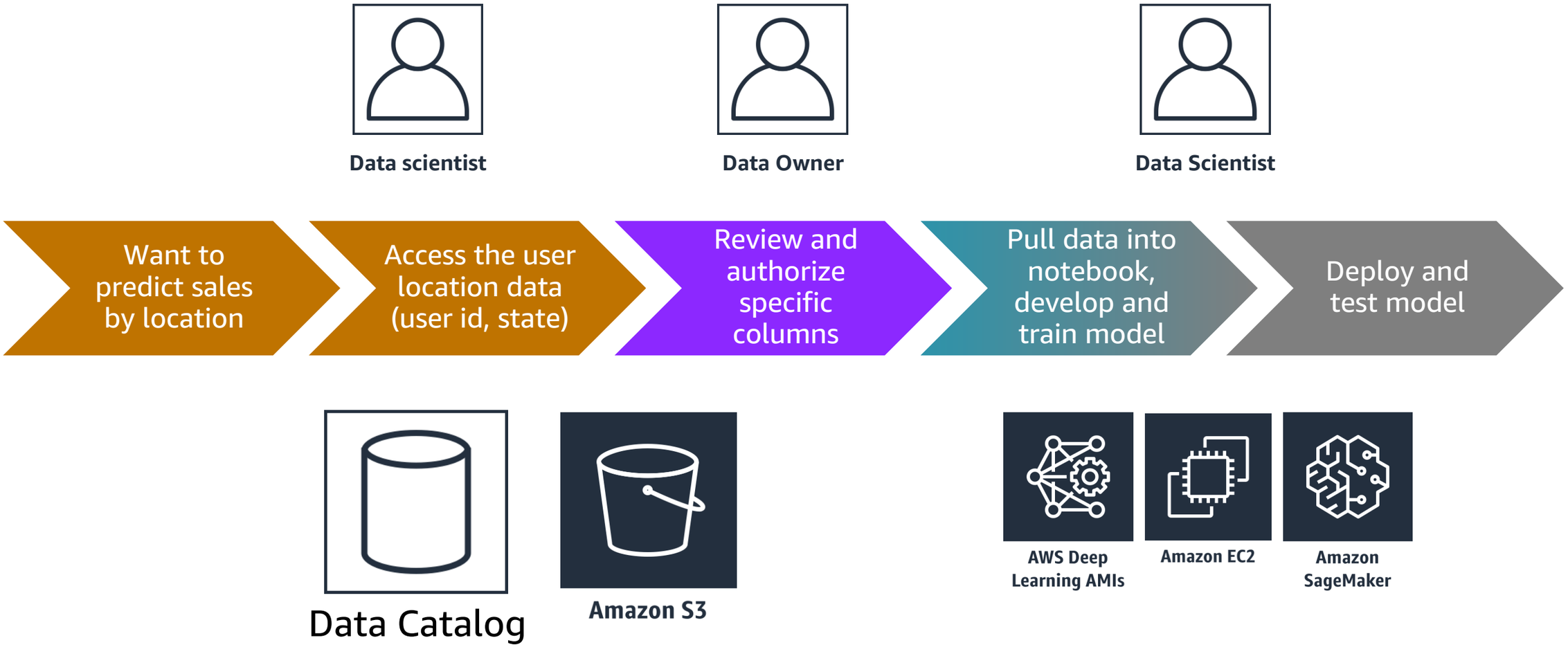
- SSL (ACM) - Set “require_SSL = true” in parameter group
- FIPS 140-2 support

Compliance Programs

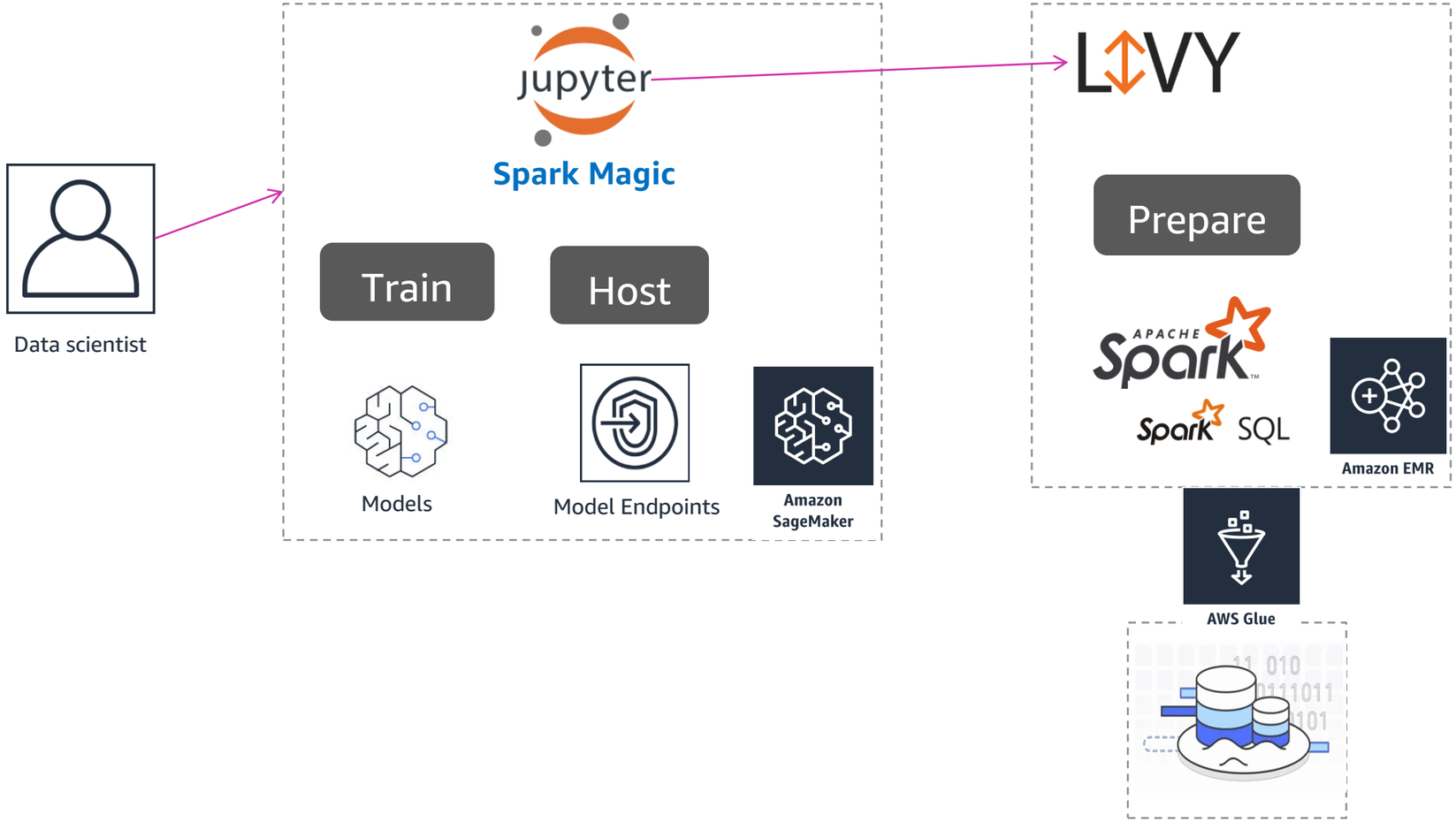
- SOC1,2,3
- PCI DSS Level 1
- FedRAMP
- HIPAA eligible with BAA

Workflow – build predictive model





Amazon SageMaker with Apache Spark



Best practices – Amazon Sagemaker security

Amazon SageMaker

Authentication

- IAM federation

Authorization

- Restrict access by IAM policy and condition keys

Audit

- API logs to Amazon Cloudtrail - exception of InvokeEndpoint

Data protection at rest

- KMS based encryption for
  - Notebooks
  - Training jobs
  - Amazon S3 location to store models
  - Endpoint

Data protection at motion

- HTTPS for API/Console
- Notebooks
  - VPC enabled
  - Interface endpoint
  - Limit by IP
- Training jobs/Endpoints
  - VPC enabled

Compliance Programs

- PCI DSS
- HIPAA eligible with BAA
- ISO

Amazon.com's vision is to be the earth's most customer—centric company; where people can find anything they want to buy online.

Challenge:

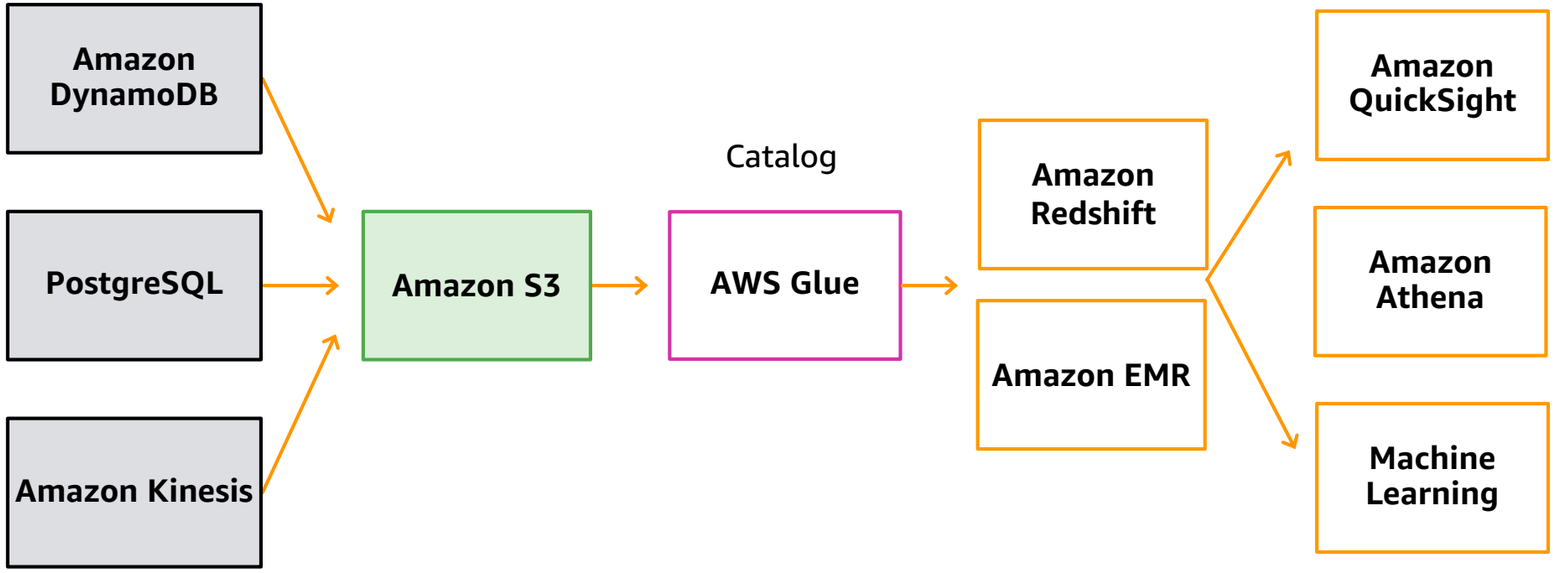
Load 500K+ transactions each day, and serve 300K+ queries/extracts each day from Amazon businsses (Amazon.com, Amazon Prime, Amazon Music, Amazon Alexa, Amazon Video, and Twitch).

Solution:

- Land data in S3 as a data lake
- Use Redshift as preferred SQL based analysis by business users, and EMR for machine learning

Amazon.com uses AWS for data lakes & analytics

amazon.com



- DynamoDB capturing all Amazon.com transactions
- Everything from DynamoDB, RDS PostgreSQL and Kinesis fed to a Amazon S3 data lake
- AWS Glue used to catalog the data
- Amazon Redshift used for all SQL-based queries, and Amazon EMR for all machine learning and big data processing
- End-users use Amazon QuickSight for visualizations

Summary

- Federate access
- Setup roles and responsibility matrix within your organization
- Leverage centralized data catalog
- Use both preemptive and detective controls
- Perform regular audits
- Secure storage, catalog and processing layers
- Incentivize teams to register datasets to catalog
- Streamline process between data producers and data consumers

Thank you!

Varun Rao Bhamidimarri
vbhamidi@amazon.com
Tony Nguyen
aanwin@amazon.com

AWS re:Invent

aws

Please complete the session survey in the mobile app.