AWS re:Invent

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

ANT405-R2

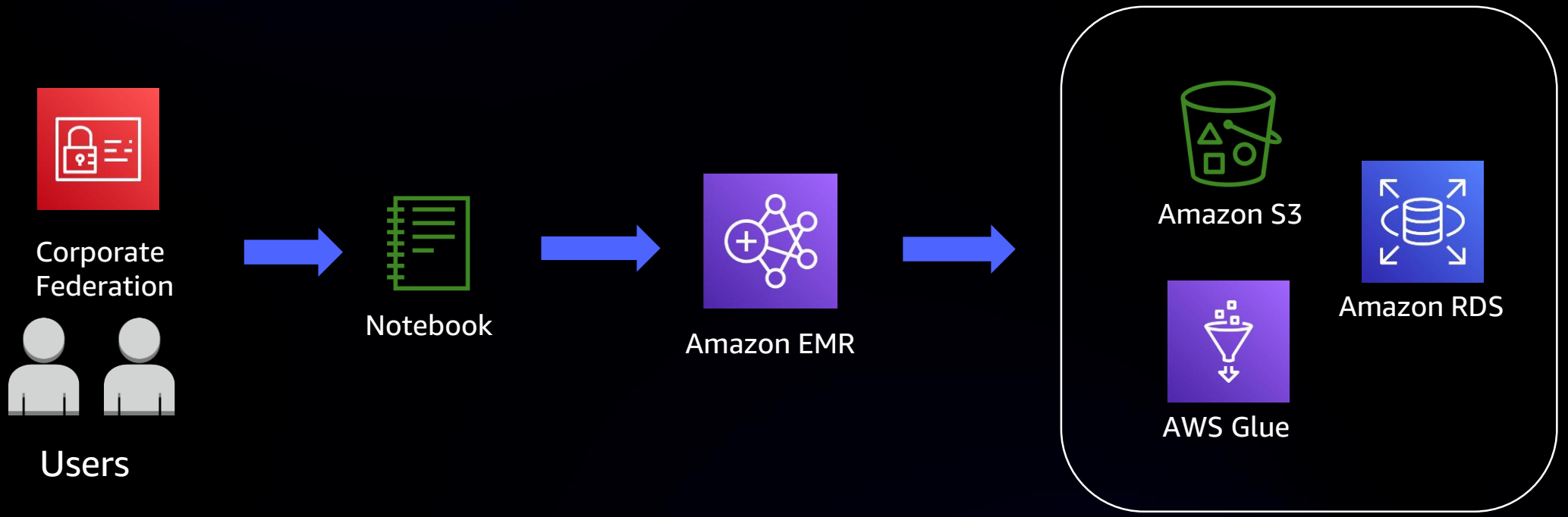
Enforcing data access control on Amazon EMR

Jalpan Randeri
Senior Software Development Engineer
AWS

Matthew Liem
Senior Big Data Architect
AWS

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

EMR Job pipeline



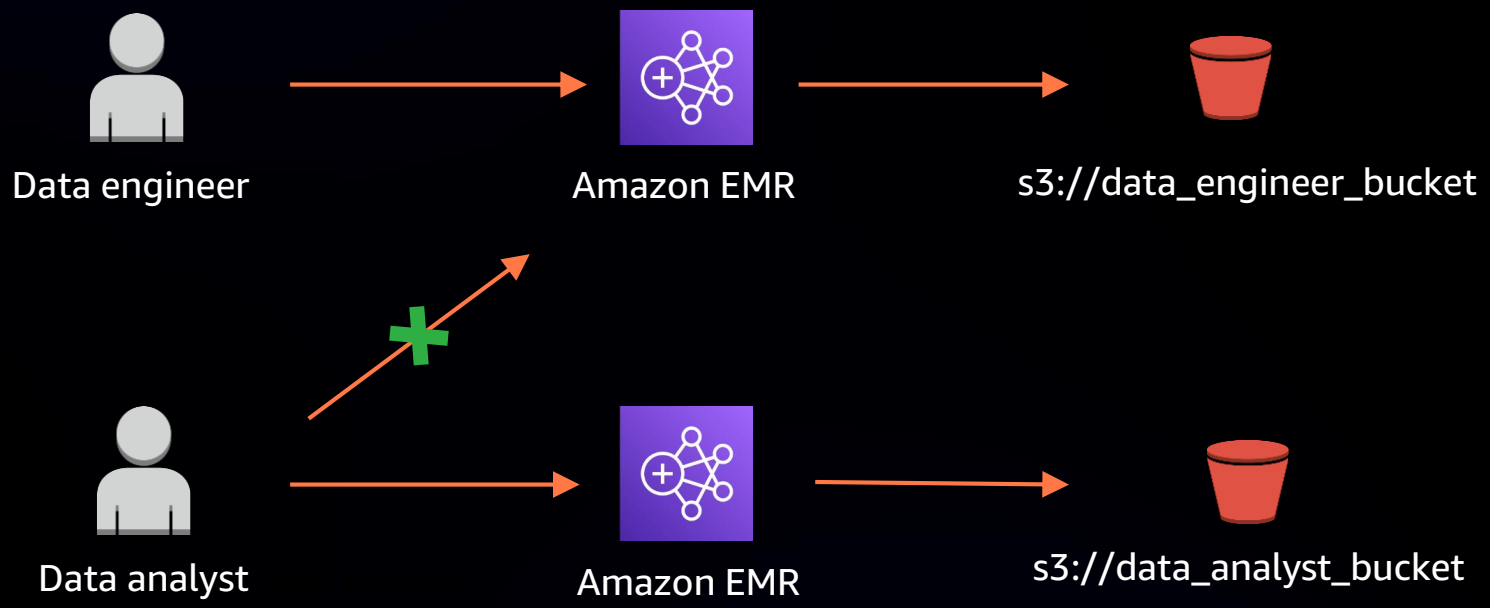
© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

What is access control



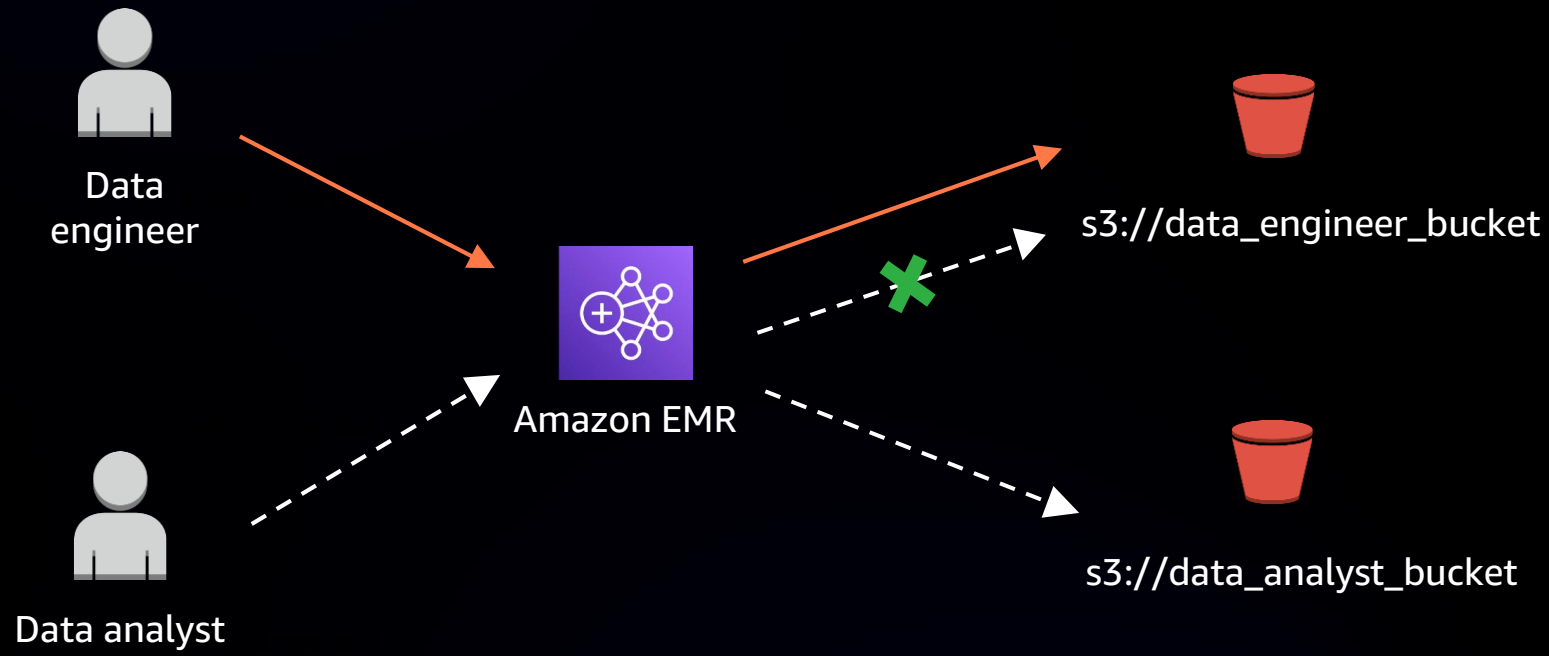
© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Single tenancy on Amazon EMR



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

Multi-tenancy on Amazon EMR

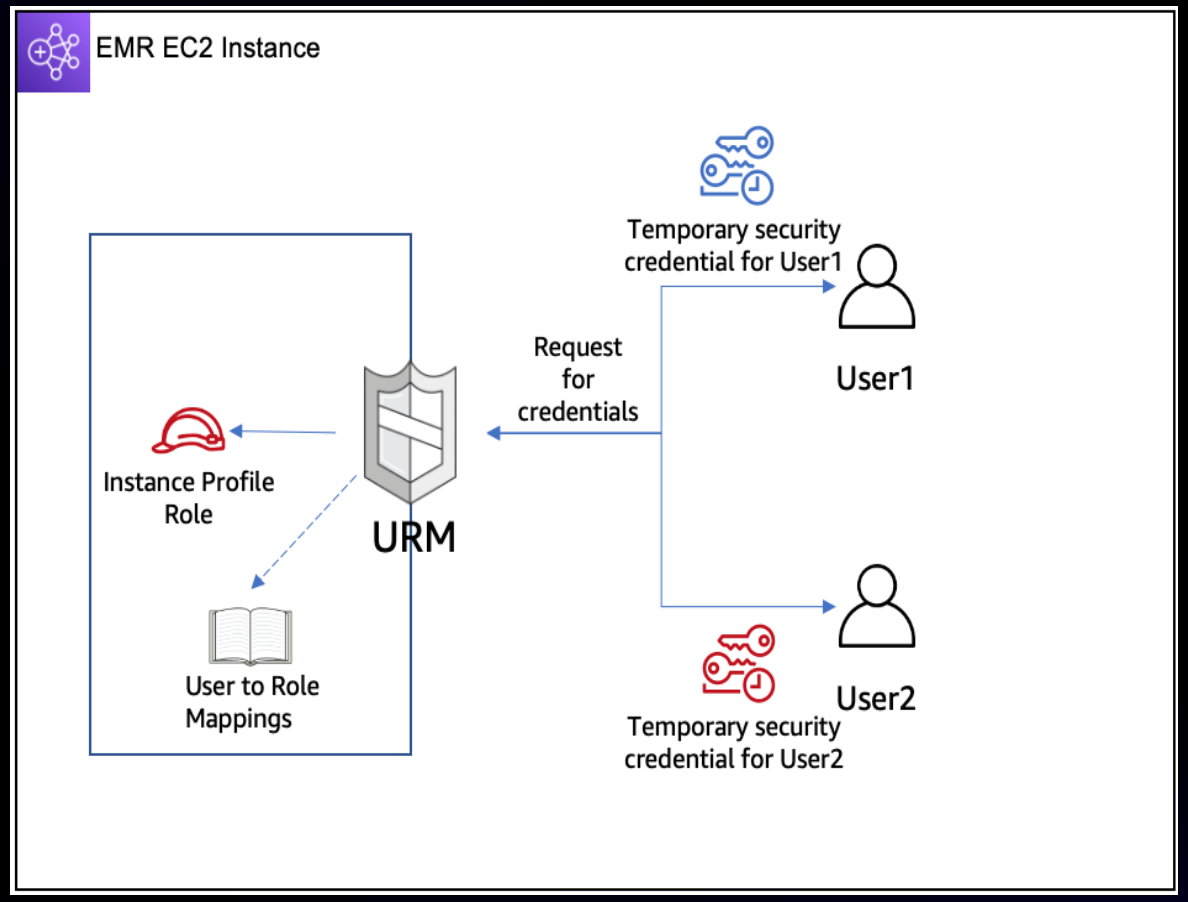


© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Access control via Amazon EMR User Role Mapper

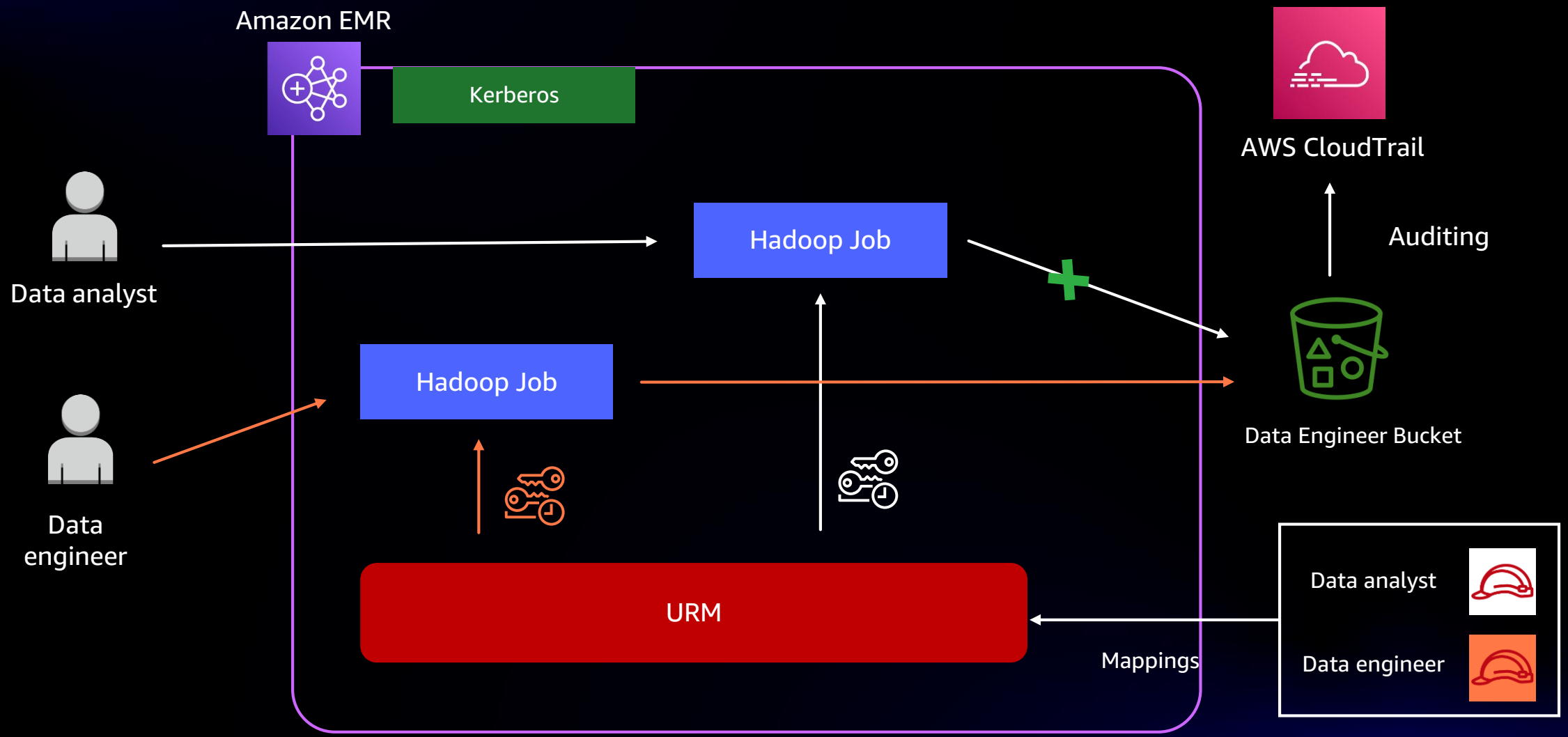
© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon EMR User Role Mapper (URM)



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

Amazon EMR User Role Mapper (URM)



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

Considerations and limitations

- Works with both EMR 5.29+ and EMR 6.x
- Currently only available via AWS Labs (https://github.com/awslabs/amazon-emr-user-role-mapper)
- Requires Kerberos
- Supports Spark, Hive, and PrestoSQL
  - Hive requires enabling user / group mapping and authorization like SQL standards-based authorization
  - Both Hive and PrestoSQL require changes in order to impersonate the end user

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

Fine-grained access control

Databases	Entities		Sales	
Tables	Users	Products	Transactions	Purchases
Data engineer				
Data analyst				

 Full Access

 Read Only Access

 Read Only Access (subset of columns)

 No Access

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

Fine-grained access control with Apache Ranger

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

What is Apache Ranger

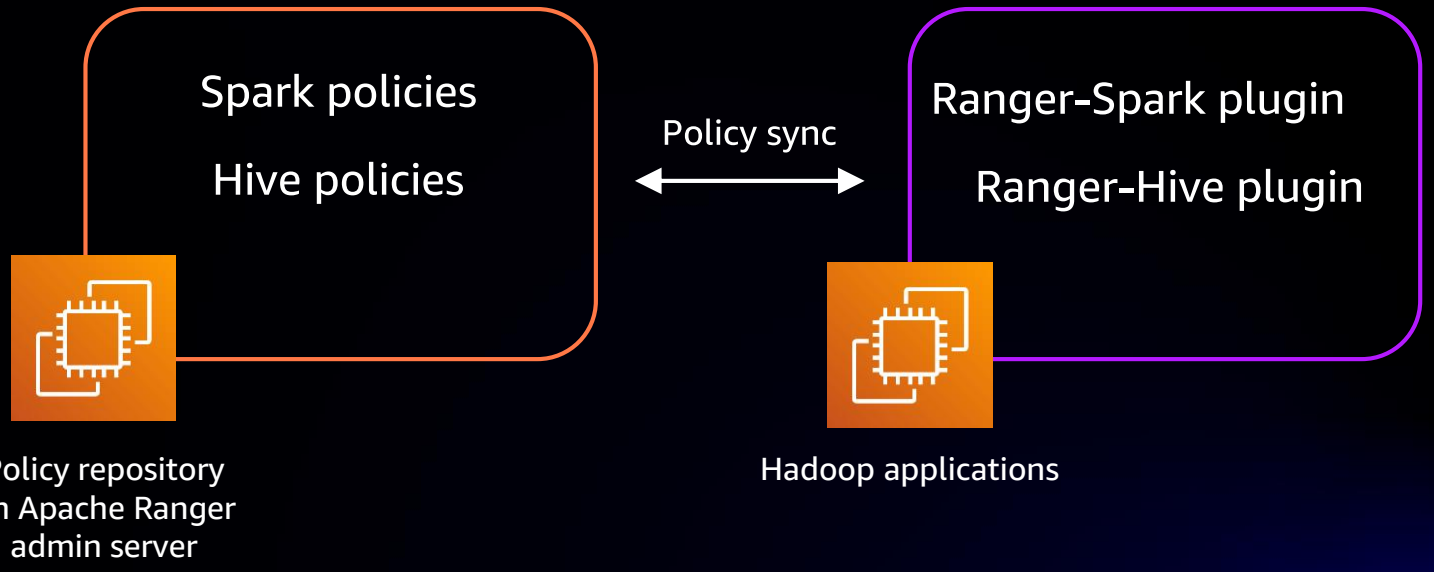
- Apache Ranger is an open-source application to enable, monitor, and manage comprehensive data access control across the Hadoop platform
- Allows you to enforce fine-grained access control (e.g., table and column level access control)

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

How Apache Ranger works

Two key components

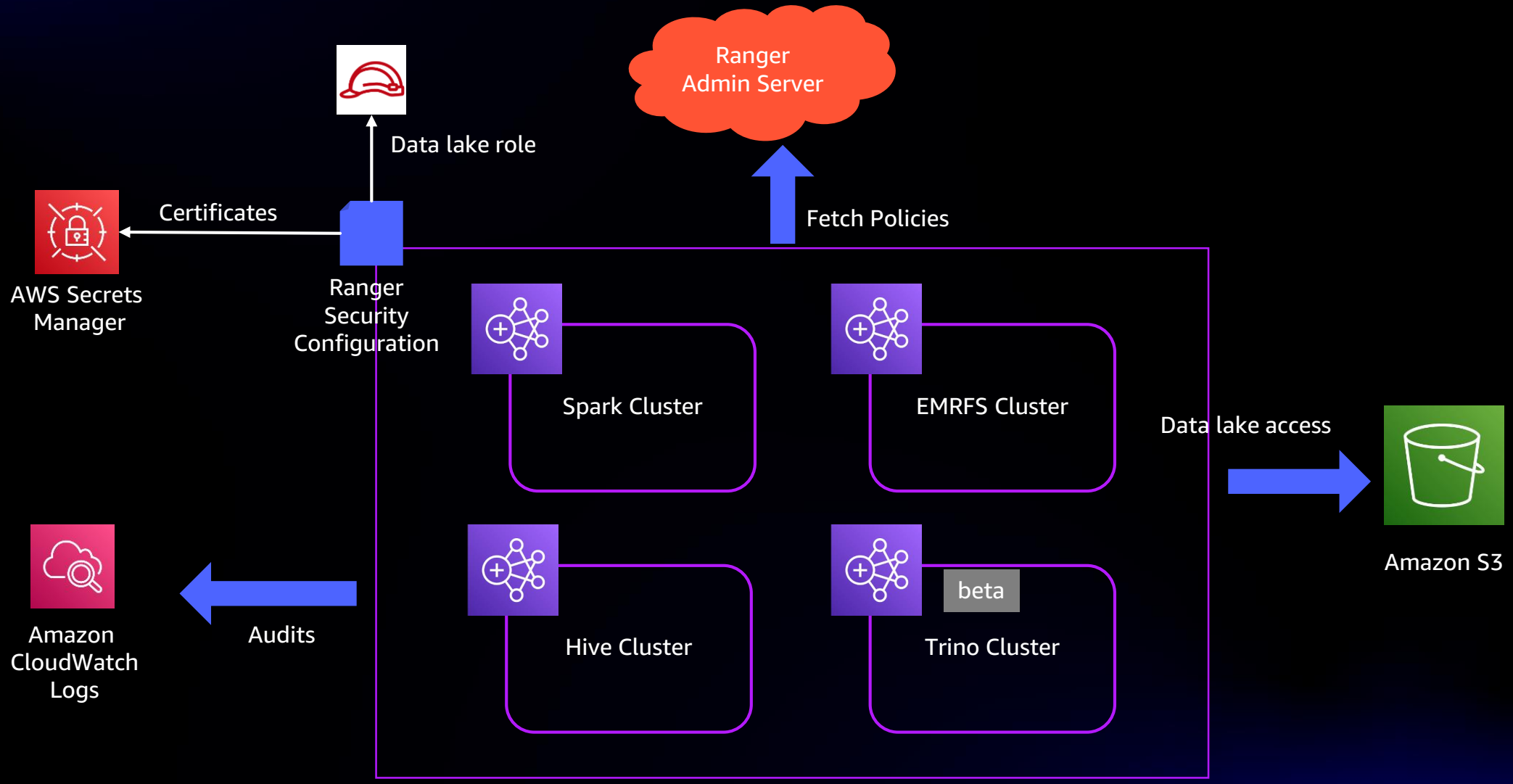
- Apache Ranger admin server – a policy repository for different Hadoop applications
- Apache Ranger plugin for a Hadoop application



aws

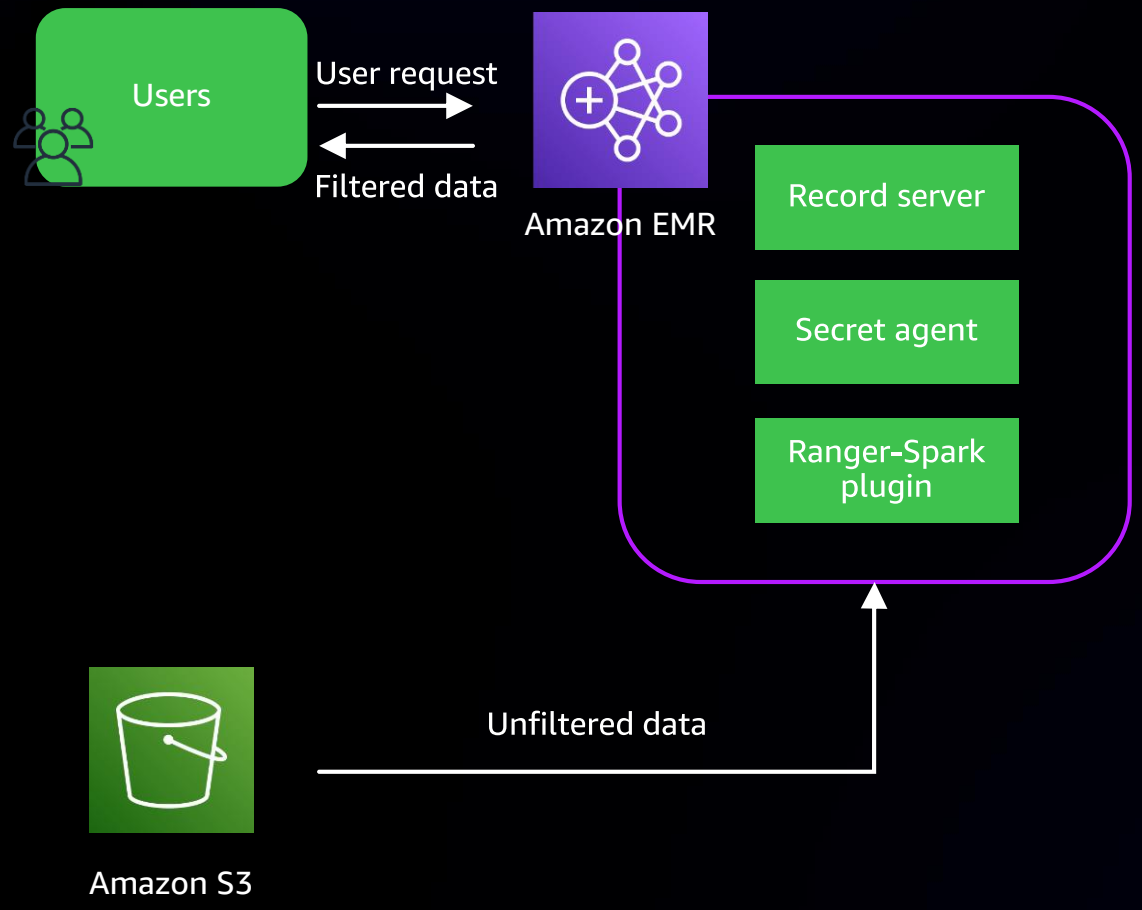
© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Configure integration with Apache Ranger



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

Amazon EMR Ranger-Spark plugin



Amazon EMR Record Server

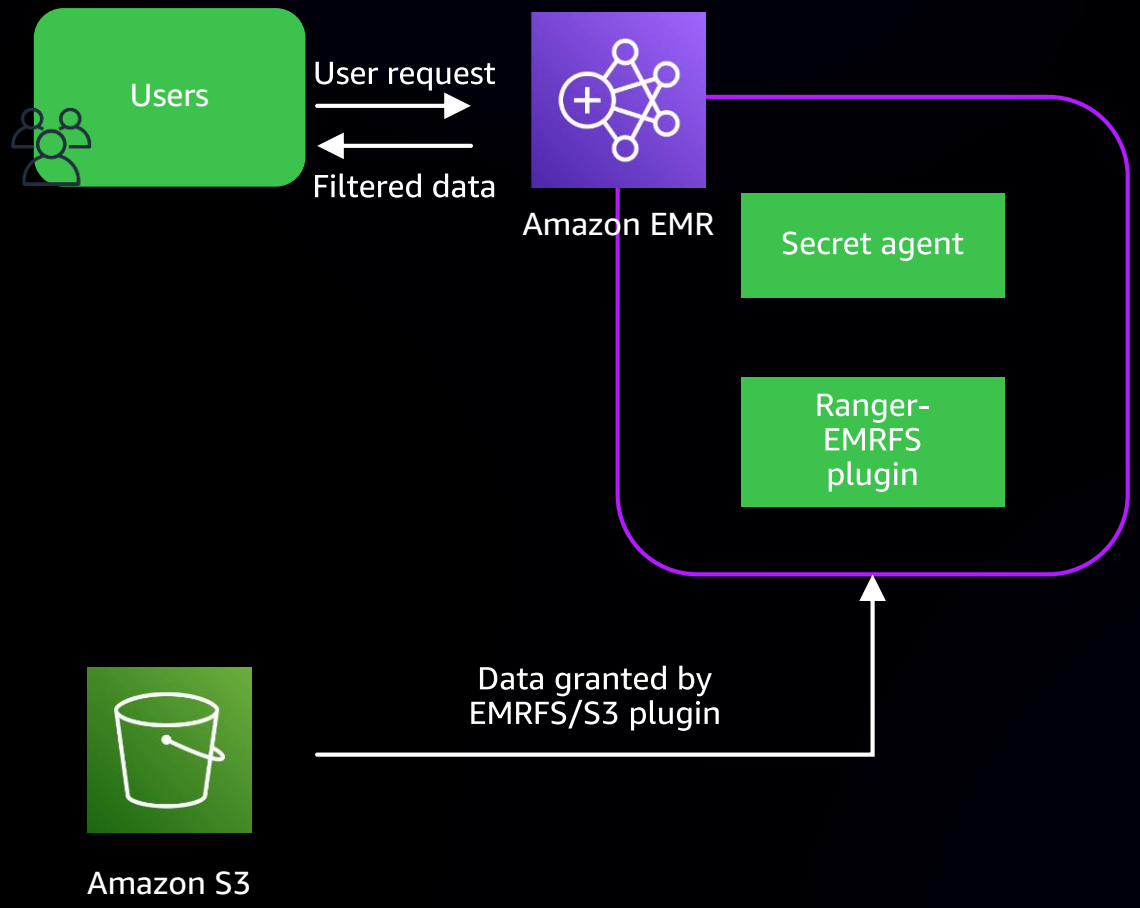
- Forces Spark to delegate action of reading data
- Reads and filters data

Amazon EMR Secret Agent

- Intercepts user requests
- Identifies the right IAM credentials
- Manages temporary credentials

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

Amazon EMR Ranger-EMRFS plugin



Manages access to Amazon S3 via EMR File System (EMRFS)

- Required as a minimum Apache Ranger configuration
- Define prefix-level and object-level authorization policies

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

Considerations and limitations

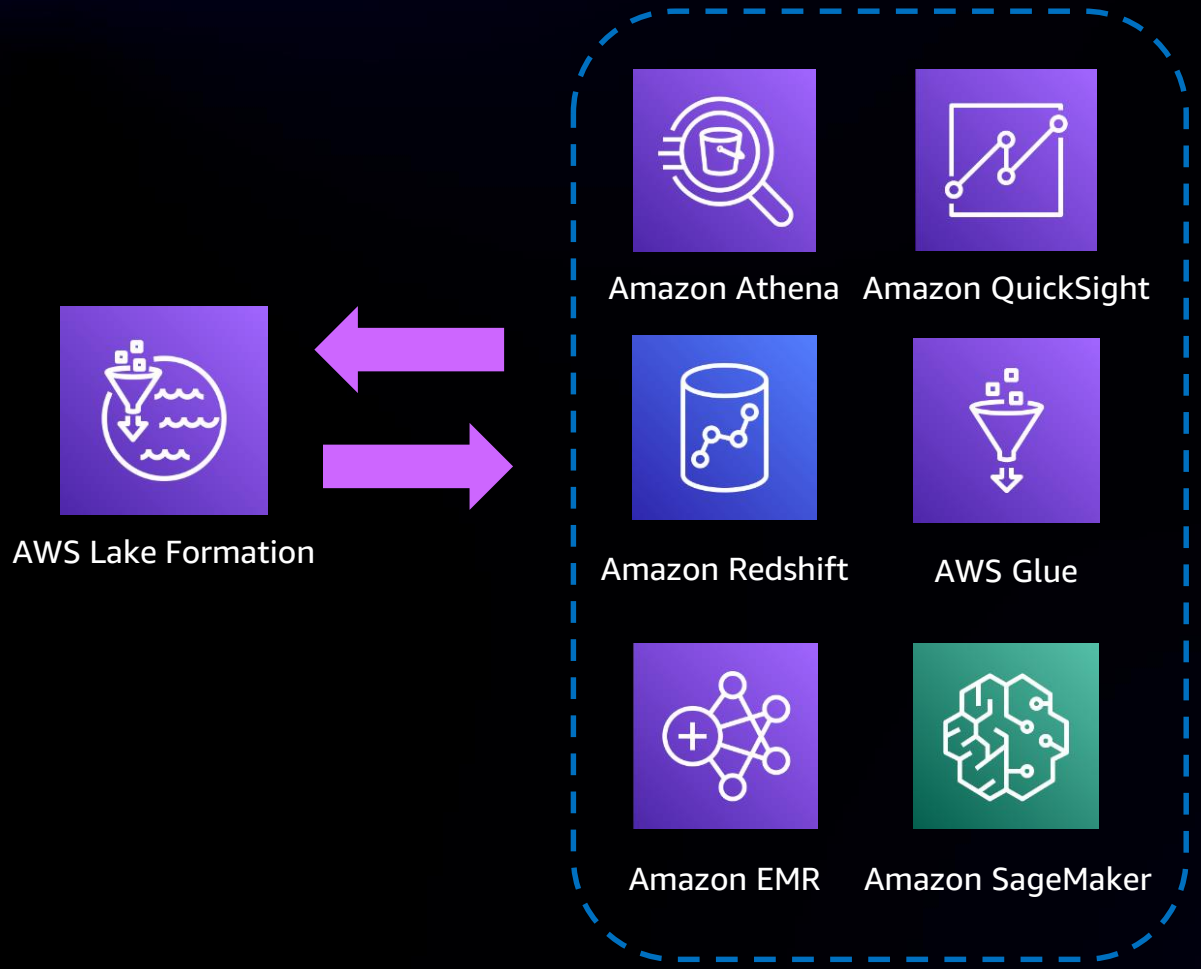
- Use with EMR 6.3+ and 5.32+
- Compatible with Apache Ranger admin server 2.x
- Supports Hive metastore only
- Supports Apache Zeppelin, Hue, and SSH
- Only supports INSERT INTO/OVERWRITE and ALTER TABLE for Spark SQL DML and DDL statements (starting from 6.4)
- Does not support Spark row-level filtering and data masking

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

Fine-grained access control with AWS Lake Formation

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

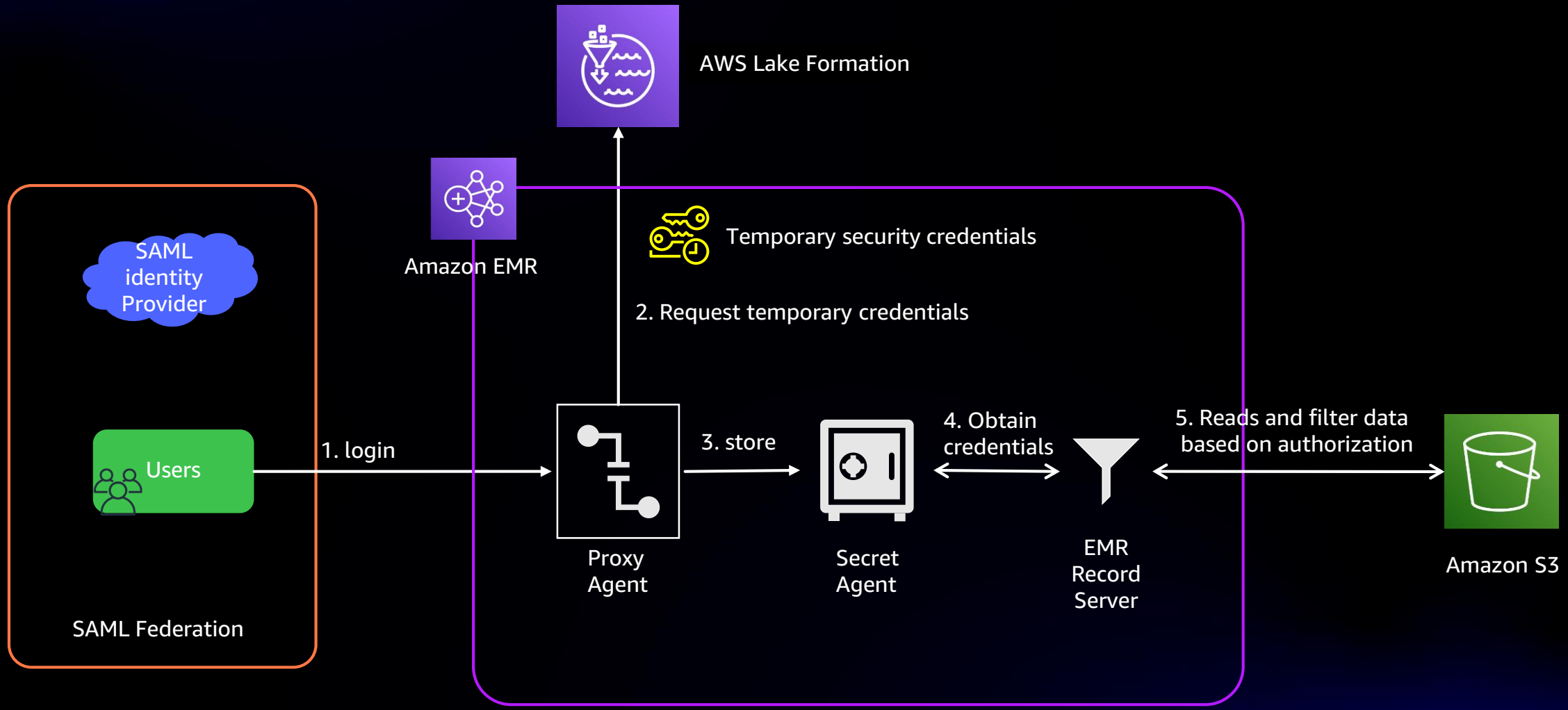
What is AWS Lake Formation?



- Fully managed AWS service that discovers, catalogs, cleanses, and secures data in an Amazon S3 data lake
- Works with various AWS analytics services
- Works with AWS Glue Data Catalog
- Centralized auditing with AWS CloudTrail

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

Integration with AWS Lake Formation



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

Considerations and limitations

- Supports identity providers via SAML only
- Supports AWS Glue Data Catalog only
- Supports Spark only
- Supports Zeppelin and EMR Notebooks via SAML authentication

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

Summary

Feature	User Role Mapper (URM)	Apache Ranger	AWS Lake Formation
Fine-grained access control database / table / column	N/A	Yes	Yes
Metadata store	All	Hive metastore	AWS Glue Data Catalog
Direct control on S3	Yes	Yes, via EMRFS plugin	No
EMR release requirement	5.29+	5.32+	5.31+
Kerberos authentication	Required	Required	Required
Auditing	AWS CloudTrail	Amazon CloudWatch Logs	AWS CloudTrail
Query engine	Apache Spark, Hive, PrestoSQL	Apache Spark and Hive	Apache Spark
Job submission	All	Apache Zeppelin, Hue, SSH	Apache Zeppelin and EMR Notebooks
Multi-primary node	Yes	No	No

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

Thank you!

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