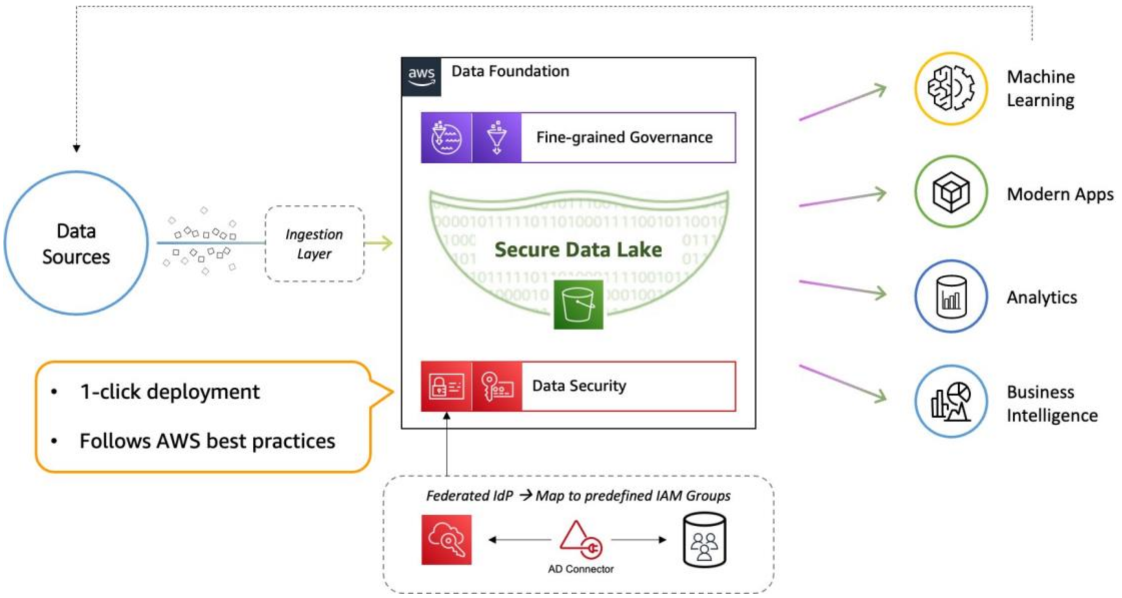
Build Documentation

MDA Architecture Overview



The Data Foundation asset sits at the core of the MDA architecture. It consists of a secure data lake built on S3, augmented with security through KMS and IAM, and fine-grained data governance through Glue and Lake Formation.

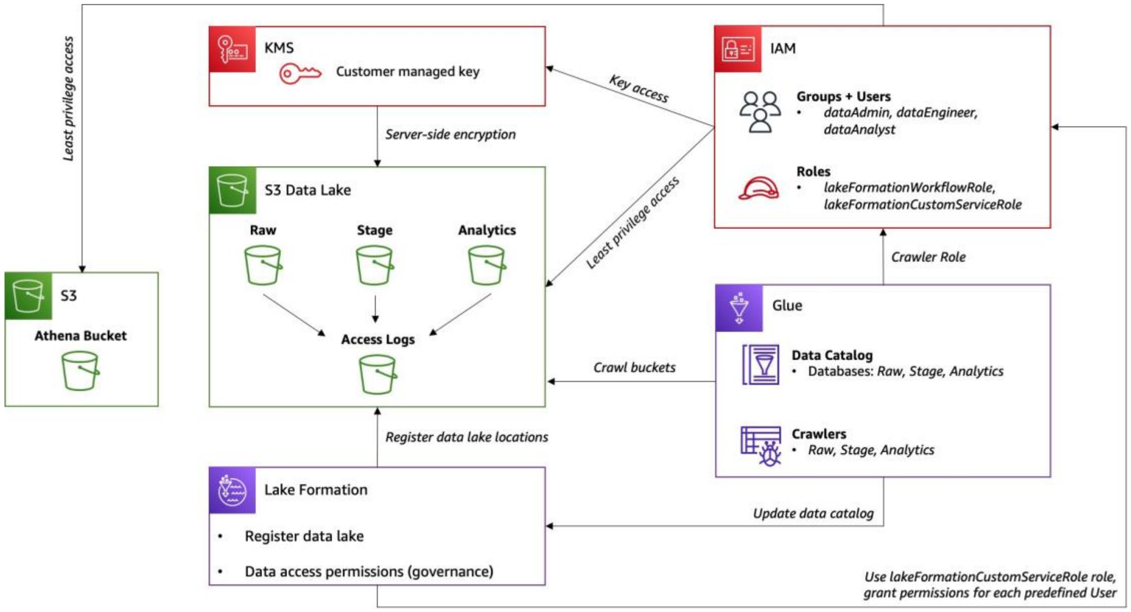
Once deployed, this data foundation allows a functional data system to be built on top of it, with the flexibility to customize the system to the customer's needs.

For instance, fine-grained access controls can be applied to existing user accounts by mapping them to the predefined IAM Groups using federated IdP, which in turn enforces pre-configured fine-grained data governance through Lake Formation. These data access permissions should of course be fine-tuned to meet the customer's specific organizational policies and compliance requirements.

Thereafter, data sources can be ingested into the data lake through an ingestion layer that is built using AWS native services (e.g. DMS, Glue, etc) or ISV solutions. This again can be customized to the customer's requirements.

Finally, analytics and insights can be created to derive business value from the data. For example, visualizing descriptive statistics using dashboards, or creating actionable insights through predictive ML models.

Data Foundation Deployment Overview

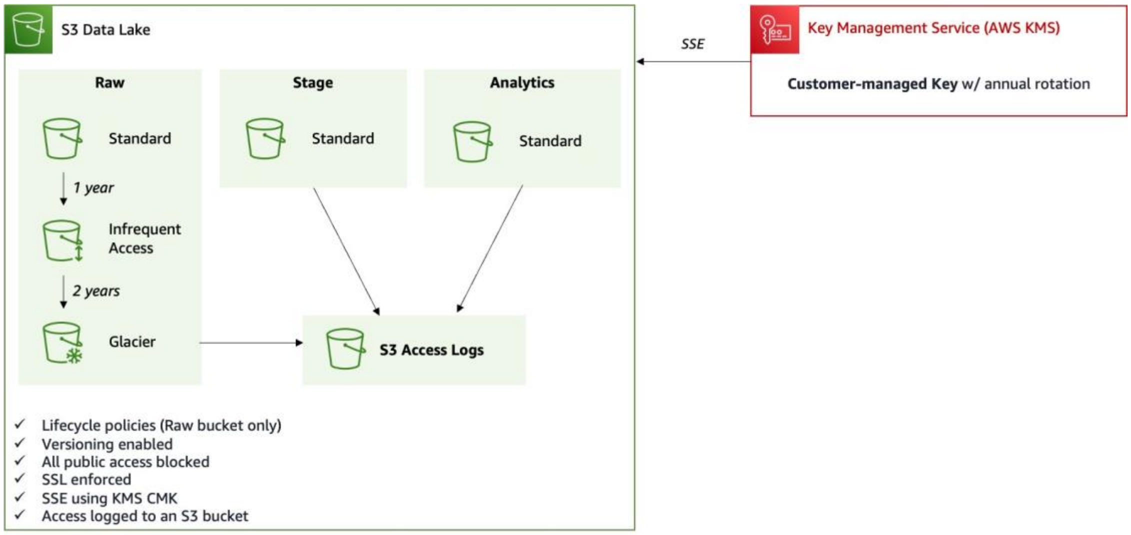


The Data Foundation asset builds the foundational components of a data system shown above, following published AWS best practices.

- A customer-managed key (CMK) is created in **KMS** to encrypt the data lake buckets.
- In **S3**, three data lake buckets are created, with server access logs written to a fourth bucket. Furthermore, a fifth bucket is created for use with Athena.
- **IAM** Groups, Users, and Roles are created with least privilege access to the S3 data lake buckets, CMK, and other essential services such as Lake Formation and Glue.
- In **Glue**, three Databases are created in the Glue Data Catalog, each pointing to their respective data lake buckets in S3. Crawlers are also created to crawl each of the data lake buckets and update the corresponding Database in Glue Data Catalog.
- Lastly, **Lake Formation** is pre-configured to register the S3 data lake locations, create and associate tags to each Database, grant least privilege access permissions to each IAM User and Role, and also assign Lake Formation Admins.

Note: Prior to deploying this asset, the default Lake Formation permissions model must be changed via the AWS Console. The required steps are outlined in the README that is included in the package.

S3 Data Lake

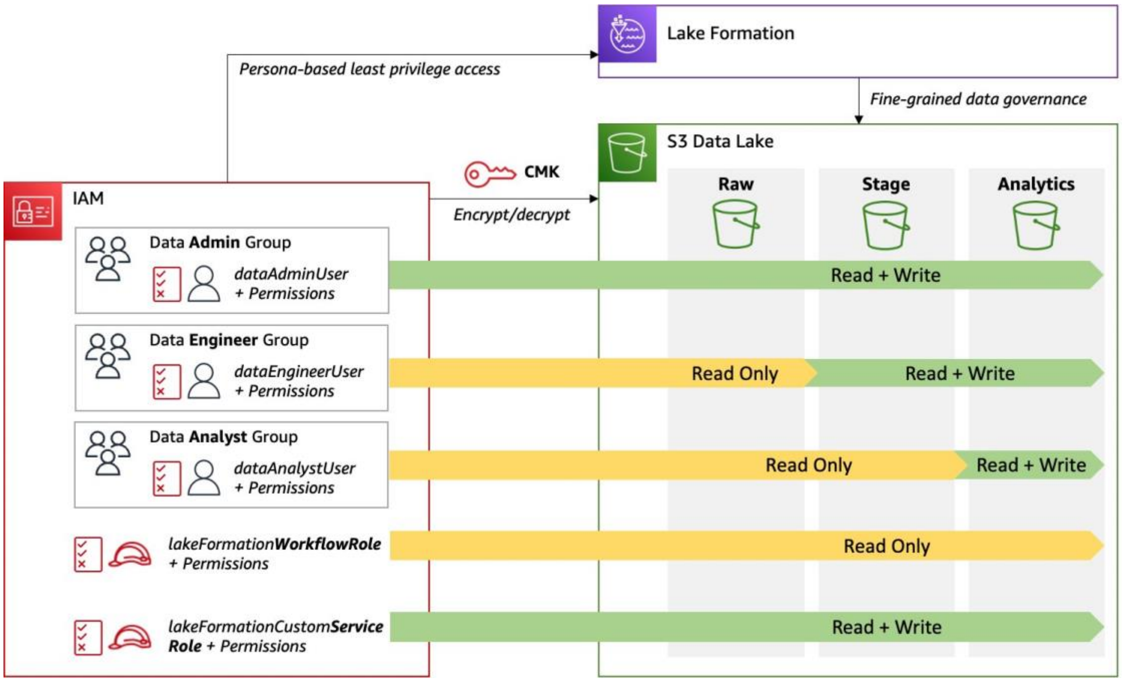

S3 Data Lake
Raw
Standard
1 year
Infrequent Access
2 years
Glacier
Stage
Standard
Analytics
Standard
S3 Access Logs
SSE
Key Management Service (AWS KMS)
Customer-managed Key w/ annual rotation
✓ Lifecycle policies (Raw bucket only)
✓ Versioning enabled
✓ All public access blocked
✓ SSL enforced
✓ SSE using KMS CMK
✓ Access logged to an S3 bucket

The S3 data lake is configured following the specifications outlined in AWS Prescriptive Guidance, such as the bucket naming strategy and lifecycle policies.

Furthermore, security best practices are also configured, including versioning, all public access blocked, SSL enforced, server-side encryption (SSE) using the KMS-managed CMK, and server access logs written to a separate bucket.

Note: Additional buckets and/or lifecycle policies may need to be configured based on the customer’s organizational policies and compliance requirements, for example to handle sensitive data.

IAM



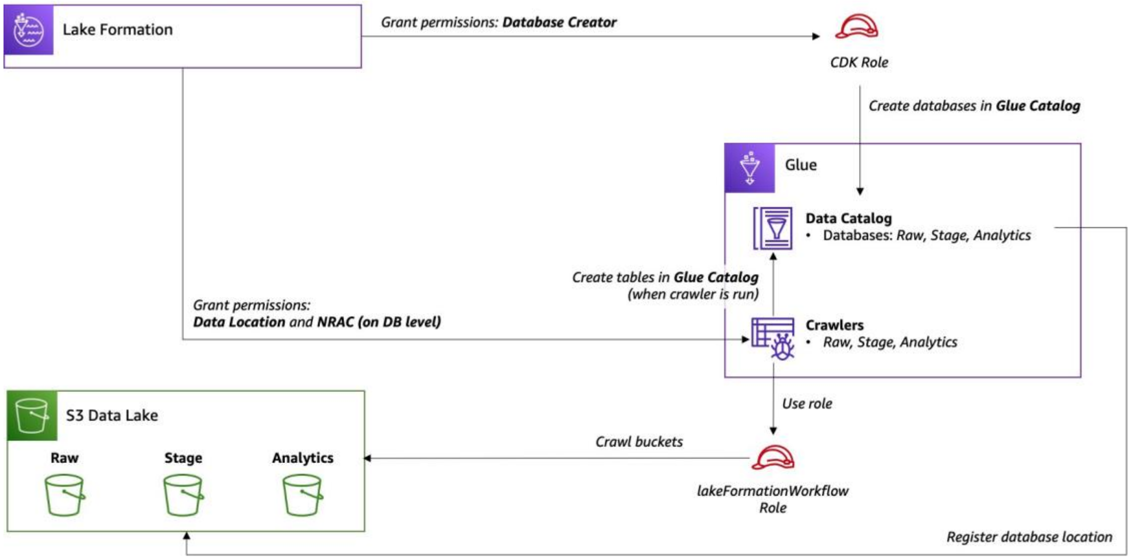
Three IAM Groups are created, each with one User who inherits all permissions of the Group. The Groups are assigned persona-based IAM permissions to enable fine-grained data governance in Lake Formation, and also minimum permissions to the data lake buckets and CMK in order to add/remove files directly in S3.

Since a CMK is used to encrypt the data lake buckets, a custom Lake Formation service role is needed. The role is granted appropriate permissions and also access to the CMK.

Lastly, a Lake Formation workflow role is created and granted S3 read-only access to the data lake buckets and the CMK. This role will be assumed by Glue Crawlers as well as used to execute Lake Formation blueprints and workflows.

Note: The Lake Formation custom service role is granted access to the CMK via an IAM inline policy. This is functionally equivalent to adding the role as a principal to the CMK via KMS.

Glue



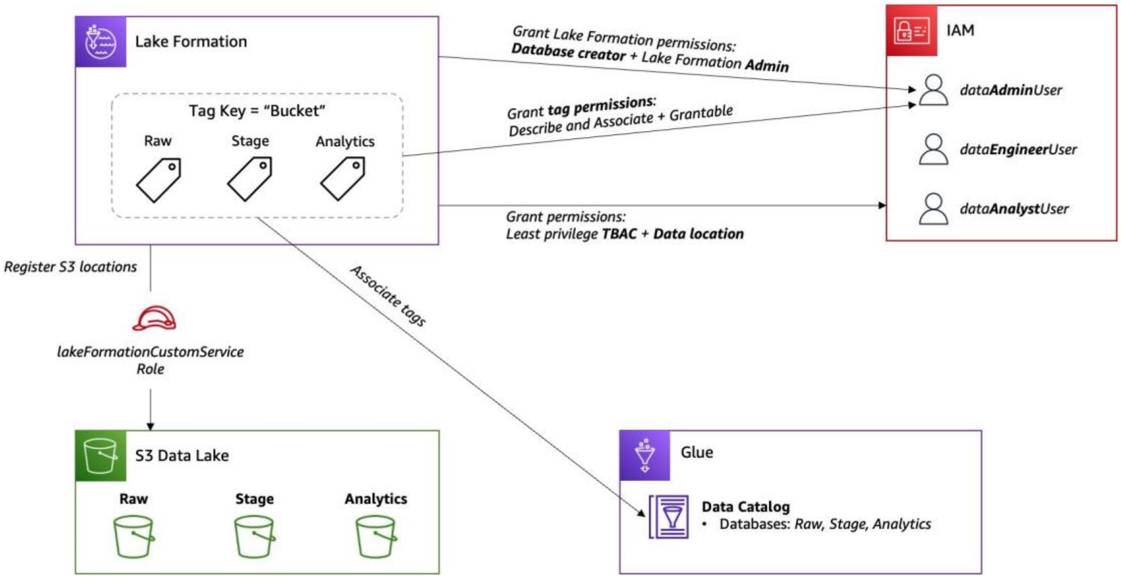
Lake Formation
Grant permissions: Database Creator
CDK Role
Create databases in Glue Catalog
Glue
Data Catalog
• Databases: Raw, Stage, Analytics
Create tables in Glue Catalog (when crawler is run)
Grant permissions: Data Location and NRAC (on DB level)
Crawlers
• Raw, Stage, Analytics
Use role
S3 Data Lake
Raw
Stage
Analytics
Crawl buckets
lakeFormationWorkflow Role
Register database location

Since Lake Formation manages the Data Catalog, the CDK Role must first be granted Database Creator permissions in order to create Databases in the catalog during the build process. With these permissions in place, three Databases are created in Glue Data Catalog, each pointing to their respective data lake bucket in S3.

Next, three Crawlers are created to crawl each of the data lake buckets using the Lake Formation Workflow role created previously. The workflow role is also granted required Lake Formation permissions, namely Data Location and Named Resource Access Control (NRAC) permissions for each of the Databases to allow it to update the Glue Data Catalog.

Note: Instead of granting the Lake Formation Workflow role permissions to S3 and the CMK, Lake Formation credentials can be used instead. This option can be enabled via the AWS Console by editing the “security settings” of a Glue Crawler to “Use Lake Formation credentials for crawling S3 data source”.

Lake Formation



Each S3 data lake bucket is registered as a Data Lake Location in Lake Formation. Furthermore, to register buckets that are encrypted with a CMK, the previously-created custom service role is also supplied.

Next, a tag with the key “bucket” and values [“raw”, “stage”, “analytics”] is created and associated to each of the Databases previously created in the Glue Data Catalog.

Then, each IAM User that was previously created is assigned least privilege Tag-Based Access Control (TBAC) permissions to tag values at the Database and Table levels. Furthermore, users with permissions to create tables are also granted Data Location permissions to allow them to update the Data Catalog.

Permissions on the tag itself are granted to the dataAdminUser to allow it to Describe and Associate this tag, and also Grant these permissions to other users. The dataAdminUser is also added as a Database Creator and Data Lake Admin in Lake Formation.

Note: Since tags are associated at the Database level, all child Tables will also inherit the same tag. However, TBAC permissions must be granted at both the Database AND Table levels. In other words, having TBAC permissions on a Database does not automatically allow a user to have access to its child Tables, even though they share the same tag.