CODING CHALLENGE

TRAINEE NAME: Swathi Baskaran

Unity Catalog:

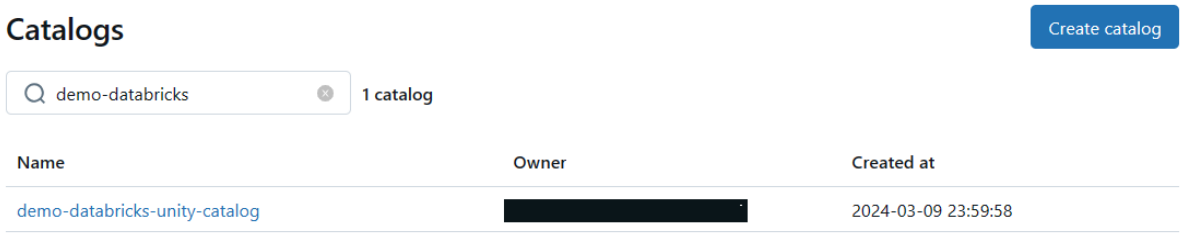
Unity Catalog, offered by Databricks, is a unified governance solution for data and AI assets across the Databricks Lakehouse Platform. It provides a centralized approach to managing data, models, and files, while simplifying access control and auditability.

Key Data Governance Capabilities of Unity Catalog:

1. **Data Discovery:**

   Unity Catalog provides a centralized metadata repository where users can search, explore, and understand datasets across the organization. It helps in:

   - Easily locating datasets through a unified catalog.
   - Understanding schema, data types, and usage.
   - Improving collaboration by providing visibility into available data assets.



2. **Data Audit:**

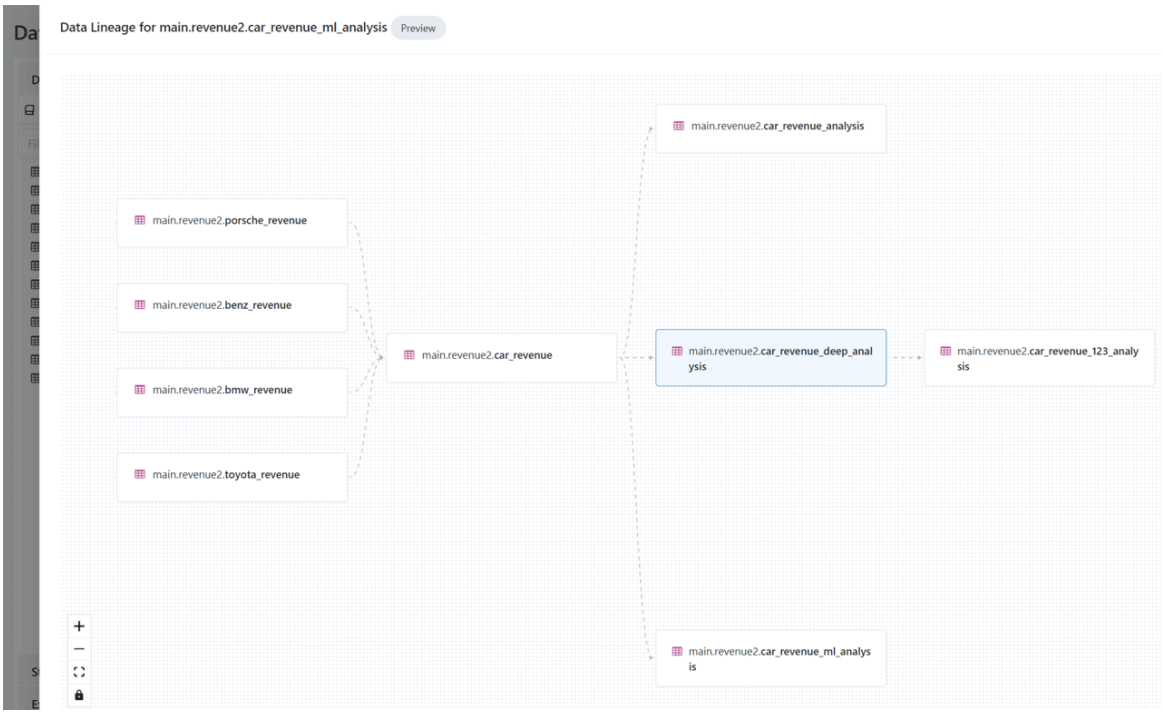
   Auditing is critical for compliance and monitoring. Unity Catalog offers:

   - Detailed logging of data access activities.
   - Audit trails showing who accessed what data and when.
   - Insights for compliance with regulations like GDPR, HIPAA, and CCPA.

3. Data Lineage:

Unity Catalog automatically tracks the flow of data, ensuring transparency in how data is transformed and used. Benefits include:

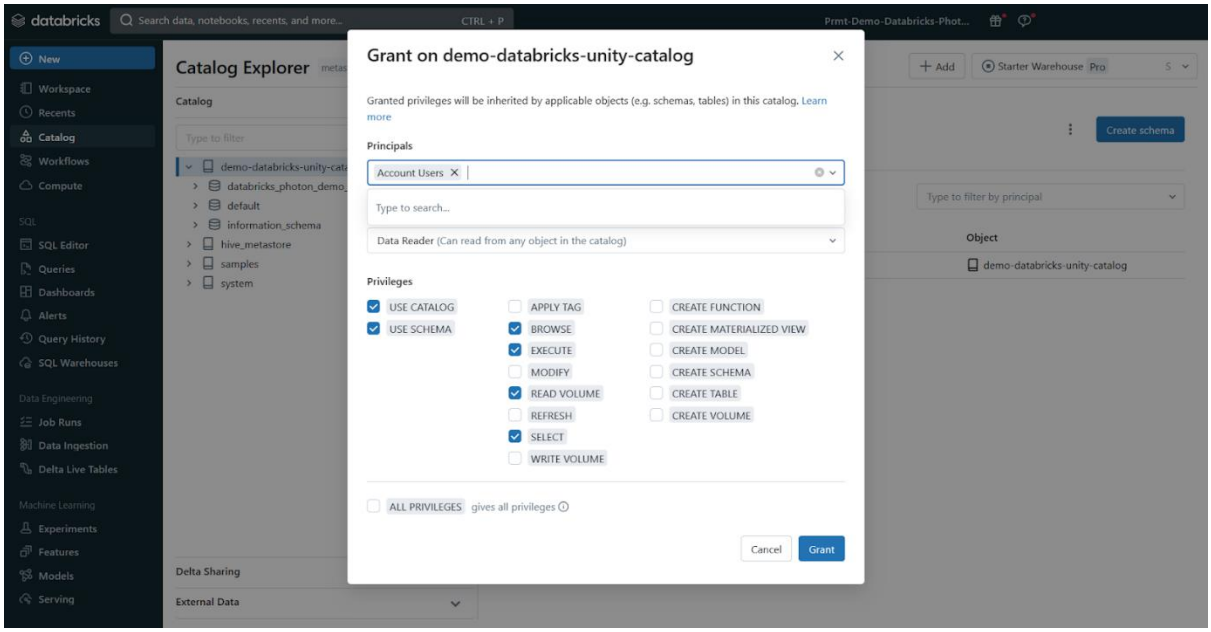
- End-to-end visibility of data pipelines and dependencies.
- Easier debugging of data issues by tracing transformations.
- Improved trust and reliability of analytical results.



4. Data Access Control

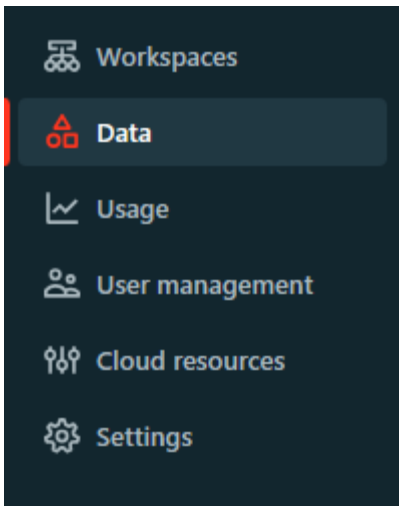
Unity Catalog introduces fine-grained, role-based access controls (RBAC). This ensures that the right users have the right access. Key features include:

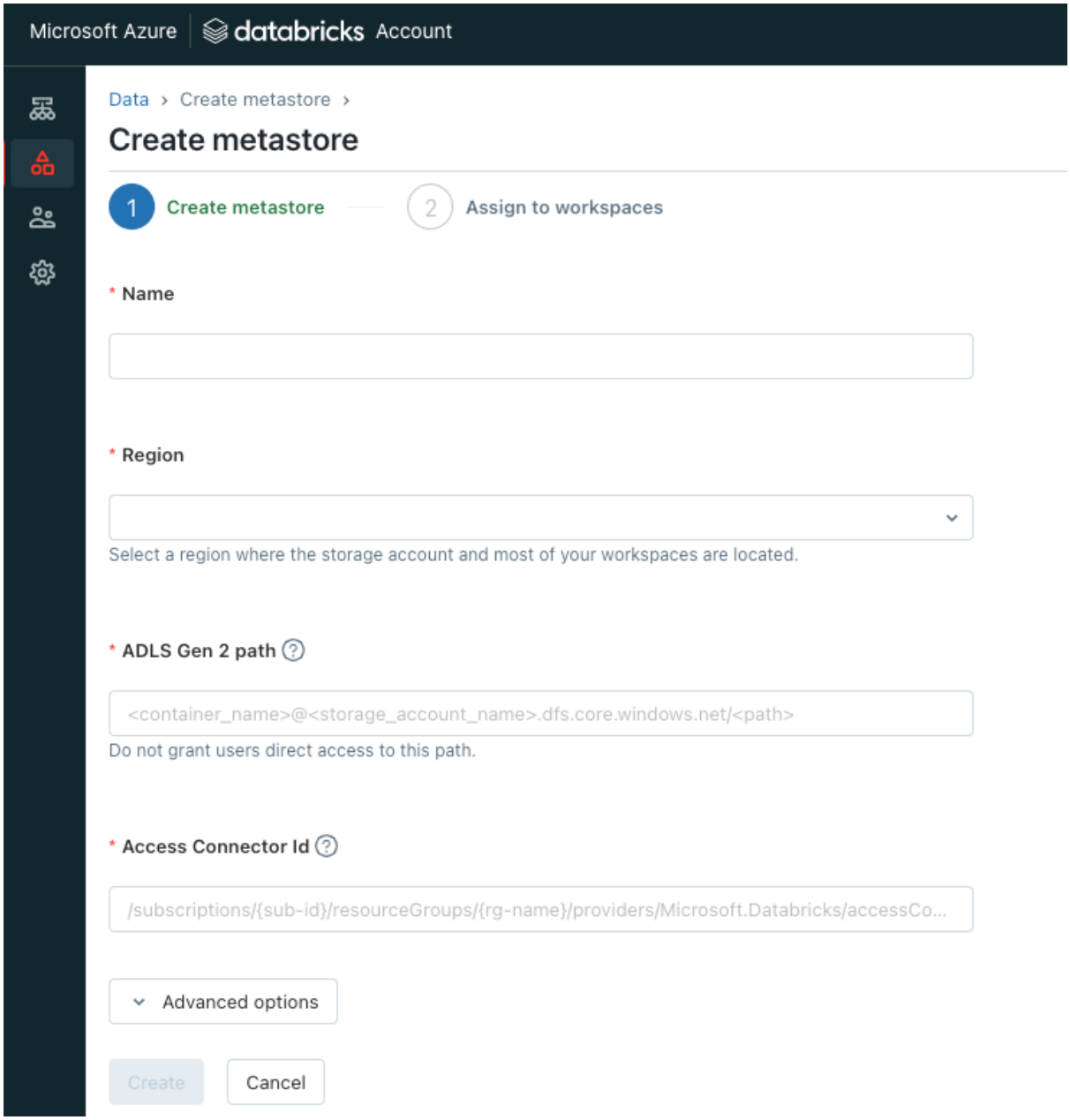
- Centralized policies that apply across workspaces.
- Attribute-based and table/column-level permissions.
- Secure sharing of data between teams and external partners.



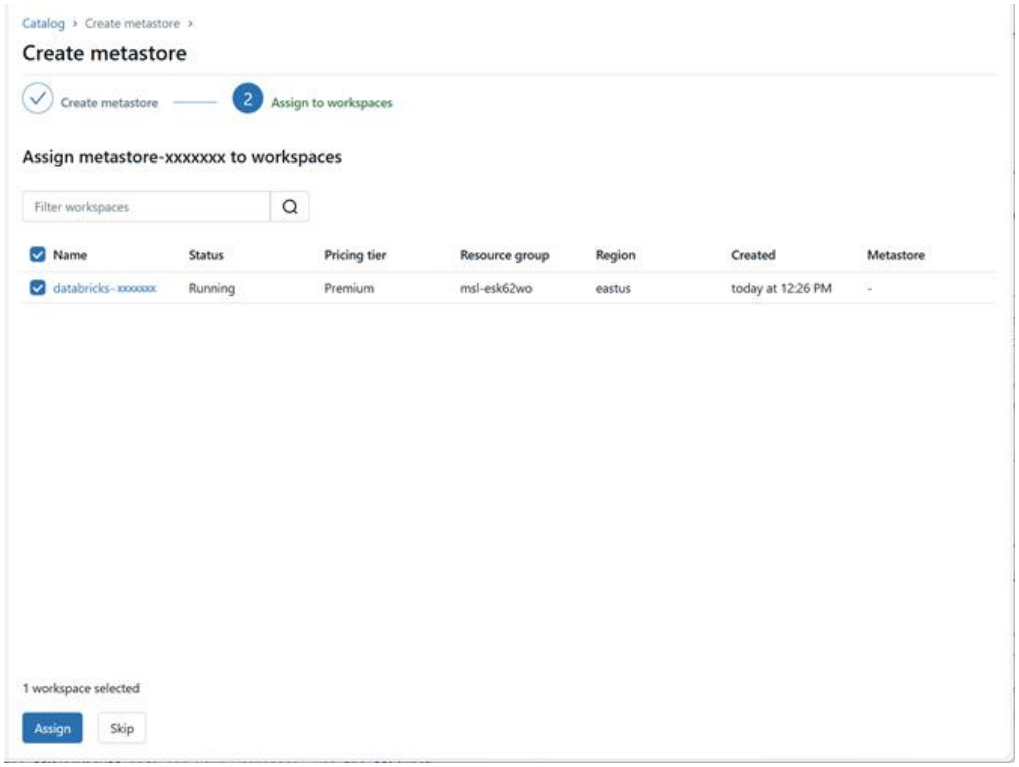
Steps to Set Up Unity Catalog:

1. **Enable Unity Catalog**
   - Ensure that the Databricks account admin has enabled Unity Catalog for the account.
   - Verify the region compatibility and prerequisites.
2. **Create a Metastore**
   - A metastore is the central metadata repository for all Unity Catalog objects.
   - It must be created and assigned to workspaces.





3. **Configure Storage**
   - Attach a cloud storage bucket (e.g., AWS S3, Azure Data Lake Storage, or GCP Cloud Storage).
   - Set up storage credentials and external locations to securely store managed and external tables.
4. **Assign Workspaces to Metastore**
   - Link one or more Databricks workspaces to the created metastore for centralized governance.



5. **Set Up Identities and Permissions**
    - Configure user groups and roles via the identity provider or Databricks account console.
    - Apply role-based access controls (RBAC) at catalog, schema, table, or column levels.
6. **Configure Audit and Monitoring**
    - Enable audit logs to track user activity.
    - Set up monitoring dashboards for compliance and security.
7. **Leverage Governance Capabilities**
    - Start using data discovery, lineage, and fine-grained access policies.
    - Continuously monitor and refine governance practices.