Cloud Data Governance & Catalog – Azure Case Study

Srinivasa Gopal – Principal Technologist, Customer Success
Sachin Jain – Principal Technologist, Customer Success
Housekeeping Tips

• Today’s Webinar is scheduled for 1 hour

• The session will include a webcast and then your questions will be answered live at the end of the presentation

• All dial-in participants will be muted to enable the speakers to present without interruption

• Questions can be submitted to “All Panelists” via the Q&A option and we will respond at the end of the presentation

• The webinar is being recorded and will be available on our INFASupport YouTube channel and Success Portal - where you can download the slide deck for the presentation. The link to the recording will be emailed as well.

• Please take time to complete the post-webinar survey and provide your feedback and suggestions for upcoming topics.
Feature Rich Success Portal

- Bootstrap trial and POC Customers
- Enriched Customer Onboarding experience
- Product Learning Paths and Weekly Expert Sessions
- Informatica Concierge
- Tailored training and content recommendations
More Information

Success Portal
https://success.informatica.com

Communities & Support
https://network.informatica.com

Documentation
https://docs.informatica.com

University
https://www.informatica.com/in/services-and-training/informatica-university.html
Safe Harbor

The information being provided today is for informational purposes only. The development, release, and timing of any Informatica product or functionality described today remain at the sole discretion of Informatica and should not be relied upon in making a purchasing decision.

Statements made today are based on currently available information, which is subject to change. Such statements should not be relied upon as a representation, warranty or commitment to deliver specific products or functionality in the future.
Introduction to CDGC
Cloud Metadata Platform

Cloud native open metadata platform that allows pervasive use of all enterprise metadata across Informatica and non-Informatica applications. Our goal is to extract and collect all types of metadata, making you more productive by providing in-context metadata intelligence.

- **Metadata Inventory and Index**
  - Metadata Knowledge Graph with objects and relationships – search

- **Metamodel and Model Ext.**
  - Metamodel store which allows customer specific models

- **Cloud Native**
  - Multi-tenant, Elastic, Security, Availability, Serverless, Live Telemetry

- **Security**
  - Metadata Security and Permissions across CMP Applications

- **Scanner Framework**
  - Agent and Cloud based metadata extraction

- **Metadata Admin**
  - Administration for Data Resources across CMP Applications

- **Scanners**
  - Metadata Extraction from Cloud and On-Premise Data/Metadata Sources

- **Ingestion Service**
  - Scalable multi-tenant ingestion

- **Workflow and Notifications**
  - Supports end user workflows and notifications
Cloud Data Governance and Catalog

- Single solution for data intelligence
- Automated to save time on manual tasks
- Start immediately with a cloud-native solution
- Put accurate information at your team’s fingertips
Cloud Data Governance and Catalog
Core Solution Tenets

**SERVERLESS**
Serverless compute to work at any scale—there is no infrastructure to deploy or manage

**MULTI-CLOUD**
Works across all cloud ecosystems and on-premises data sources

**INTEGRATED**
Deep Integration with data management and data consumption tools

**AI**
Uses state of the art machine learning and intelligence to ensure end user productivity

**OPEN & EXTENSIBLE**
API-based access to metadata repository and functions
Application Stack
Two Applications in CDGC

**Data Governance and Catalog**
- Focused on end User consumption
- All users Contributors, Collaborators and Consumers will be on this App
- Create/Update/Delete/View all activities managed from here

**Metadata Command Center**
- Focused on Admin Activities
- All Admin activities and definition of Data Classifications and adding of New Resources will happen here
- One Admin App for multiple products
Data Governance | Data Catalog

Collaboration between the Business and Technology

Looking to define, understand and trust data to support regulatory compliance, analytics, risk management, customer experience, etc.

**BUSINESS**

**TECHNOLOGY**

Managing the growing complexity of data in multiple systems & applications. Creating a data digitalized platform to support business consume-ability

**Business Personas**
- Data Analyst
- Data Scientist*
- Privacy Officer
- Business Stewards
- Governance Managers

**IT Personas**
- Data Architects
- Data Engineers
- System Owners
- Database Admins
- Technical Stewards
DGC Model

Business

System
  Dataset
  Data Element
  (Parent / Child)

Technical

Glossary
  Data Classification
  (Glossary Associations)

Resources
  Table
  (Parent / Child)

Logical

(Data Element)

Glossary Associations

(Discovered)
Security and Secure Agent
Security

• IICS Services including CDGC are SOC2 compliant
  - https://www.informatica.com/in/trust-center.html

• IICS Security Architecture Overview
  - https://network.informatica.com/docs/DOC-18220

• IICS Services IP Addresses to whitelist on customer’s Firewall for Secure Agent to communicate
Common Security Questions

Q1: Where is metadata and data profiles stored?
A1: INFA Cloud, each Tenant data is encrypted with unique Tenant specific key

Q2: Where are Connection credentials stored?
A2: INFA Cloud or Secure Agent, customer’s choice

Q3: Is data in transit and at rest encrypted?
A3: Yes, all communication from agent to IICS is over SSL

Q4: Is there any inbound communication between IICS and Secure Agent?
A4: No, it is only outbound, Secure Agent communicates to IICS always
Runtime Deployment Location
Choose the right Runtime Environment location

On-premises Secure Agent
• Catalog Source is on-premises
• Customer is already using Secure Agent
• Large Data Sources or number of data sources
  - Co-location of metadata extraction or profiling compute with Data Sources

Serverless
• Catalog Sources are in the cloud
• Customer is okay providing a public end-point for accessing the data source
• Ideal for AWS ecosystem - DL, DW and DBs (because CDGC is AWS native)
Secure agent for metadata extraction and discovery
Easy deployment with limited requirements

• Recommended Prerequisites
  - 16 CPU cores
  - 64 GB RAM
  - 200 GB HDD

• Easy to deploy
  - Download and install the secure agent
  - Register agent using generated token

• Scale with IICS agent grouping capabilities
  - 2 or more SA for sources more than 100k assets

• All agents can execute all the capabilities
  - Metadata extraction
  - Profiling

• Set the MAXDTMProcesses on Data Integration Service parameter to 20 to reduce the profiling execution time
CDGC Secure Agent-based execution

Cloud Data Source

1. Metadata Extraction
2. Profiling

Secure Agent #1
Secure Agent #2

Ingestion

Discovery

Serverless Cluster

CDGC Repositories

Customer’s premises ←

IDMC
Azure Case Study
Azure Reference Architecture – Data Lake House

Semi-Structured
csv, json, parquet, xml
(loosely-typed)

Relational Databases
(structured)

Cloud Security Team

Cloud Dev Team

Enterprise-grade semantic models and self-service BI

Consumer of structured data

Platform
- Azure Cost Management and Billing
- Azure Automation Provisioning
- Azure Policy
- Azure Active Directory
- Azure Security Center
- Azure Key Vault
- Azure Monitor

© Informatica. Proprietary and Confidential.
Azure Data Lake House with CDGC
Analytics Governance – Cloud Data Lake House

End User Consumption

• Data Analyst
  - View Goal Oriented Dashboards
  - Browse through Business and Technical Assets
  - Trust Data Assets with easily extracted and powerful lineage views
  - Trust Data through Integrated Data Quality
Analytics Governance – Cloud Data Lake House
Governance and Catalog Program

**Governance Setup**

- Data Steward
  - Set Governance Program
    - Set up Org Specific Attributes
  - Configure Workflows
  - Content Creation
    - Create Business Glossaries and other business assets manually going through governance workflows
    - Create Relationships
    - Bulk upload business terms for CDRs
  - Search, Browse and Dashboarding
    - Basic and Advanced Search
    - Browse Experience
    - Create and Edit Dashboard

- Catalog Administrator
  - Create Classifications*
  - Configure Metadata Scanners
  - Scan CDL/DW - Sql Server, ADLS, ADF, PBI and Synapse
  - Perform Connection Assignments

- Data Steward (Discovery)
  - Curates IGA recommendations*
  - Technical Assets without Business terms
  - Assets without Stakeholders

- Data Steward (Curation)
  - Create Systems/Datasets
  - DQ Linkage
  - Business Lineage
  - AI Model Governance

* Denotes optional tasks.
DEMO

CDGC – AZURE CASE STUDY
References

- Cloud Data Governance and Catalog: Click Here
- Introduction and Getting Started: Click Here
- Governance: Bulk Upload Business Metadata
- Metadata Command Center:
  - Catalog Scanner Configuration
  - Creating Catalog Source
  - Configure Runtime Environments