ENTERPRISE DATA CATALOG

Vijay Bhasker Reddy Vedire[1]

Dr. Aruna Varanasi[2]

[1]Student, Department of Computer Science and Engineering, Sreenidhi Institute of Science and Technology, Hyderabad, TS, India

[2]HOD, Department of Computer Science and Engineering, Sreenidhi Institute of Science and Technology, Hyderabad, TS, India

ABSTRACT

Data is the lifeblood of our economy, and data-driven companies turn their data assets into revenue and profits. The first step in any data-driven digital transformation initiative is to manage your data as an enterprise asset: take inventory of it, assess its value, and maximize its use—just like you do with other significant capital and operational investments. Data is diverse and distributed across many different departments, applications, and data warehouses and data lakes (some on-premises, others in the cloud), making it a challenge to know exactly what data you have and where. As data sources proliferate, the data landscape becomes even more complex. Informatica® Enterprise Data Catalog is an AI-powered data catalog that provides a machine learning-based discovery engine to scan and catalog data assets across the enterprise—across multi-cloud and on-premises. Enterprise Data Catalog is powered by the CLAIRE® engine, which provides intelligence by leveraging metadata to deliver recommendations, suggestions, and automation of data management tasks. This enables IT users to be more productive and business users to be full partners in the management and use of data. Informatica Enterprise Data Catalog provides data analysts and IT users with powerful semantic search and dynamic facets to filter search results, detailed data lineage, profiling statistics, data quality scorecards, holistic relationship views, data similarity recommendations, and an integrated business glossary.

Keywords: Warehouse, Catalog, Enterprise

I. INTRODUCTION

Enterprise Data Catalog helps you analyze and understand large volumes of metadata in the enterprise. You can extract physical and operational metadata for a large number of objects, organize the metadata based on business concepts, and view the data lineage and relationship information for each object.

The key concepts in Enterprise Data Catalog include catalog, resource, resource type, scanner, and schedule. Catalog stores all the metadata extracted from sources. Resource type represents different metadata source systems. Resource is a representation of a resource type, such as Oracle, SQL Server, or PowerCenter. Scanners fetch the metadata and save it in the catalog. Schedules determine the intervals at which scanners extract metadata from the source systems and save the metadata in the catalog.

II. ENTERPRISE UNIFIED METADATA ARCHITECTURE

The Enterprise Unified Metadata architecture consists of applications, services, and databases. The applications layer consists of client applications, such as Enterprise Information Catalog. The services layer has application services, such as the Catalog Service, Data Integration Service, and Model Repository Service. Enterprise Unified Metadata requires the Catalog Service to extract metadata from data sources and manage the administrative tasks. The databases layer consists of the Model repository and internal or external Hadoop cluster for metadata storage and analysis. Data sources and metadata sources include source data repositories, such as Oracle, Microsoft SQL Server, PowerCenter repository, and SAP Business Objects.

III. MODELLING AND ARCHITECTURE

Basic architecture of Enterprise Data Catalog

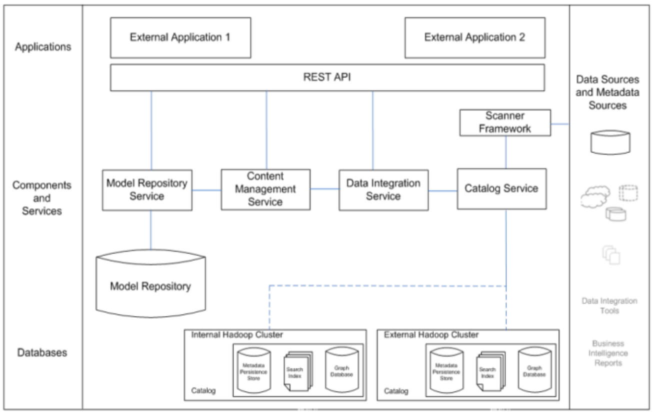

Figure 1: System Architecture.

IV. DATA SOURCES SUPPORTED INCLUDE

- Databases/Data warehouses: Oracle, MS SQL Server, SQL Scripts, Sybase ASE, IBM Netezza, Teradata, JDBC, SAP HANA, SAP BW, SAP BW/4HANA, Snowflake, Stored Procedures
- Big Data: Cloudera Navigator, Hive (Cloudera/Hortonworks/MapR/IBM BigInsights/EMR), HDFS, Hortonworks Atlas, Cassandra, MongoDB, Kafka, Greenplum
- Mainframes: DB2 z/OS, DB2 i5/OS, COBOL, JCL
- BI and Analytics: SAP BusinessObjects, Tableau, Microsoft Power BI, Cognos, MicroStrategy, OBIEE, QlikView, Qlik Sense, Microsoft SSRS and SSAS, SAS
- ETL: Informatica PowerCenter® , Informatica Data Engineering Integration, Informatica Intelligent Cloud Servicessm, Informatica Data Integration Hub, Microsoft SSIS, IBM InfoSphere DataStage, Oracle Data Integrator, Talend Data Integration, AWS Glue
- Business Glossary: Informatica Axon Data Governance, Informatica Business Glossary
- Data Modeling: Erwin Data Modeler, SAP PowerDesigner Enterprise Applications: Salesforce, Oracle, Workday, Informatica MDM, SAP ECC, SAP S/4 HANA
- File Systems: Microsoft SharePoint, Microsoft OneDrive, Windows/Linux Filesystems
- File Formats: MS Excel, MS Word, MS PowerPoint, Adobe PDF, Flat Files, CSV, Delimited, XML, JSON, Avro, Parquet
- Cloud Platforms: AWS S3, AWS Redshift, Azure SQL DB, Azure Synapse Analytics, Azure ADLS, Azure ADLS Gen 2, Azure Blob, Google Cloud Storage, Snowflake, Google BigQuery

V. CONCLUSION

When EDC was launched ,there were only few resource types with only Profiling and Scheduling Options.

Now there are 35 resource types , UCF Resources – Erwin, Greenplum, clique, SAP BO, bteq, data domains have been added with data rules and metadata rules ,data domain groups have been added, composite data domains have been added, more system scanners have come, access control has been added to give permission to users and search i.e. the GOOGLE of Enterprise level has made customer's life easier as they can effectively extract metadata from the large set of data. End to end scenarios for a particular customer has been looked into showing the end – to – end lineage including relationships between objects.

Now with UI Automation we are moving towards Build Integration and Jenkin Integration. Customers have been very satisfied with our product and we are working towards improving the quality of our product and fulfilling every need of our customers.

VI. REFERENCES

[1] The above project is owned by Informatica®

[2] https://www.informatica.com/

[3] [online] Available: https://hadoop.apache.org/.

[4] [online] Available: https://hive.apache.org/.

[5] F. Chang et al., "Big table: A distributed storage system for structured data", OSDI, pp. 205-218, 2006.

[6] [online]Available:hadoop.apache.org/docs/r1.2.1/hdfs_design.html.

[7] M. Zhu and T. Risch, "Querying Combined Cloud-Based and Relational Databases", International Conference on cloud and service computing (CSC) 2011, pp. 330-335.