Tiger Analytics

Tiger Data Fabric on AWS

Date: 06/12/2022 Author: Muthu G

Problem Statement and Proposed Solution

Problem Statement

- Need for the Standard Frameworks to Ingest data quickly into the Datalake Platforms
- Need for the Standard Frameworks to support Data file Schema, Data Quality Check , Data Validation and Verification.
- Need for Standard Frameworks to perform data pipeline monitoring and reporting
- More Effort spend on Onboard Source files

Solution

- Easier to ingest data into AWS data lake
- AWS Cloud native and Open Source technologies
- **Data Provenance**: data quality, data masking , lineage, Recovery and Replay Audit Trail, Logging, Notification
- **Intuitive** User Interface to quickly focus on key data elements

Benefits

- Onboard Data Quickly with configuration entered using UI
- Auto ELT data pipeline Generation.
- Reduction in documentation and Testing Effort for onboarding different source files
- Lesser time spend in Data Ingestion and project team can focus more on Product Features and KPI generations

Automated Datalake Framework

Standardized Metadata & Audit process

Orchestration using Airflow

Easy Deployment

Time driven/Event Driven process

Accelerated pipeline build process

Data Ingestion, Data quality, Masking & Standardization

Data published as Parquet, Athena Tables, Redshift

Data consumption & publishing from disparate sources

In-built standard Data quality rules

UI Driven - setup & configurations

Data masking

Overview

FilesSupported

CSV

json

XML

Parquet

DatabaseSupported

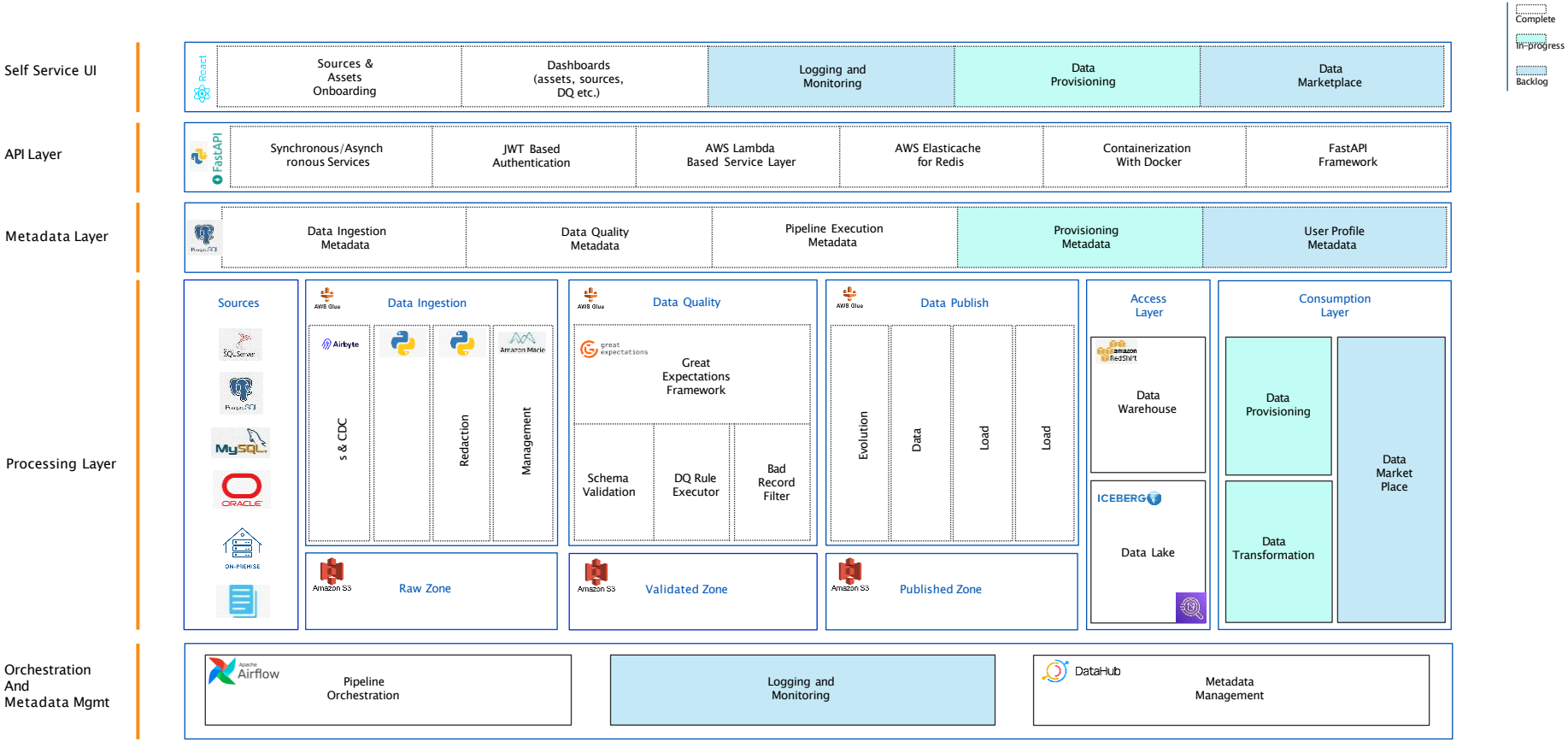
MySQL

PostgreSQL

ORACLE

Microsoft SQL Server

High Level Design



Large Level Components

- Data Ingestion: AWS Data Fabric connects with a multitude of data sources and accommodates diverse formats and databases. The platform leverages Airbyte, an open-source data integration platform, to streamline the connection to various sources and support Change Data Capture (CDC) for real-time data updates. In addition, it employs Format-Preserving Encryption (FPE) to mask data securely during ingestion and utilizes AWS Macie for automated data scans to ensure data privacy and security.
- Data Quality Framework: Utilizing the robust Great Expectations framework, the platform ensures and monitors the quality of ingested data. Interactive dashboards provide users with a comprehensive view of data health status, flagging anomalies and potential quality issues.
- Data Standardization & Publishing: AWS Data Fabric uses reference data lookup for data standardization, supports ACID transactions powered by Apache Iceberg on AWS Athena, and implements Slowly Changing Dimensions (SCD) Type-2 within AWS Redshift for preserving historical data accuracy.
- Data Provisioning: With automated data provisioning and access control mechanisms, AWS Data Fabric guarantees the timely delivery of appropriate data to the right users, ensuring robust data security.
- Metadata Management & Lineage: Using LinkedIn's open-source tool, DataHub, the platform offers advanced metadata management capabilities. DataHub provides centralized repositories for technical and business metadata and an organized lineage view, enabling effective tracking and data governance.
- Fast-API based API Framework: The Fast-API framework serves as the application programming interface layer, enabling seamless interaction between the user interface and backend systems.
- UI Driven Data Management: The platform features a React-based UI, providing a low/no-code platform for users with varying levels of technical expertise. This allows for more efficient and intuitive data management.
- Medallion Architecture: The implementation of the Medallion architecture facilitates the handling of complex data structures and provides a logical view of data storage and access.
- Pipeline Orchestration: The platform orchestrates data pipelines using Apache Airflow templates, ensuring streamlined handling of intricate data workflows. These pipelines are built using AWS Glue, Apache Spark, and a variety of open-source libraries, thereby promoting efficient data processing and management.
- RDS Postgres Metadata Store: For a reliable and robust metadata management solution, all metadata is securely stored in an RDS Postgres database.

TIGER ANALYTICS

Thank You

Questions & Feedbacks

www.tigeranalytics.com