Accelerate Analytical Data Lake (ADL) with vDataHub framework powered by AWS

Organizations operating on legacy data platforms face an increased total cost of ownership, lack of business agility, and slower time to market, making it challenging to meet their SLA commitment. Furthermore, complex legacy technologies make it difficult for companies to drive innovation. Hence, businesses that depend on legacy data platforms need to modernize their current data ecosystem for faster market response and reduced operational costs.

Virtusa's vDataHub framework is a modern data warehouse solution built using AWS native services, which is robust, easily configurable, reusable, faster, and efficient. It automates data pipelines from source to ADL, including reporting and visualization for the end customer. It is a spark-based integration framework on AWS EMR orchestrated using the Amazon Managed Workflow for Apache Airflow (MWAA), which integrates the entire process.

The core features include:

- Foundation blueprint to set up end-to-end ADL on the cloud
- Fully configuration-based data ingestion framework
- Configurable standardization and data transformation
- Automated reconciliation, data integrity check, data pipeline, DataOps, DevOps, and data governance

Why clients choose Virtusa to automate the data pipeline?

Virtusa leverages the vDataHub framework to automate ingestion, standardization, and configuration to reconcile processes in the data pipelines. Clients choose Virtusa's vDataHub framework for:

15-30% faster configuration and development with accelerated frameworks

Faster loading of large multilayer source data

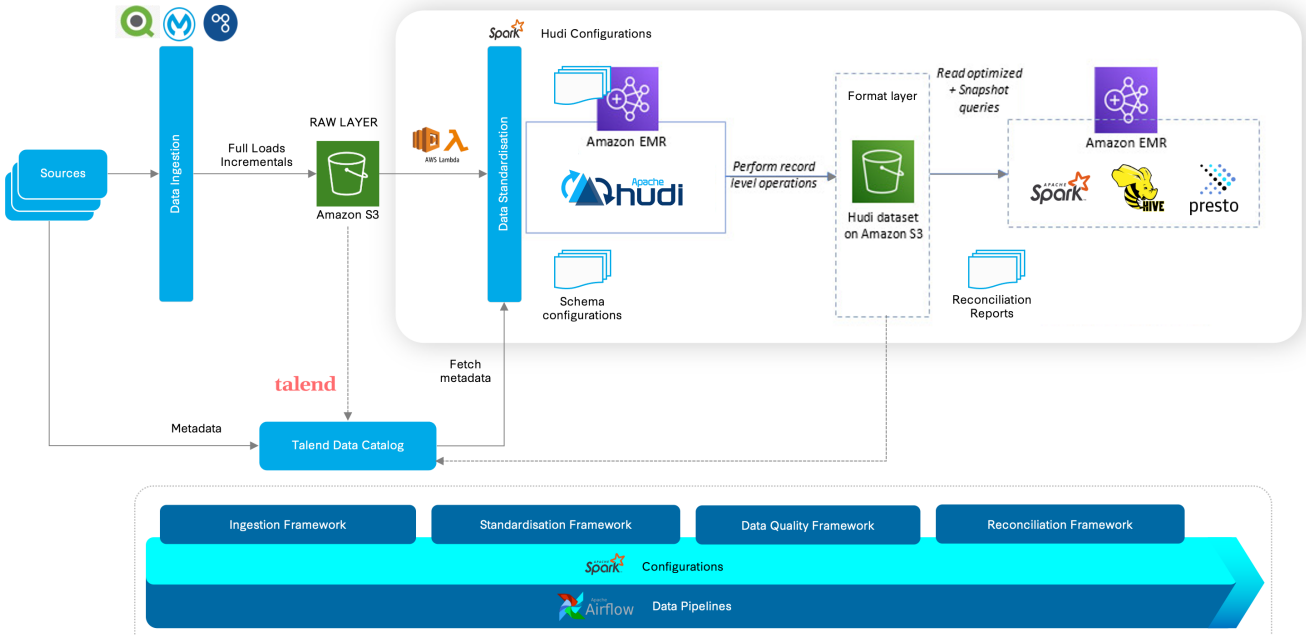
Seamless integration with AWS components

Ready to use templates for projects

Capabilities

Modernize the current data ecosystem with the vDataHub framework

vDataHub framework offers a data management tool to modernize the current data ecosystem for quick market response and lower operational costs. It automates various possibilities in the data lake environment throughout phases using key accelerators and framework components. vDataHub framework ensures complete automation for the data ingestion, validation, schema creation, and incremental processing. The frameworks also include DevOps and DataOps to automate development, integration testing, and deployment.



The key accelerator components of the vDataHub framework include:

Foundation (Blueprints)

Terraform IaC as blueprints for AWS infrastructure provisioning for AWS landing zone, AWS services, EMR cluster, data lake, S3, MWAA, etc.

Ingestion framework

ELT framework: PySpark-based configuration driven ELT framework can be customized to load data on the cloud

Data Lake as a Service: Accelerate data lake built with data services, ingestion, and transformation solutions. It allows business users to extract, select, schedule, and integrate source systems into the data lake

Data quality framework

ML-based DQ Engine: To identify key trends, patterns, and anomalies DQ engine is used in the vDataHub framework. It is integrated with data lineage and data profiling tools. It ensures sanity checks, statistical checks, and AI/ML-based advanced checks on the source data before loading into ADL

Standardization framework

Standardizes data sets in the RAW layer. It includes setting up schema by integrating data lineage tools, performing data quality checks, defining and configuring change data capture (CDC) from RAW to format layer

Reconciliation framework

Data reconciliation framework verifies the target data against source data to validate mappings and any transformation logic. It covers any missing records, missing data, duplicate records, and poorly formatted data. In the end, it produces a data reconciliation report

DataOps

The formal discipline of data gathering, assembly, management, curation, and publication that includes:

- Data validation includes cleanliness, completeness, alignment, accuracy, granularity, timeliness,
  and latency
- Metadata enrichment including descriptive, structural, administrative, reference, and statistical metadata
- Governance to ensure availability, usability, integrity, security, and usage compliance across use cases
- Data cataloging and indexing so the data can be easily searched, found, accessed, understood, and re-used

For more information, please contact us at **marketing@virtusa.com**