virtusa

vDataAid

Simplifying business complexities for a better experience

Organizations are increasingly prioritizing data modernization to deliver actionable insights and API-driven application integrations and respond to an ever-evolving business environment. Lack of reusable and configurable code, insights into the data loading process and metrics, and integrated data validations make the data unreliable for analytics and result in increased manual effort and errors during the development/maintenance phases of the project.

Virtusa's vDataAid solution helps organizations accelerate data modernization platform development by deploying a configurable and metadata-driven solution. The solution covers different phases, including data ingestion, data validation, and Slowly Changing Dimensions (SCD) data processing. It combines multiple data frameworks, such as Generic Data Ingestion, Data Validation, and SCD Type 1 and Type 2, that are easily configurable, customizable, and deployable for any Microsoft Azure platform.

Features

vDataAid is a low-code solution that audits each level in the data loading process to provide in-depth insights and control the pipeline behavior. Some of its key features include:

Data ingestion from all sources without creating multiple ingestion pipelines

- Ingests data from various data sources like RDBMS (Oracle, DB2, SAP, Teradata, etc.) and files (Txt, CSV, Excel, JSON, XML, etc.)

Auditing

- Stores and manages statistics, including start time, end time, batch, and job status, etc.
- Captures the source & target count, data validation count, data validation summary.

Job control and dependencies

- Job control is provisioned by performing the pre-checks in batch/jobs before proceeding with execution, and also it enables the restartability feature for failed jobs.
- Maintains the metadata of the jobs and their dependencies.

Job and Object lineage

- The solution helps in maintaining the Job and Object lineage across multiple data layers in a data lake.

Data validation framework

- Integrated, configurable, and customizable data validation framework for data quality checks like not null, uniqueness, min, max, is between, length, mean and median, text matches, etc.

SCD Type 1 and Type 2 framework

- Integrated and configurable SCD Type 1 and Type 2 framework for maintaining the most up-to-date data in the data lake.

Error logging and notifications

- The solution captures the error at every activity and logs in the auditing tables.
- It also sends the error notification emails on errors in the pipeline.

Dashboards and reports for operational insights

- Use tools like Power BI to create dashboards and reports for additional insights.

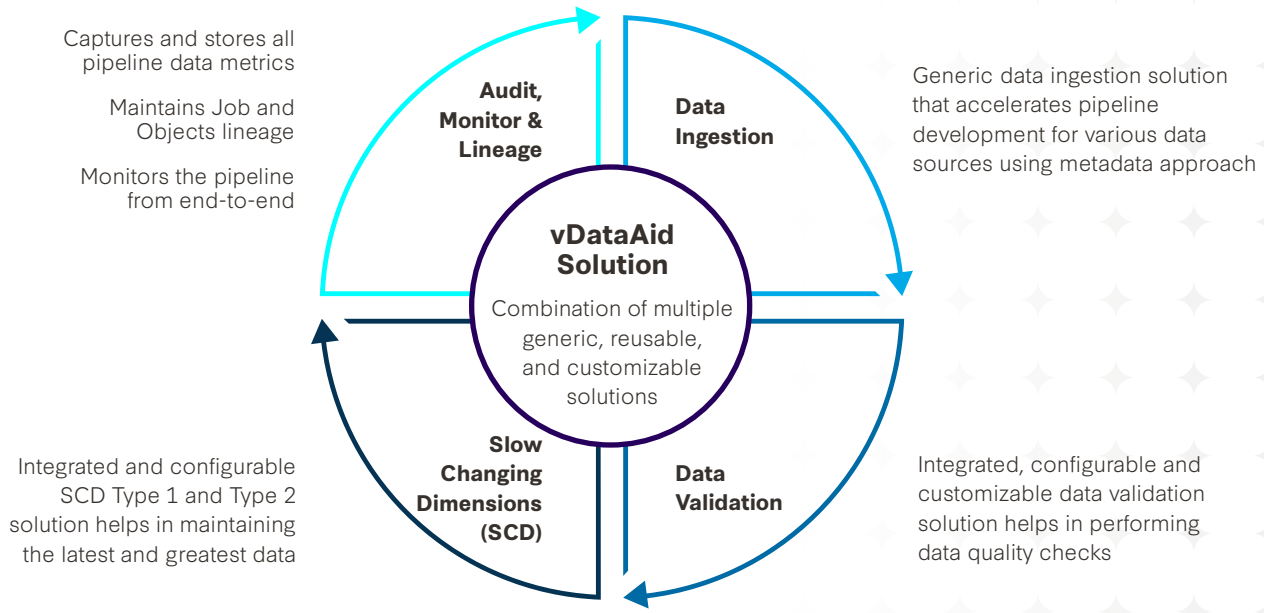
How does it work?

The vDataAid solution is developed using Azure Data Factory for data ingestion and Spark Notebook for data validation. Azure data integration pipeline is a generic pipeline used for data ingestion & data validation that is completely driven by metadata.

For instance, with any data source configuration, the first step is to capture the ingestion details like source & target paths, objects to be ingested, etc., in pre-configured metadata tables. Then, we use the single generic pipeline for ingestion, validation, and transformations (Slowly Changing Dimensions - SCD) of all the objects without creating and maintaining multiple pipelines.

To deploy and leverage the solution in a client environment, we need to:

- Upload and configure the vDataAid solution code to the client's Git repo
- Deploy the schema for metadata tables in the client database
- Use the power app screen to insert metadata tables
- Configure the pipelines based on the data sources
- Run the vDataAid integration pipeline
- Create dashboards to visualize the metrics



Benefits

vDataAid solution is a one-stop-shop, plug and play, configurable, and extensible solution for Azure enabling seamless data integration. It helps in:

- High efficiency of operations: 30-40% reduced effort on pipeline development.
- Plug and play, configurable, and metadata-driven solution helps in reducing the overall development time.
- Faster onboarding of new data set and sources.
- High code consistency and reduction of development errors.
- Processing quality data across layers in the data lake.
- Better insights to improve business process: overall pipeline health, performance, data loading, quality, and completeness using dashboards.

For more information, please contact **marketing@virtusa.com**
or visit **www.virtusa.com**