Hard Rock Digital

Donal Gonsalves

Data Vault 2.0 Methodology

18th Nov 2024

Data Vault 2.0 methodology:

Q1. What technology/technologies will be used to implement this storage solution?

While implementing a Data Vault 2.0 solution, I will consider various factors such as data volume, complexity, real-time requirements, and organizational preferences. Here's how I propose to approach it:

1. **Data Warehouse (Snowflake):**
   - ➔ I will use Snowflake for data warehousing due to its scalability, performance, easy integration, flexibility, support for advanced analytics on structured and semi-structured data, and as a managed service.
2. **Data Orchestration (Apache Airflow):**
   - ➔ I will use Apache Airflow for scheduling and managing ETL/ELT pipelines, automating data loading and processing tasks efficiently.
3. **Data Integration (Apache Kafka):**
   - ➔ I will use Apache Kafka for event processing in high-throughput, scalable, and real-time data architectures. Its flexibility and replay capabilities make it ideal for processing, storing, and replaying JSON messages for analytics.
4. **Data Lake (Azure Data Lake Storage):**
   - ➔ I will use Azure Data Lake Storage because it is cost-effective, scalable, and flexible. It integrates seamlessly with modern tools and supports event-driven architectures.

Note: Databricks can be used on the top of data lake if the data is complex and required heavy workflow designing.

Q2. Describe the table structure, attribute composition, and data types. The format of the description is open-ended; use whichever is most convenient or familiar for you.

For this question, I have showcased the table structure, attribute composition, and data types by referring to the Data Vault 2.0 methodology. Please refer to the attached Excel file for reference.

Table Structure and
Data Types.xlsx

Q3. What additional components need to be developed to support your solution?

Based on the technologies implemented for Data vault 2.0, additionally I will suggest data analytical tools and monitoring and alerting tools.

Data Analytical tools:

➔ I will use Power BI or Tableau for analyzing historical data from the data warehouse and presenting meaningful insights to the business team to support operational decisions.

Monitoring Tools:

➔ I will use AWS CloudWatch or Datadog to monitor pipeline status, data quality, and system performance. These tools will help identify potential discrepancies or bottlenecks in data processing.