Building Real-Time Analytics Pipelines

A Modern Data Cloud Solution

The Problem:
Data Silos, Slow Pipelines, Costly Analytics

Companies deal with massive data stored in S3, spreadsheets, and external sources.

Traditional tools struggle with:

- Slow data ingestion.
- Complex ETL pipelines.
- High costs for scaling.

Real-World Example:
Imagine a retail company with millions of sales records in S3 and weather data impacting sales. How do they combine these datasets for actionable insights?

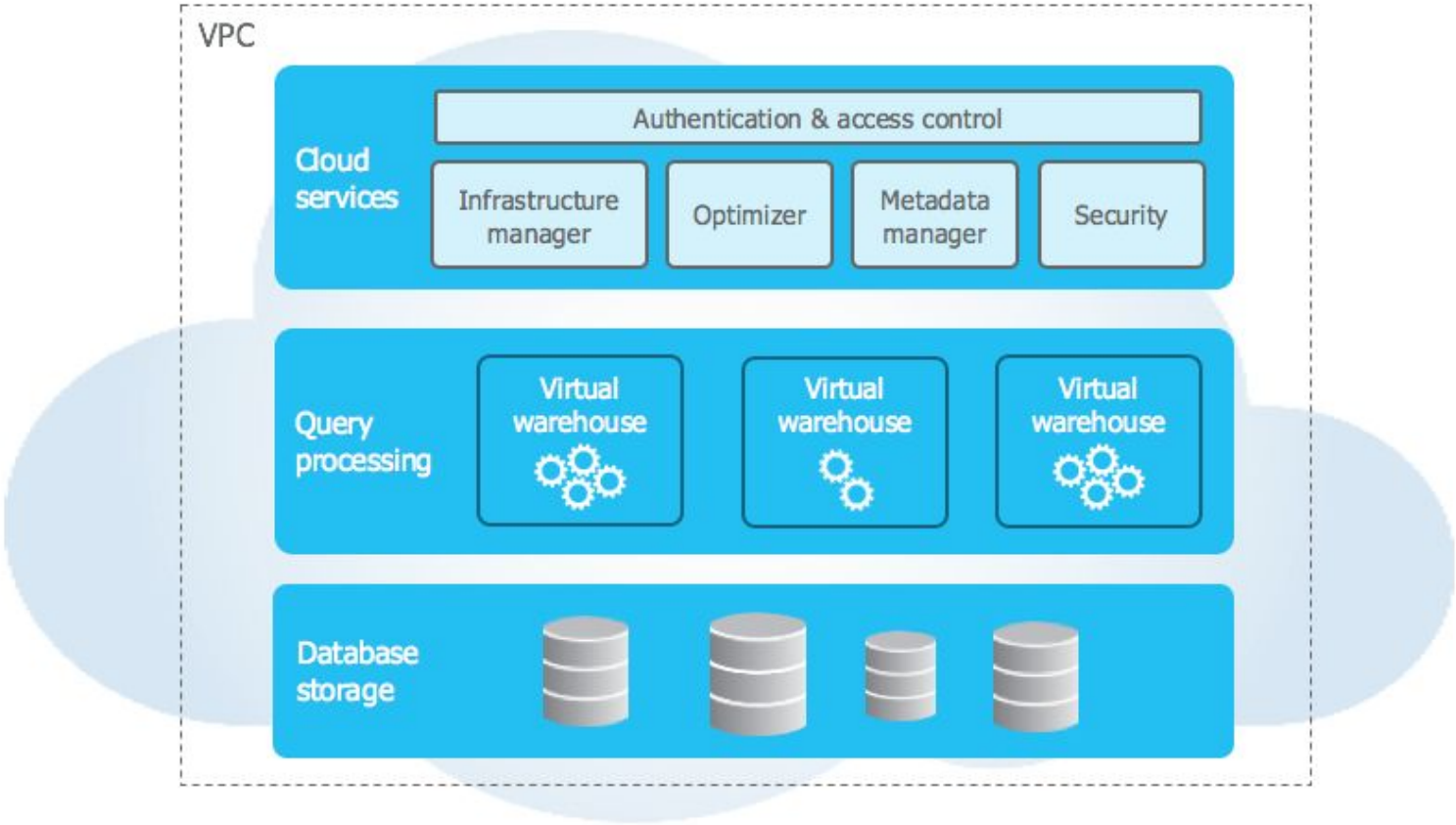
The Solution: Snowflake's Modern Data Cloud

Cloud-based data platform that allows you to store, process, and analyze massive amounts of data without worrying about the underlying infrastructure.

- Zero Management: No infrastructure to manage.
- Scalability: Handles petabytes of data effortlessly.
- Data Sharing: Access external datasets like weather data via Snowflake Marketplace.

Handles the infrastructure for you so you can focus on insights!

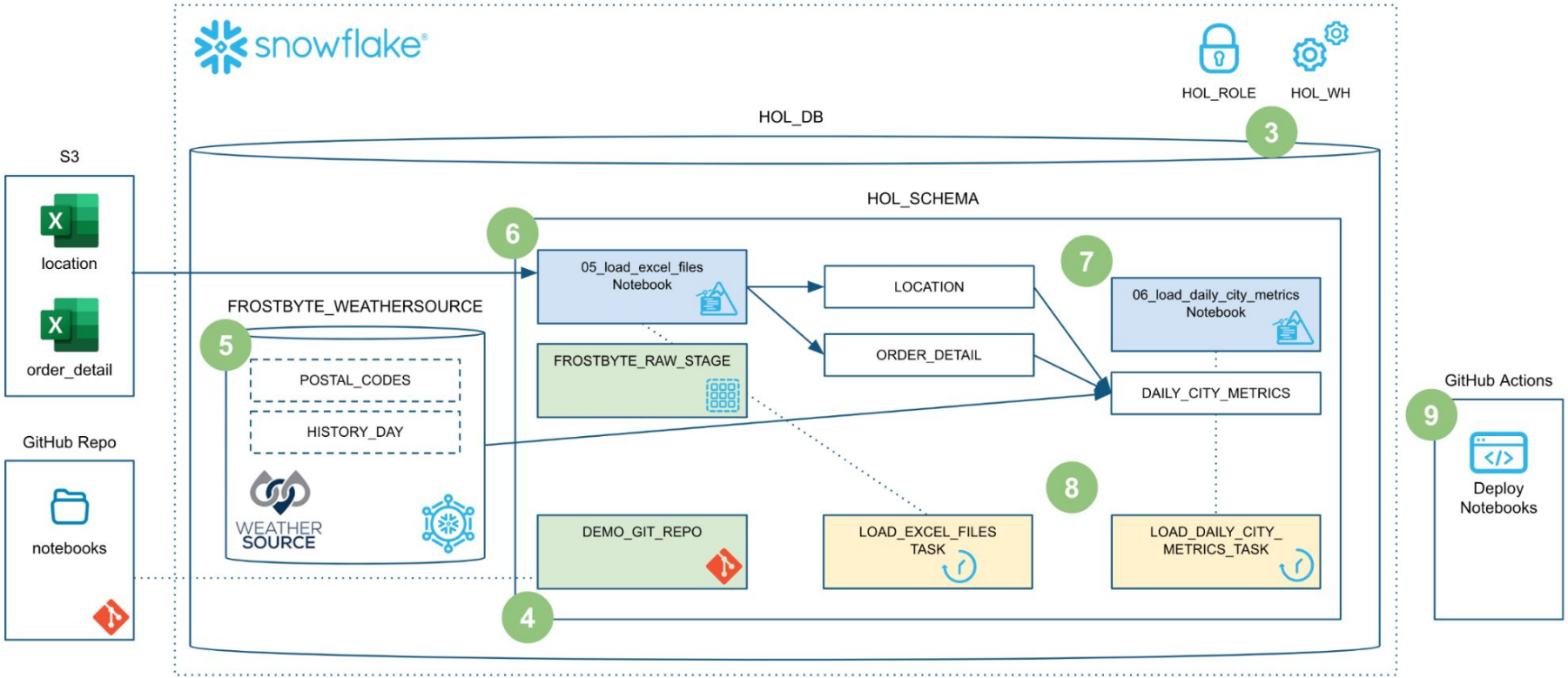
ARCHITECTURE



Source: Snowflake documentation

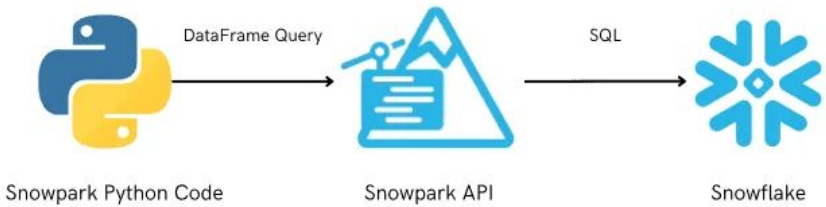
Demo Time!

A retail company analyzes how weather impacts daily sales and optimizes inventory


snowflake
HOL_ROLE
HOL_WH
HOL_DB
3
S3
location
order_detail
HOL_SCHEMA
6
05_load_excel_files
Notebook
LOCATION
7
06_load_daily_city_metrics
Notebook
FROSTBYTE_WEATHERSOURCE
5
POSTAL_CODES
HISTORY_DAY
WEATHER SOURCE
FROSTBYTE_RAW_STAGE
ORDER_DETAIL
DAILY_CITY_METRICS
GitHub Repo
notebooks
8
DEMO_GIT_REPO
LOAD_EXCEL_FILES
TASK
LOAD_DAILY_CITY_
METRICS_TASK
4
GitHub Actions
9
Deploy
Notebooks

Services Used:

- Snowpark Notebooks: For Python-based data processing.
- Snowflake CLI: For seamless deployment and management.
- Tasks & DAGs: For orchestrating pipelines (Airflow-like syntax).
- Use Python Management API to manage tasks (pause, resume, debug).



Thank you!