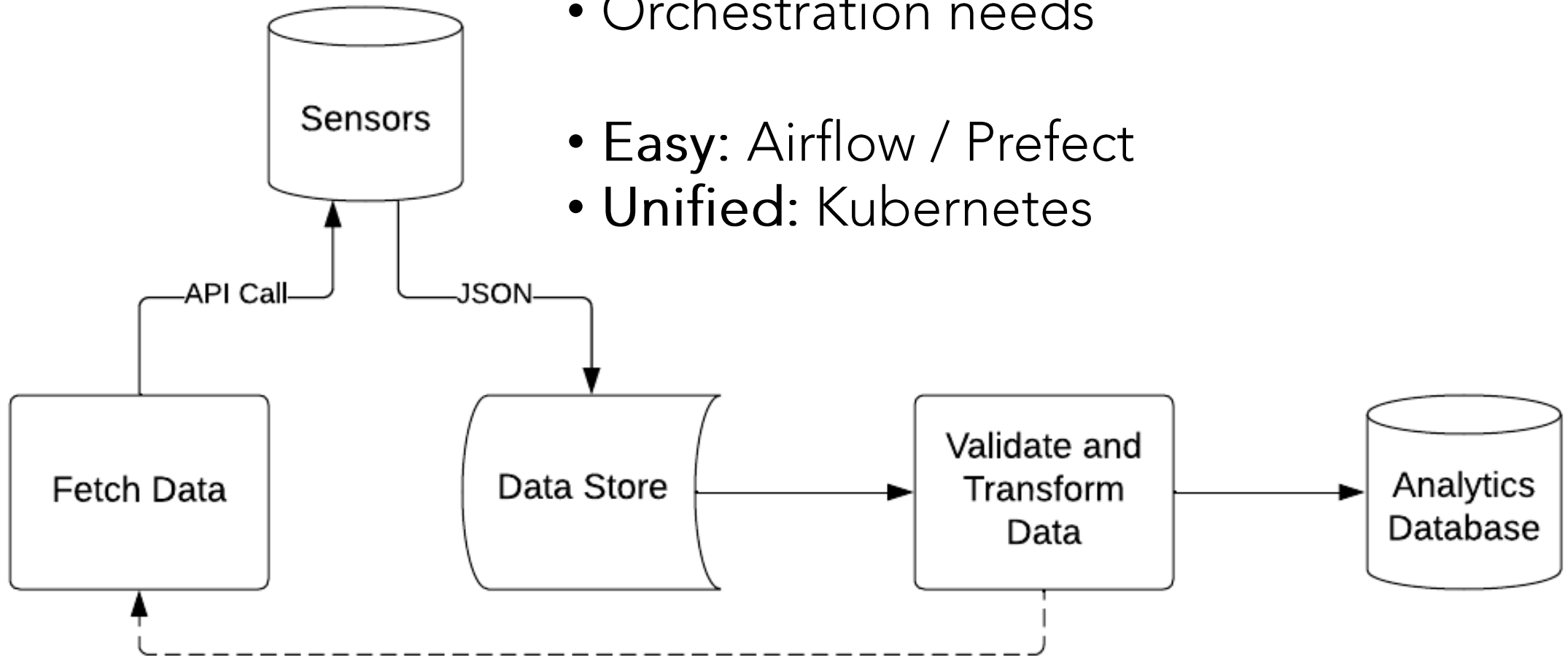# Productionizing Trash Can-alytics

A proposal by Herschel Pecker

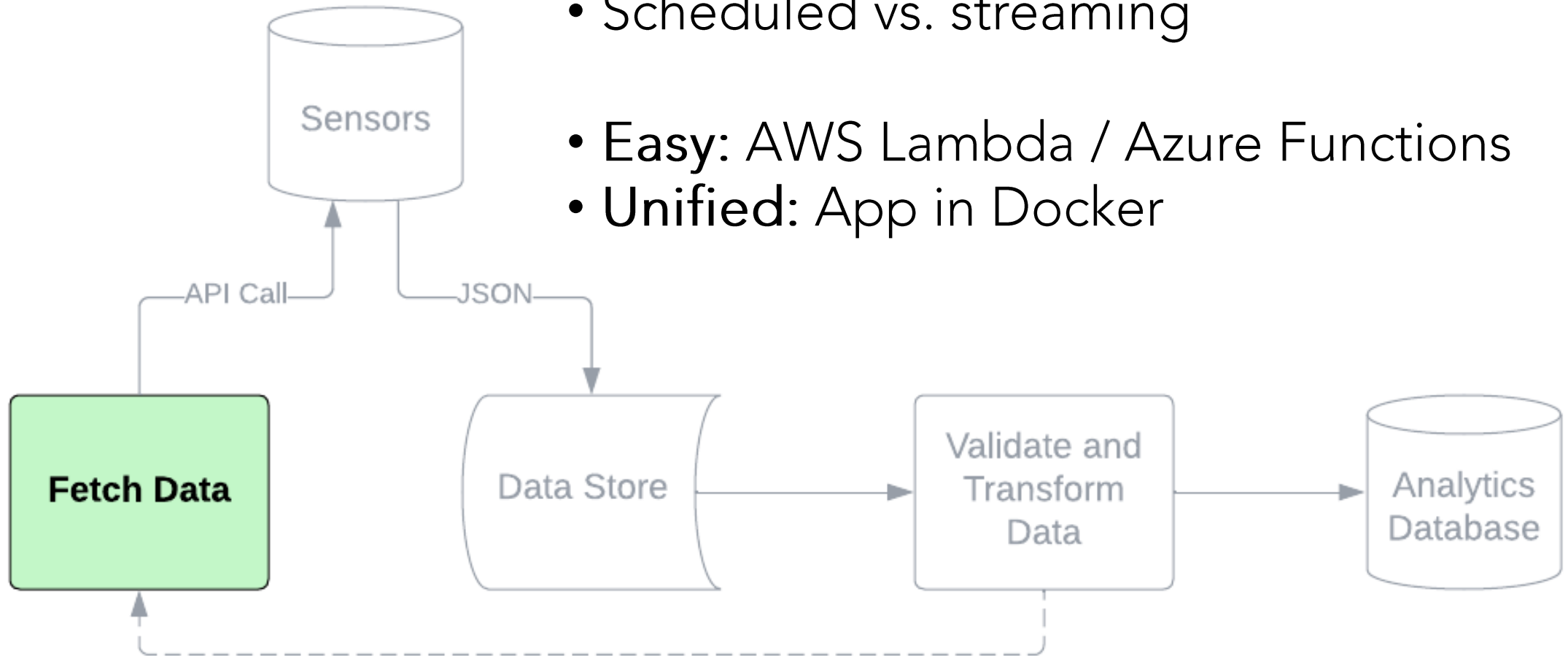# The ETL Process

- Orchestration needs

- **Easy:** Airflow / Prefect
- **Unified:** Kubernetes

Sensors

API Call   JSON

Fetch Data   Data Store   Validate and Transform Data   Analytics Database
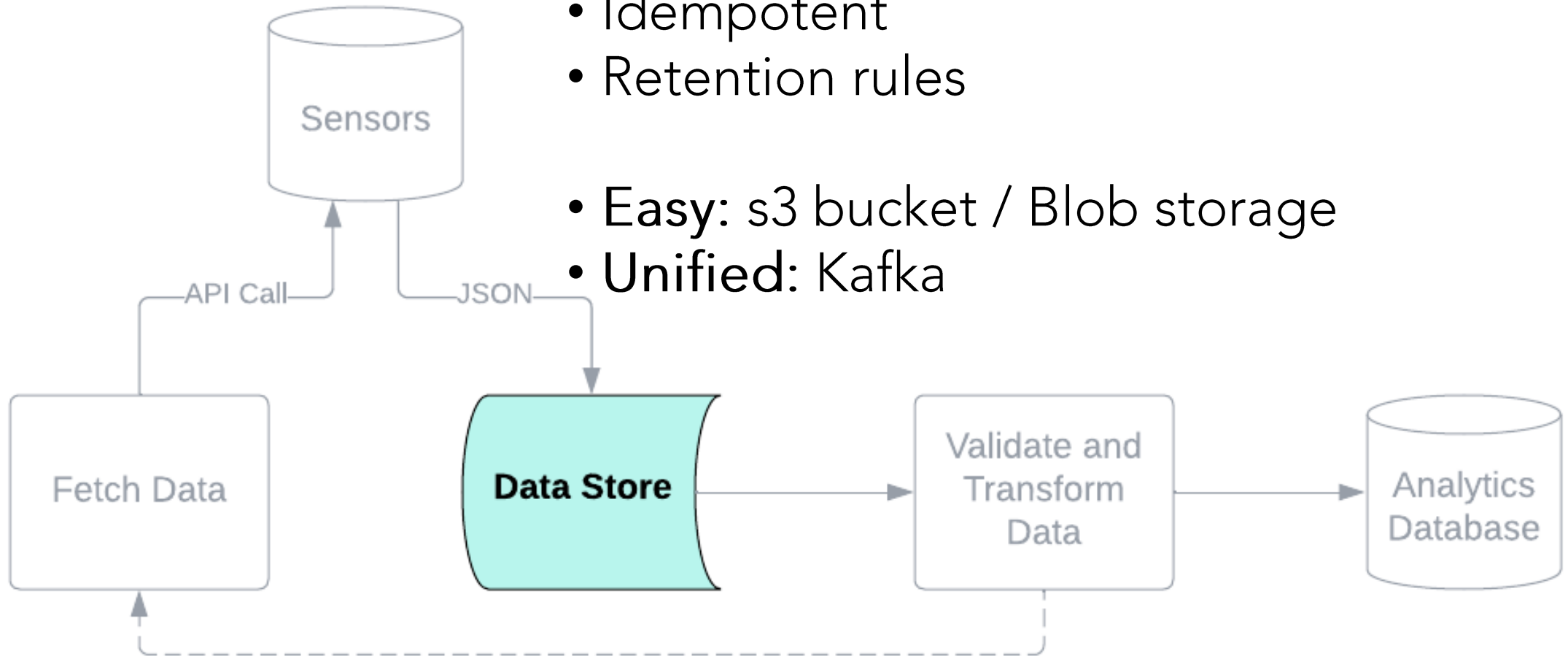
Fetching the Data

- Scheduled vs. streaming

- **Easy**: AWS Lambda / Azure Functions
- **Unified**: App in Docker

Storing the Raw Data

- Idempotent
- Retention rules

- **Easy:** s3 bucket / Blob storage
- **Unified:** Kafka

Sensors

API Call

JSON

Fetch Data

Data Store

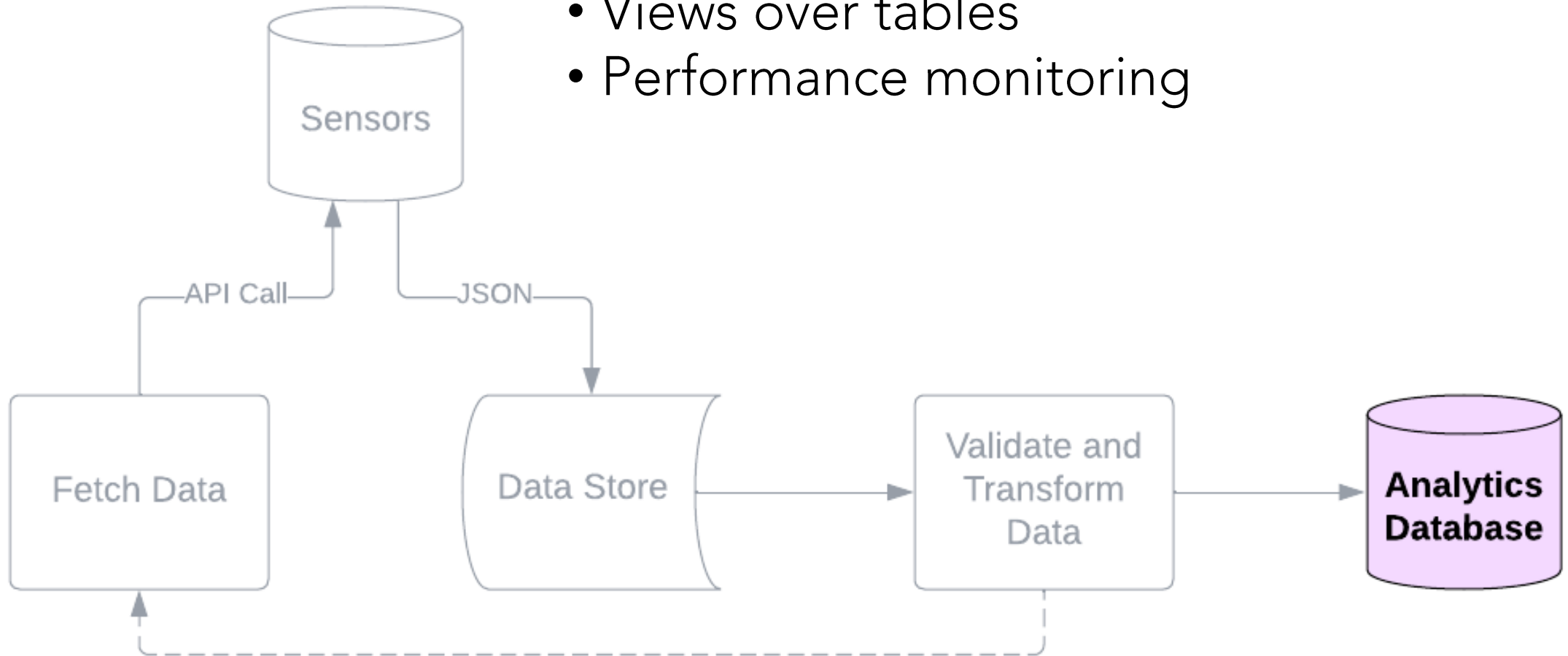Validate and Transform Data

Analytics Database

Validating and Transforming

- Invalid data – rerun fetch job
- Missing data – availability vs. consistency

- **Recommended:** Snowflake, dbt
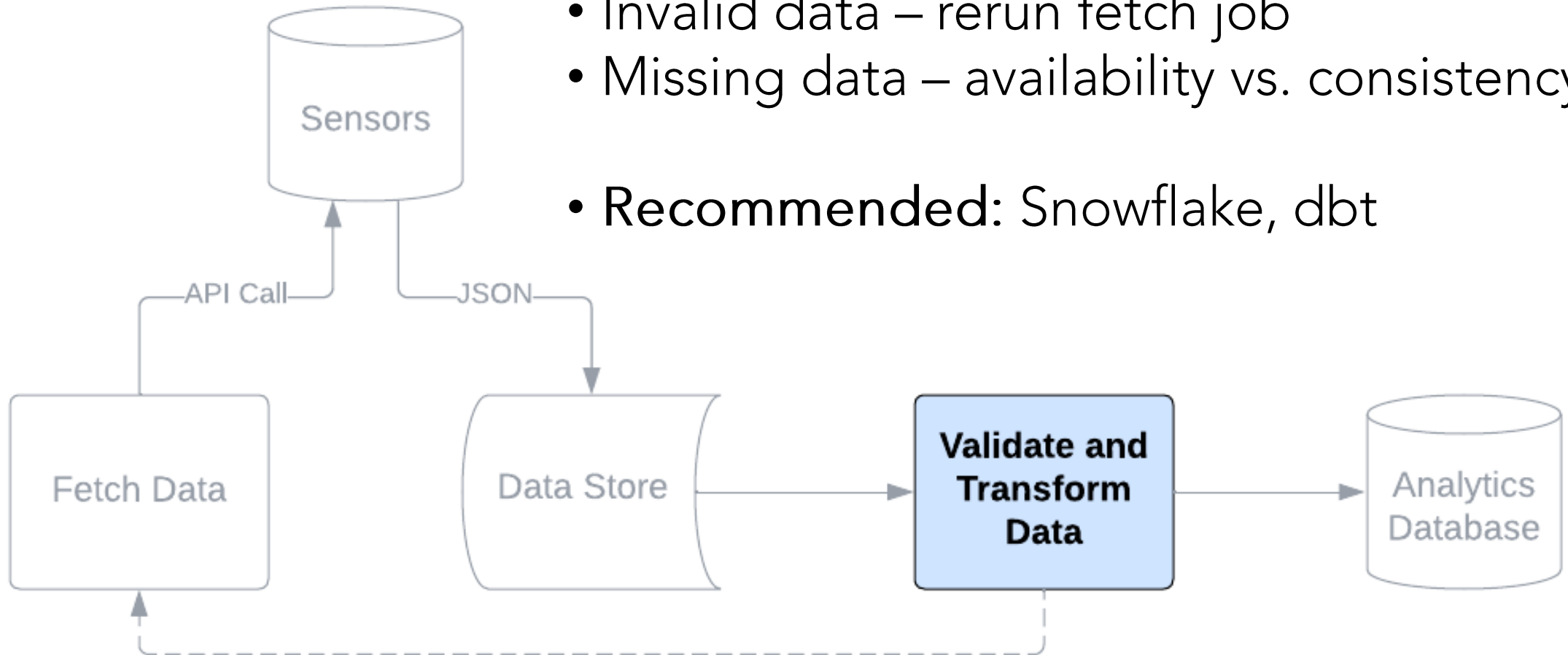
Ready for Analysis

- Views over tables
- Performance monitoring

Questions and Discussion