

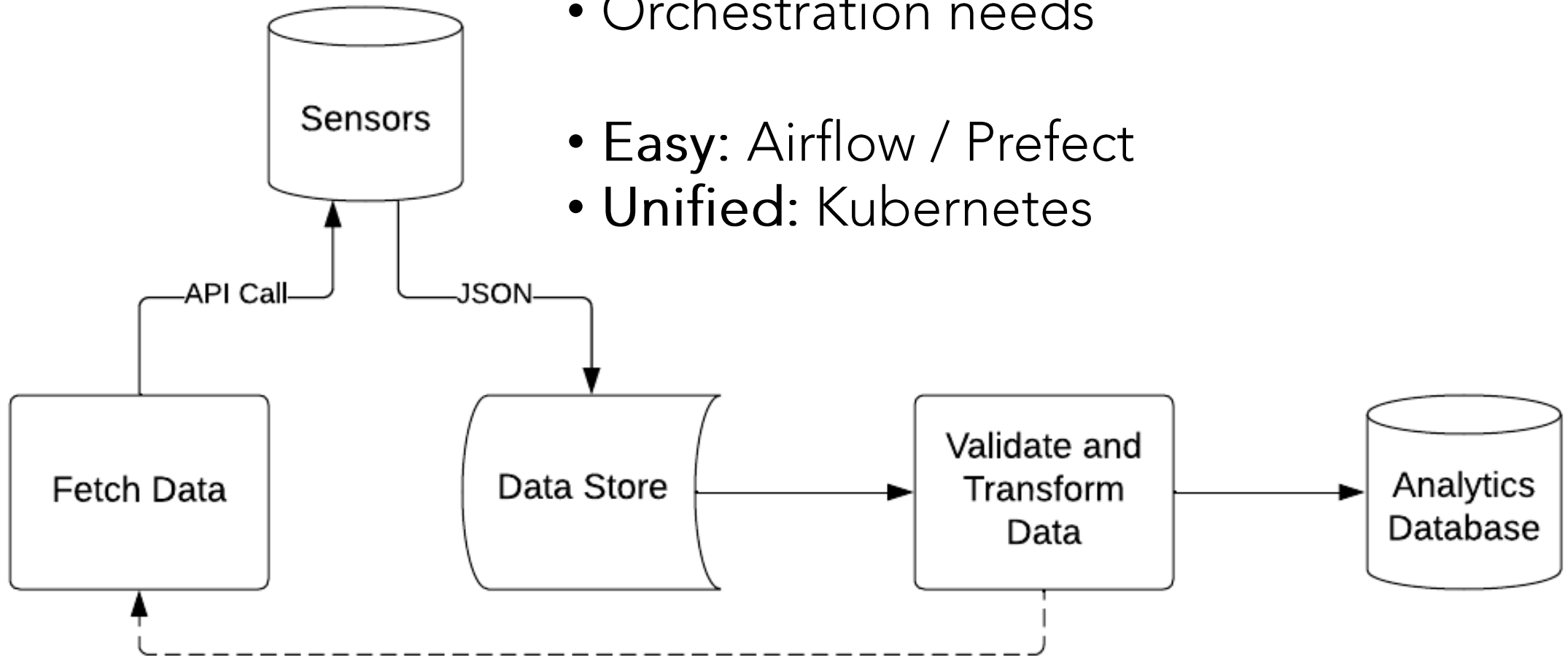


# Productionizing Trash Can Analytics

A proposal by Herschel Pecker

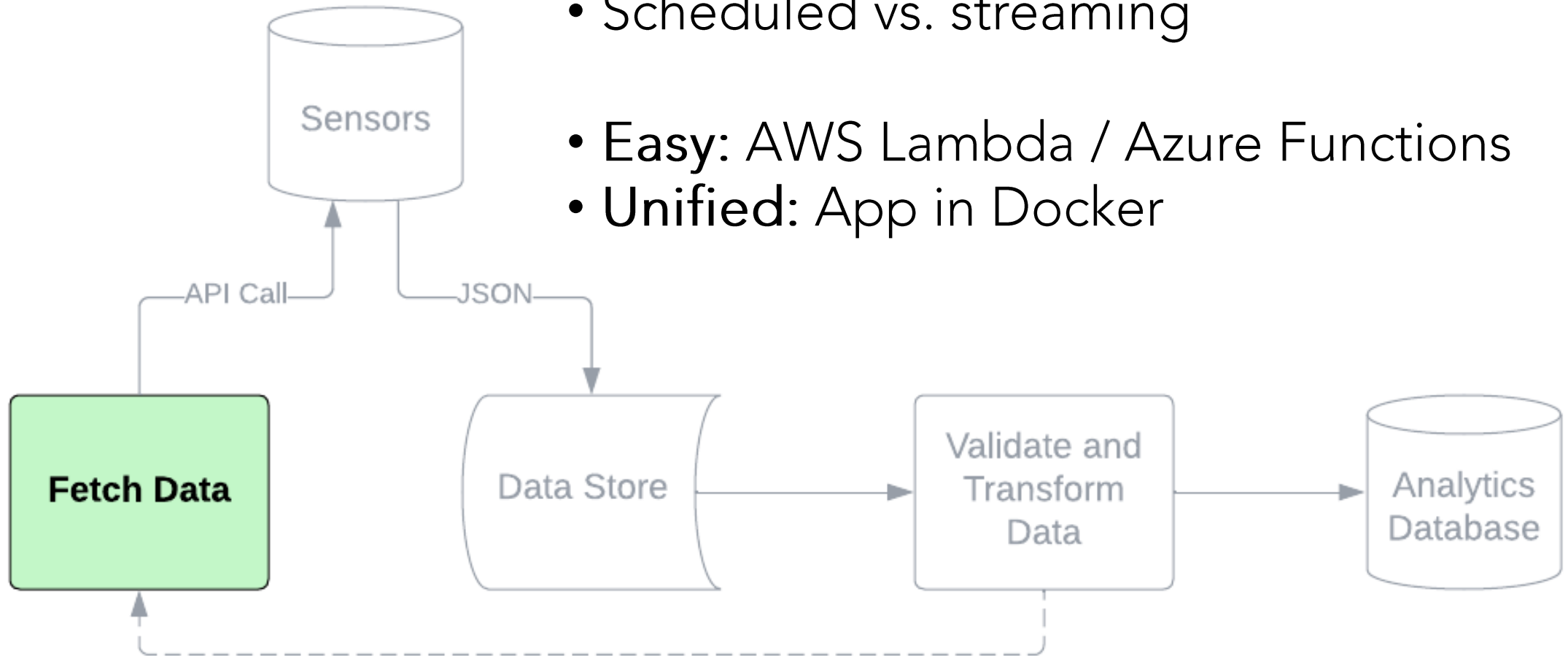
# The ETL Process

- Orchestration needs
- Easy: Airflow / Prefect
- Unified: Kubernetes



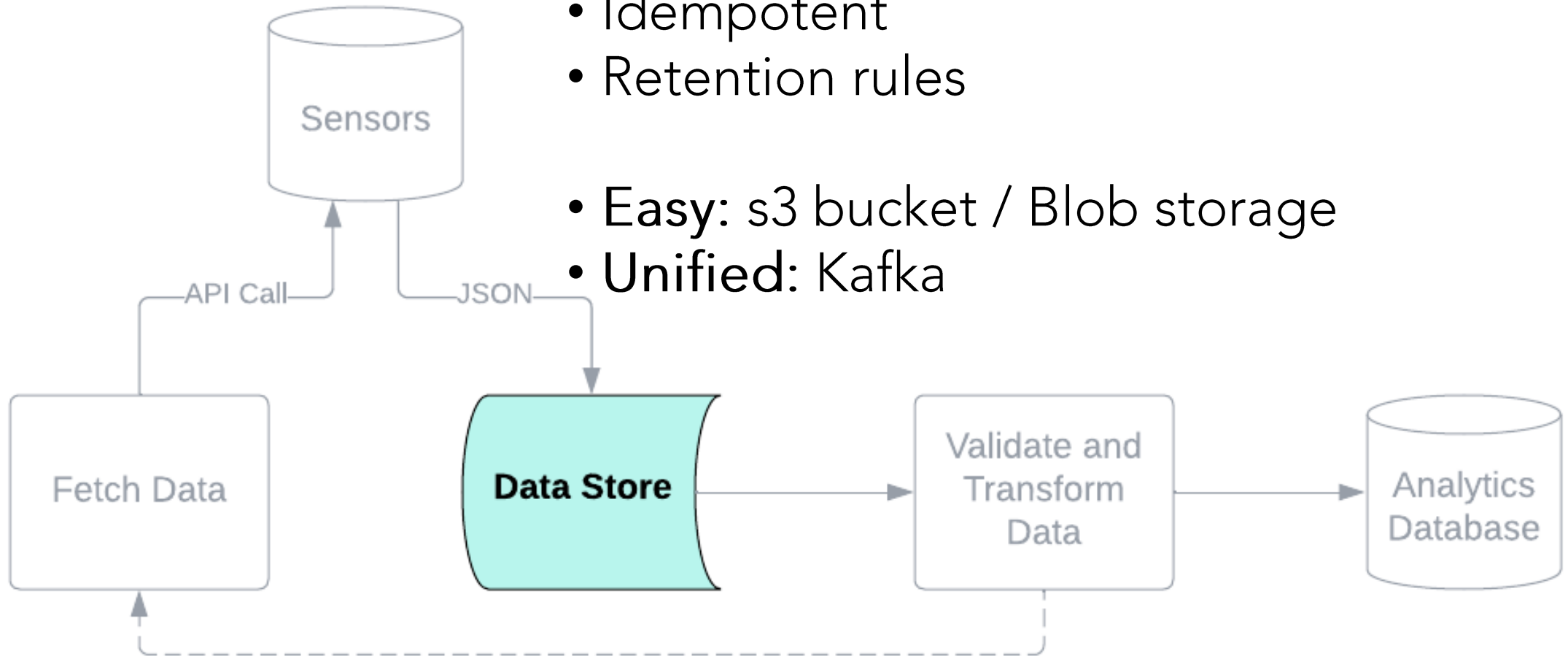
# Fetching the Data

- Scheduled vs. streaming
- Easy: AWS Lambda / Azure Functions
- Unified: App in Docker



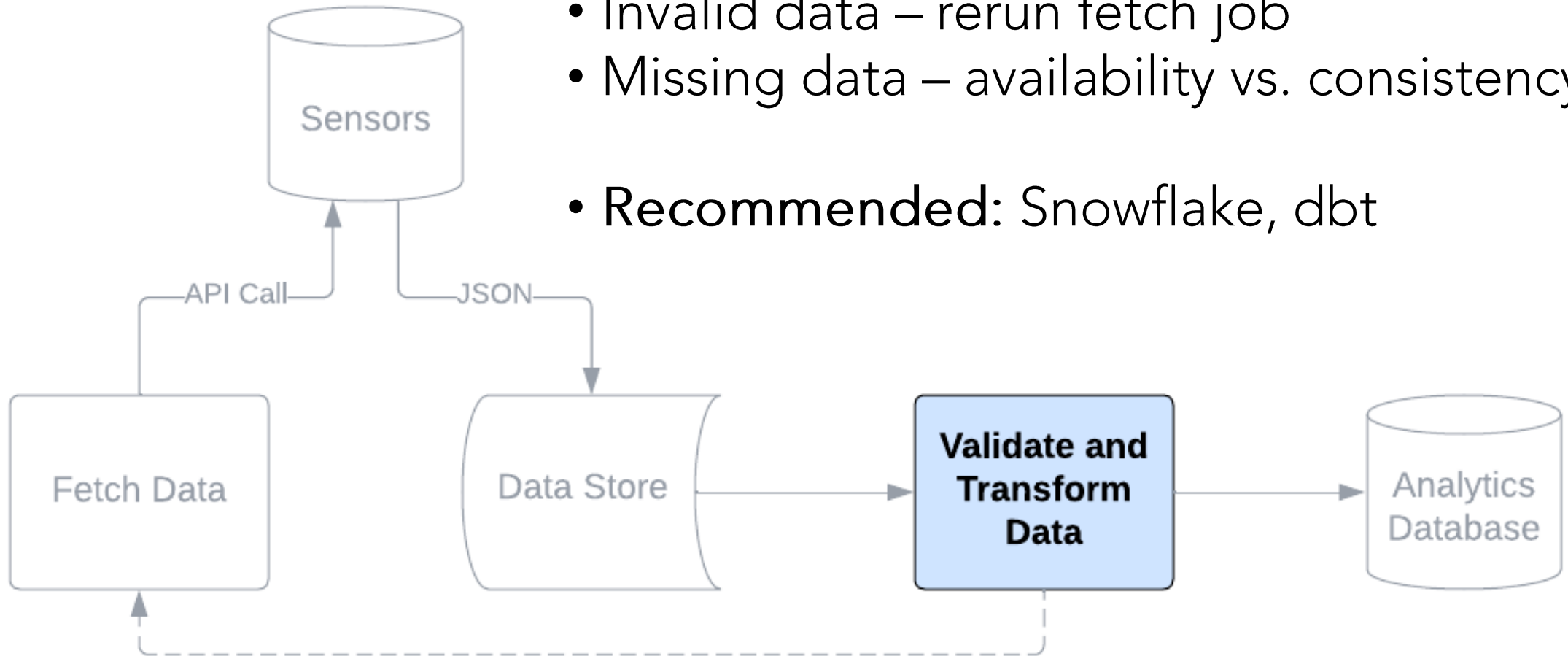
# Storing the Raw Data

- Idempotent
- Retention rules
- Easy: s3 bucket / Blob storage
- Unified: Kafka



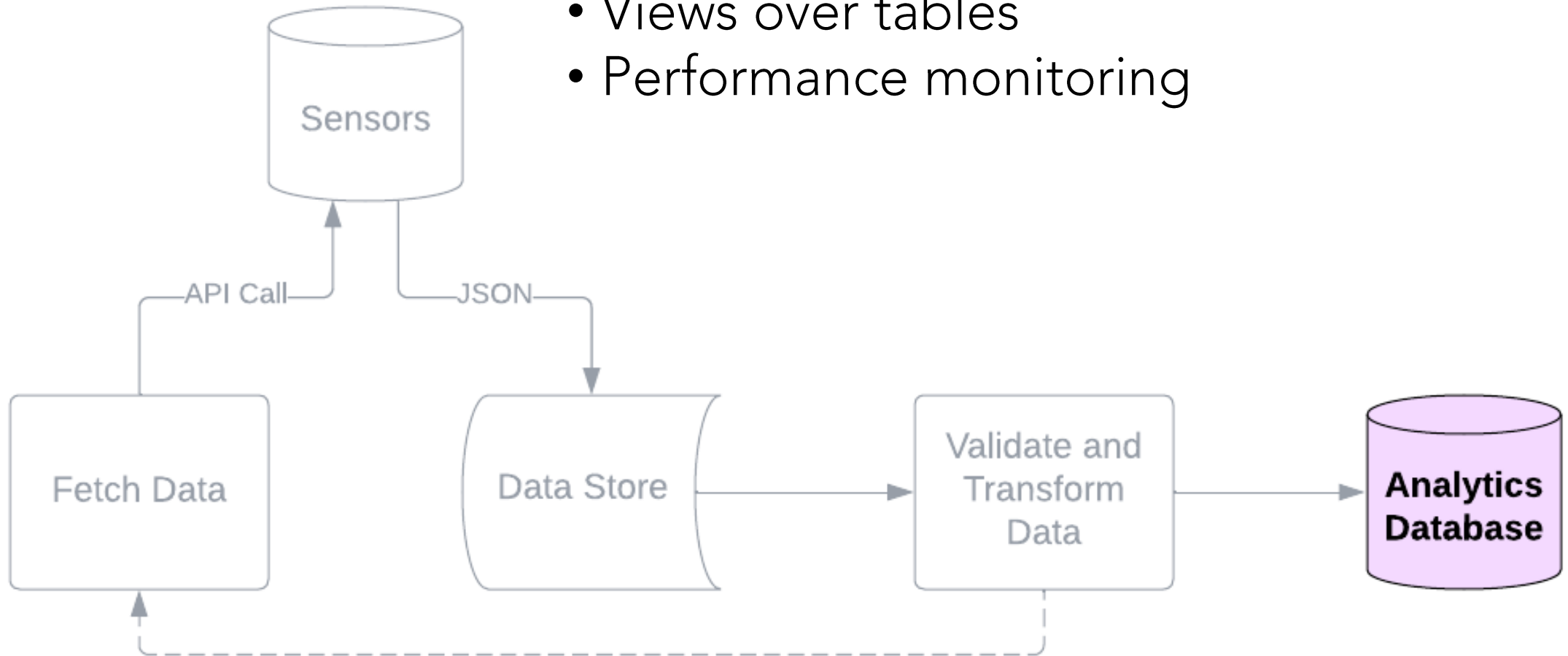
# Validating and Transforming

- Invalid data – rerun fetch job
- Missing data – availability vs. consistency
- Recommended: Snowflake, dbt



# Ready for Analysis

- Views over tables
- Performance monitoring



# Questions and Discussion