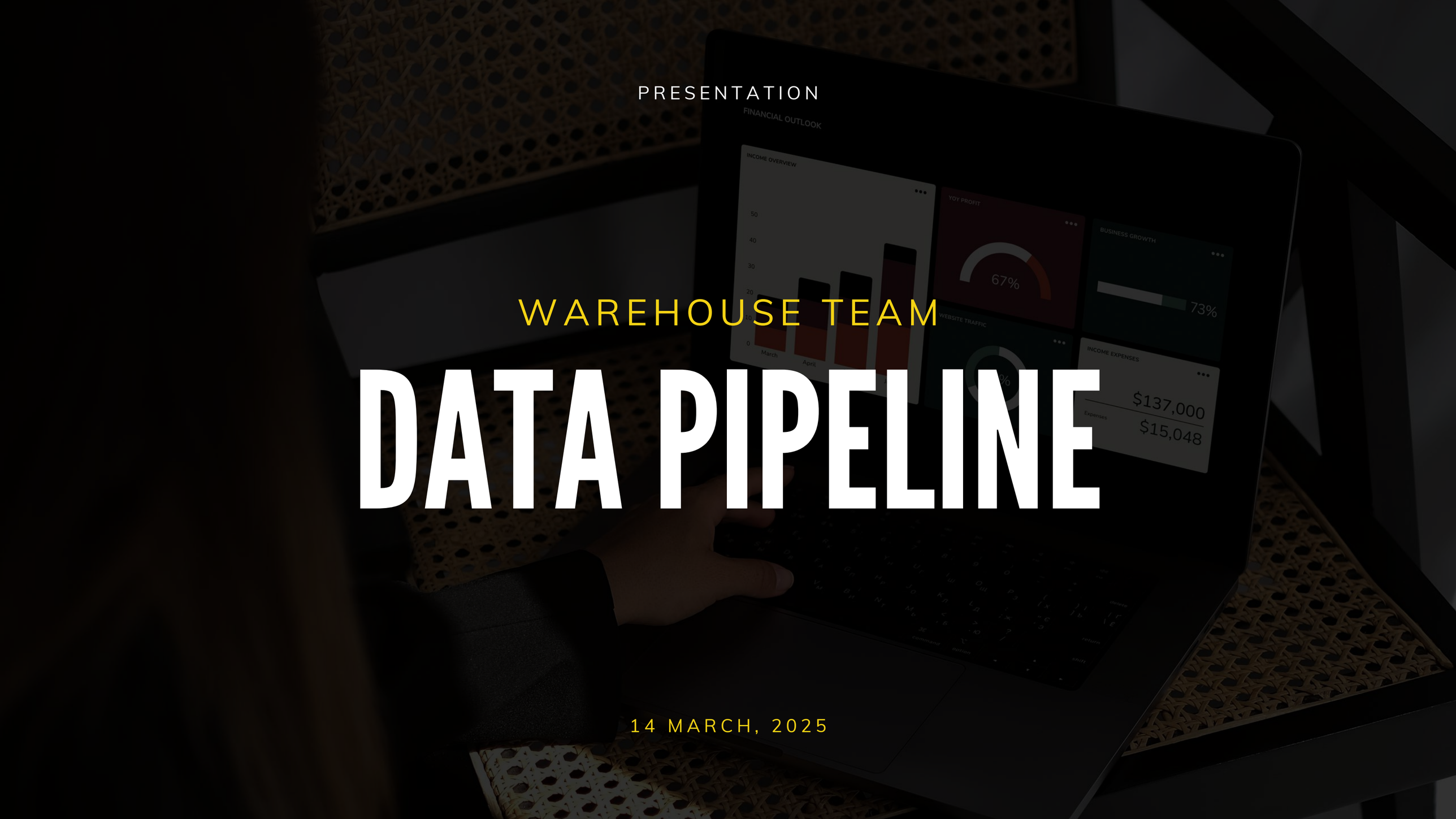
PRESENTATION

WAREHOUSE TEAM

DATA PIPELINE

14 MARCH, 2025



ADVANCE OF THIS WEEK

WAREHOUSE TEAM

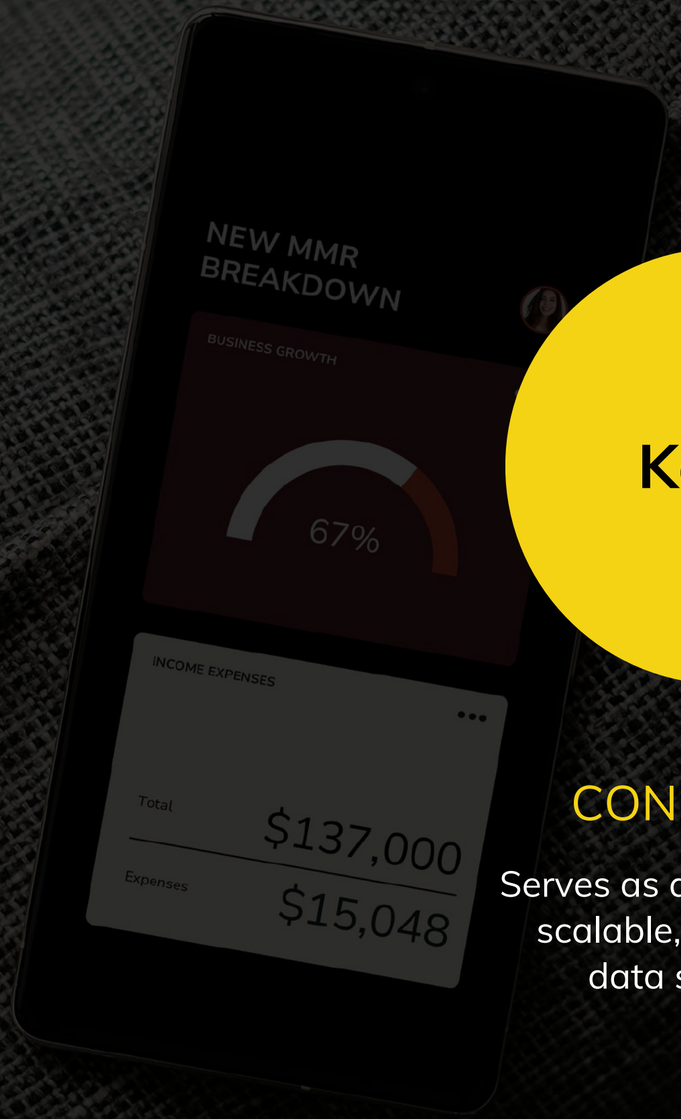
COMPLETE THE DATA PIPELINE

- Complete Data pipeline
- Found Monitoring and scheduling softwares
- Visualize Dataflow
- Compare visualization software
- Data lake or not?

INGESTION LAYER

WAREHOUSE TEAM

Iot, log, databases, ...



Kafka

DATA SOURCES

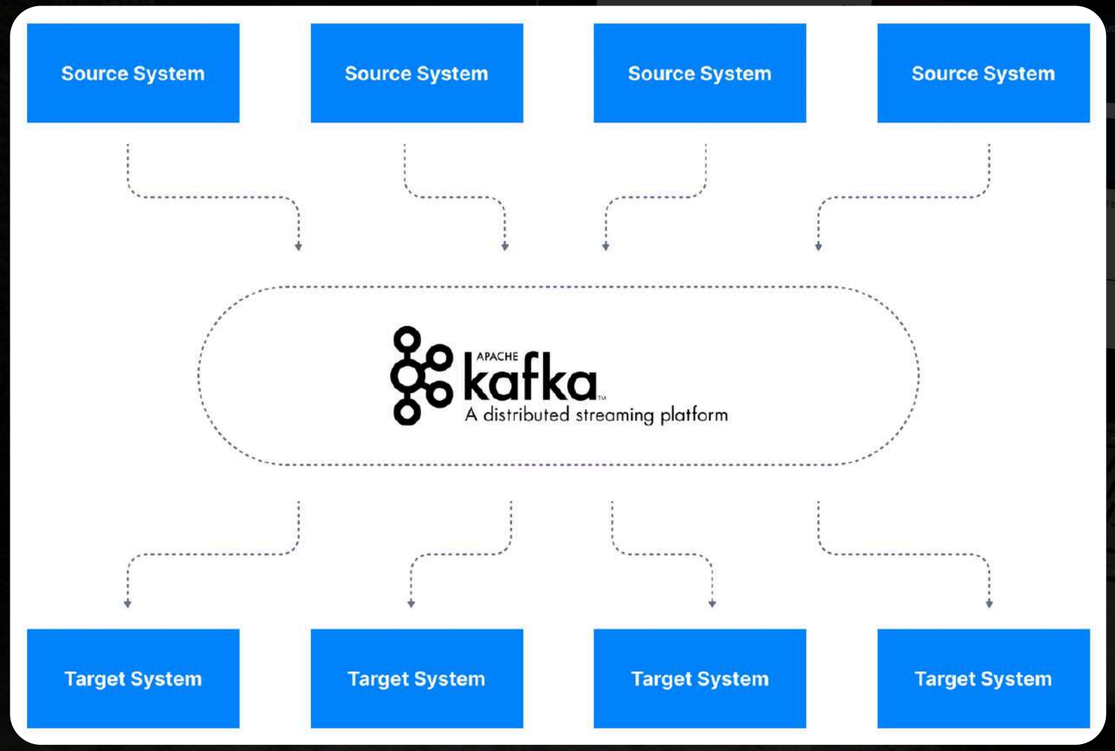
Define the sources for the IDL

CONNECTOR

Serves as a buffer. Ensures scalable, fault-tolerant data streaming.

KAFKA

WAREHOUSE TEAM



80%

FORTUNE 100

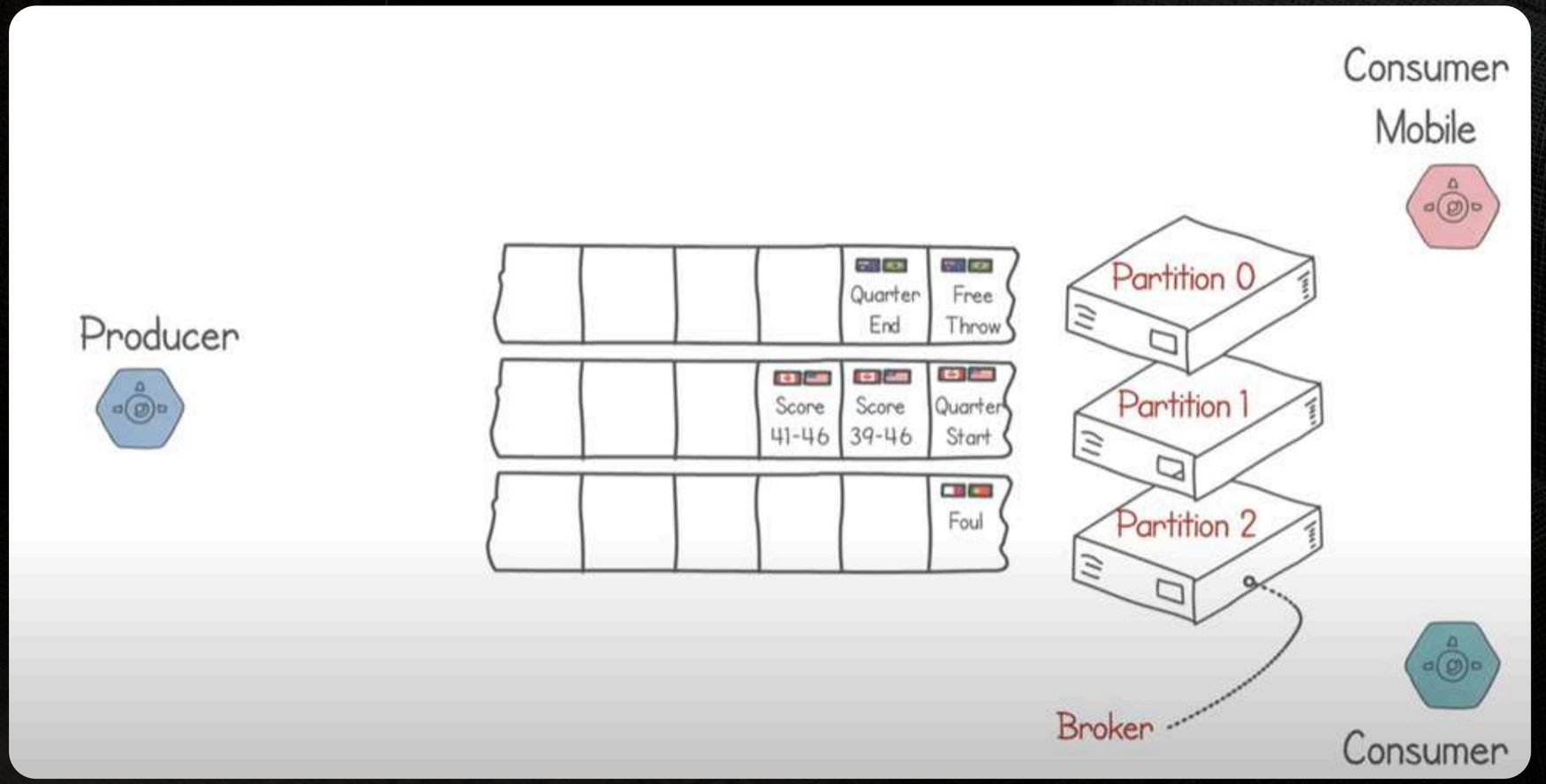
Over 80% of all Fortune 100 companies use Apache Kafka

WHAT IS IT

It's designed to handle high volumes of data and allows you to publish, store, and process streams of records. Essentially, it's a way to move and manage large amounts of data very quickly.

KAFKA

WAREHOUSE TEAM



PROCESSING AND STORAGE LAYER

WAREHOUSE TEAM

POSTGRESQL

Relational database for structured data like logs and metadata. Strong ACID compliance ensures data consistency.Robust, open-source, and widely adopted.

Disadvantages: Less optimized for real-time analytics compared to dedicated OLAP tools.

PROCESSING AND STORAGE LAYER

WAREHOUSE TEAM

INFLUXDB

Optimized for time-series data like IoT sensor readings. Fast data ingestion and querying for time-based metrics. Efficient storage design for high-frequency data points.

Disantages: Limited SQL support; requires specialized query syntax.

PROCESSING AND STORAGE LAYER

WAREHOUSE TEAM

APACHE HUDI

Manages large-scale **data lakes** with support for incremental updates. Ideal for handling streaming data and CDC (Change Data Capture). Enables efficient data versioning and time travel queries.

Disantages: Can be complex to configure and requires well-structured schemas.

PROCESSING AND STORAGE LAYER

WAREHOUSE TEAM

DORIS

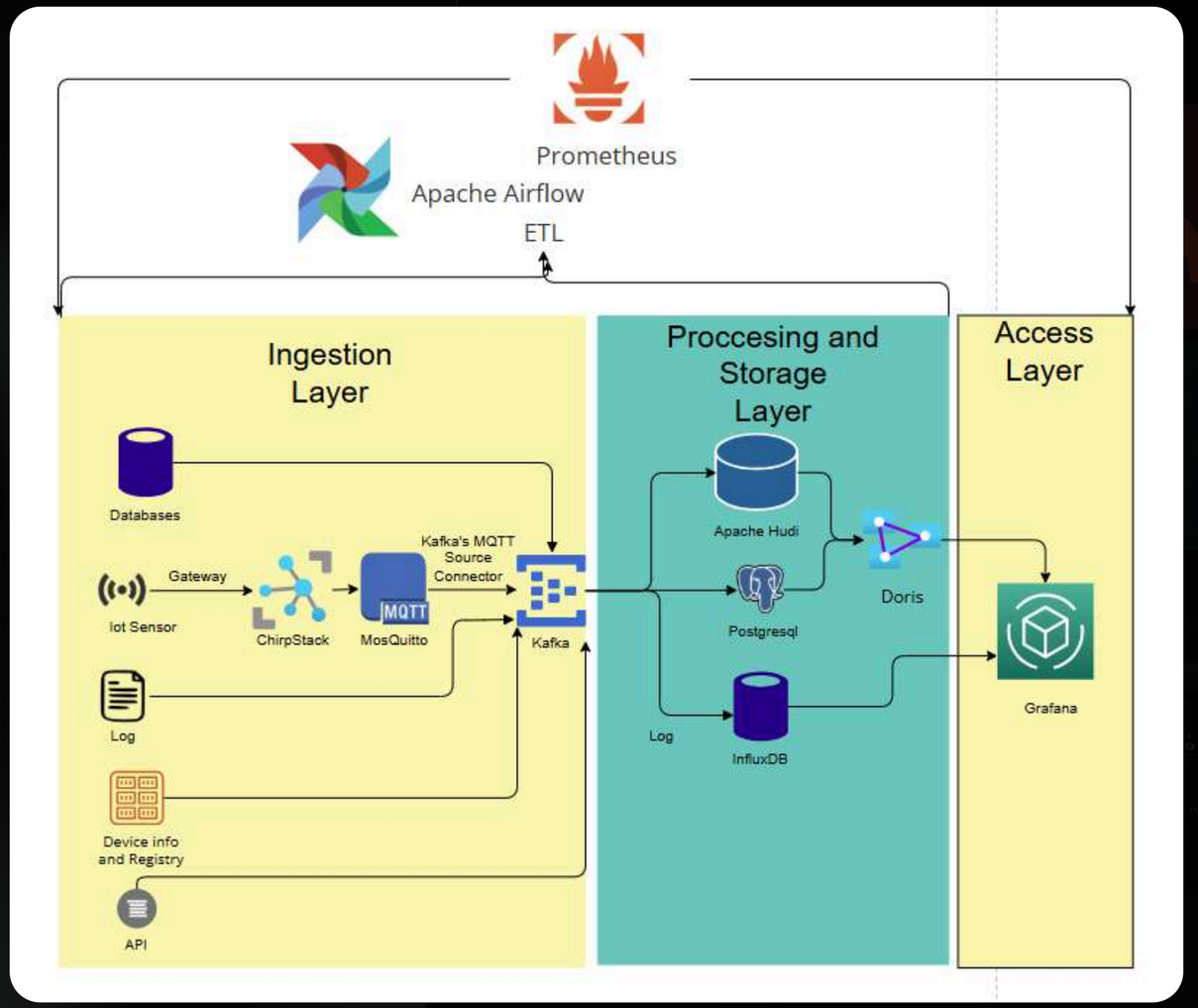
Analytical database for fast, real-time querying. Ideal for OLAP workloads. High performance for large-scale data aggregation.

Disantages: Less mature compared to other OLAP solutions like ClickHouse.

DORIS

PIPELINE

WAREHOUSE TEAM



APACHE AIRFLOW

WAREHOUSE TEAM

?

WHAT IT IS

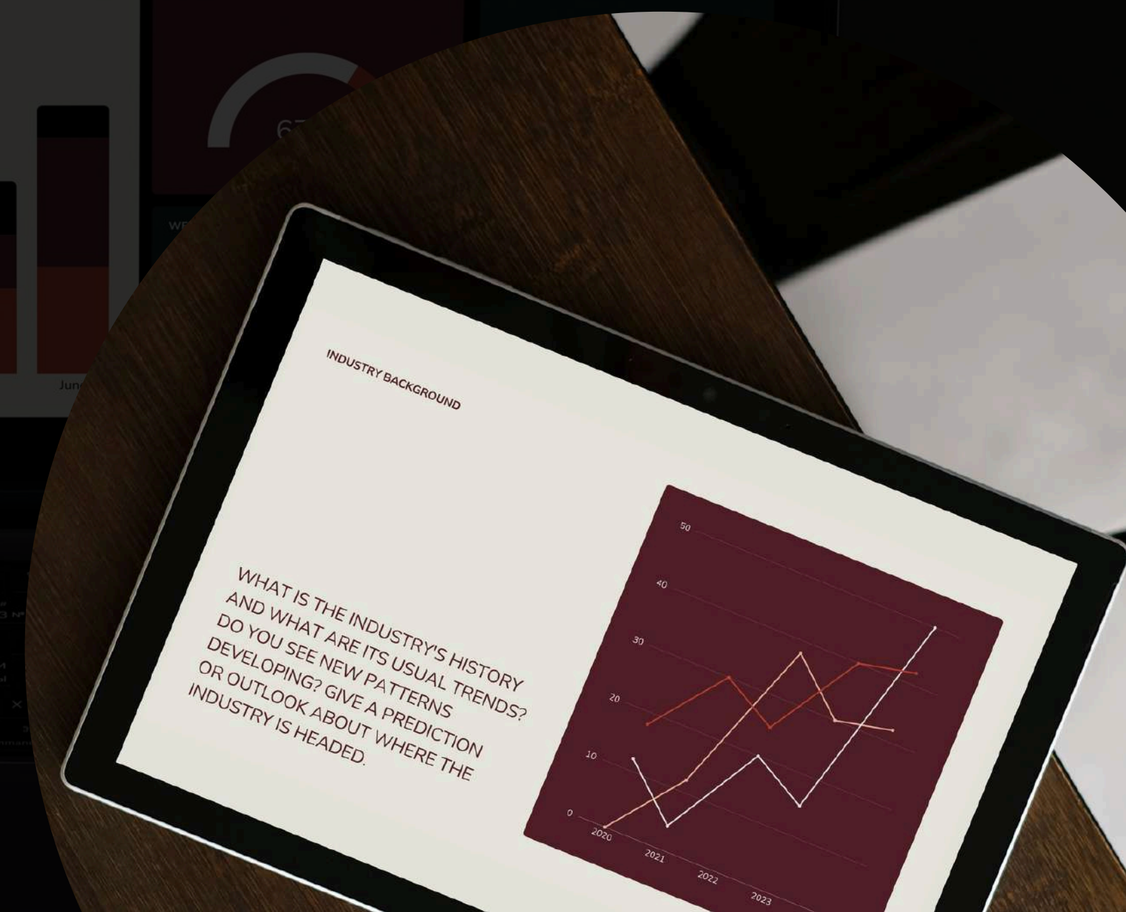
Is essentially a platform for managing and scheduling workflows. It helps you define, schedule, and monitor workflows, particularly data pipelines

DAGs

Directed Acyclic Graphs

WHY SHOULD WE USE IT?

Workflows in Airflow are represented as DAGs. These are essentially visual representations of your tasks and their dependencies.
This makes it easy to see how your tasks relate to each other.



PROMETHEUS

WAREHOUSE TEAM

?

WHAT IT IS

is a powerful open-source monitoring and alerting toolkit widely used in cloud-native environments

Why?

WHY SHOULD WE USE IT?

Prometheus ensures your pipeline is running smoothly and alerts you when it's not.

CONCLUSIONS

WAREHOUSE TEAM

NEXT YEAR PLANS

A presentation is a formal or informal communication method that involves conveying information, ideas, or a message to an audience. It often employs visual aids such as slides, charts, graphs, or multimedia elements to support and enhance the spoken content.

IMPROVEMENTS

A presentation is a formal or informal communication method that involves conveying information, ideas, or a message to an audience. It often employs visual aids such as slides, charts, graphs, or multimedia elements to support and enhance the spoken content.

WE WANT TO SAY

THANK YOU

FOR YOUR ATTENTION