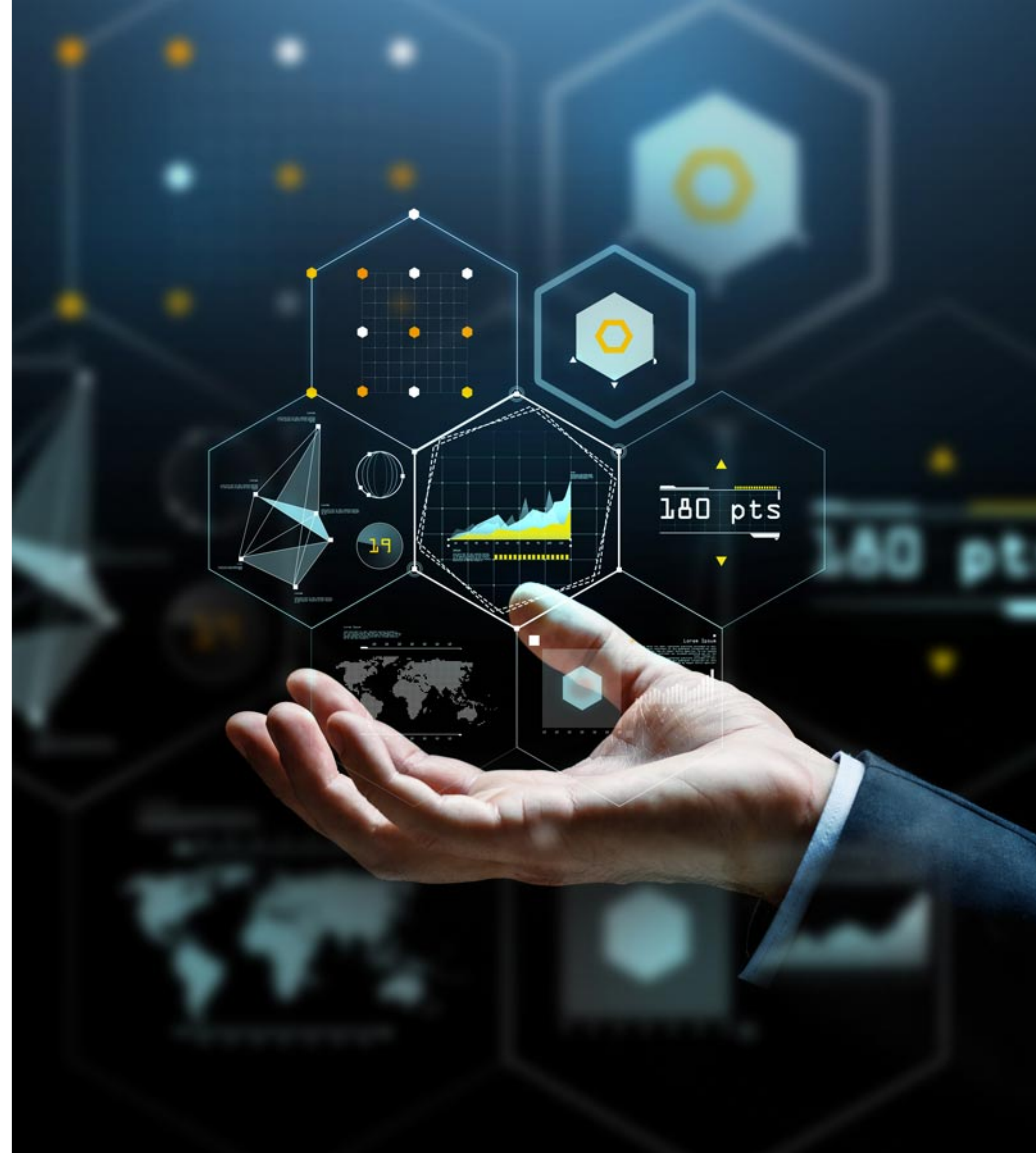# ETL

EXTRACT - TRANSFORM - LOAD

Soufiane ODF

Business intelligence and analytics today

Business intelligence is vital for enterprise since it allows:

- Gain new customer insights

- To improve visibility

- To turn data into actionable information

- To improve efficiency

- To gain sales & market intelligence

- To gain competitive intelligence

Questions the business tries to answer

- How did our marketing campaign affect the company logistic cost ?

- How did the growth of our company employees number reflect on the sales profit recorded this year ?

- How does the increase of return rate reflect on our supplier order ?

Enterprise data sources
The enterprise data comes from different resources

- ERP

- Website

- Local transaction database

- Partner/client system

- Cloud platforms

- Other systems

How does the data look like ?
The data comes in different format

- Database tables

- Excel files

- API response

- CSV files

- JSON files

- XML files

- Space/Tab delimited files

- AVRO/PROTOBUF files

- Etc...

Limitations of having different data formats

- Different ways to query and access the data sources.

- Complexity of managing data quality.

- Business questions usually limited to one data source.

Birth of Data Warehouses

- A data warehouse (DW or DWH), also known as an enterprise data warehouse (EDW), is a system used for reporting and data analysis and is considered a core component of business intelligence. DWs are central repositories of integrated data from one or more disparate sources. They store current and historical data in one single place that are used for creating analytical reports for workers throughout the enterprise.

What is ETL ?

- **Extract**: source-specific routes to pull selected data from an external system.

- **Transform**: business logic specific to your organization to serve an analytics or operational use case.

- **Load**: destination specific routines to push data where it is going to be consumed.

ETL doesn't work in today's world

Inflexible:

- friction when changing an existing pipeline.

- Hard to add new data

- Most issues force data to be re-extracted

Lack of autonomy:

- Warehouse made data consumers more autonomous

- Changes require engineering involvement

Complex:

- Custom DSL

- Force adoption of a data stack

- Address 70% of the needs, 30% still built and maintained in-house

- **Extract**: general-purpose routines to pull selected data from a source.

- **Load**: general-purpose routines to push raw data where it is going to be consumed.

- **Transform**: business logic specific to your organization to serve an analytics or operational use case with SQL/dbt/...

ELT fixes the ETL related issues

Flexibility:

- All the data available on the destination

- Data consumers are free to use what they need for the insights they want.

Autonomy:

- Data consumers can leverage SQL queries to transform the data the way they want.

- No need to involve the engineering team.

Future proof:

- Issues during transformation don't prevent access to the data

- Easy to update transformation schemas.

Airbyte as an example

Open-source data integration for modern data teams

Get all your ELT data pipelines running in minutes, even your custom ones. Let your team focus on insights and innovation.

Get Started for Free

→ Or deploy Airbyte Open-source

Extract from sources
Pre-built or custom connectors

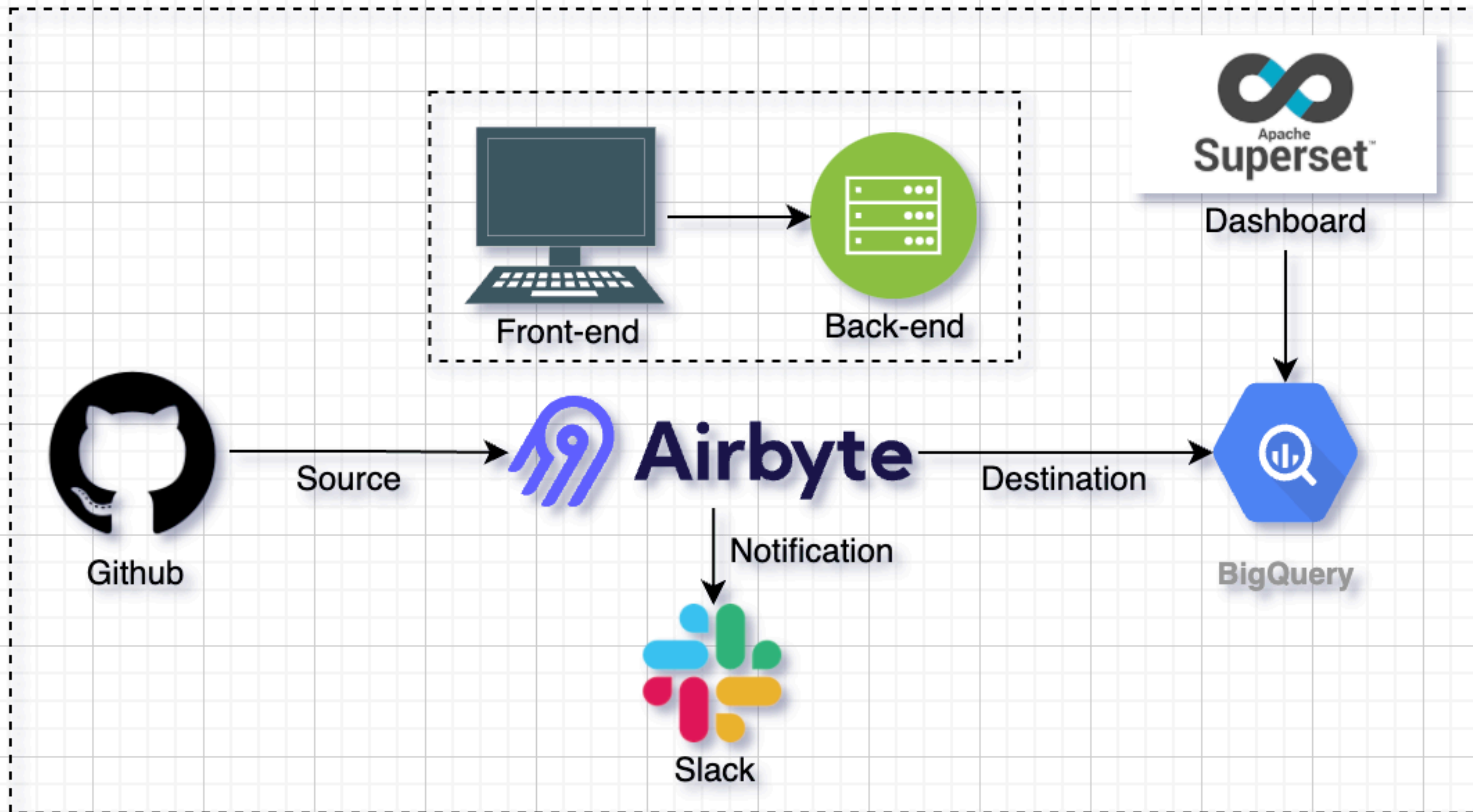### Load to destinations
Through our UI or API

Transform
Raw or normalized, with dbt-based transformations

Architecture Example

BI/Visualization

Superset

Looker

Data
Warehouse

Reverse ETL

hightouch

Census

Airbyte

salesforce

stripe

Demo

Source code and Keynote are available on **Github**

@soufianeodf/virtuocode-w1-etl

Front-end → Back-end

Apache Superset
Dashboard

Github → Source → Airbyte → Destination → BigQuery

Airbyte → Notification → Slack

Digital ocean + docker

Thanks!