UNLOCKING THE TREASURE CHEST OF DATA

A QUEST WITH CLICKHOUSE© & DBT™

EMIR KARAMEHMETOGLU

DATA ENGINEER & DUNGEON MASTER

AXIS COMMUNICATIONS

DIAGNOSTIC DATA MANAGEMENT

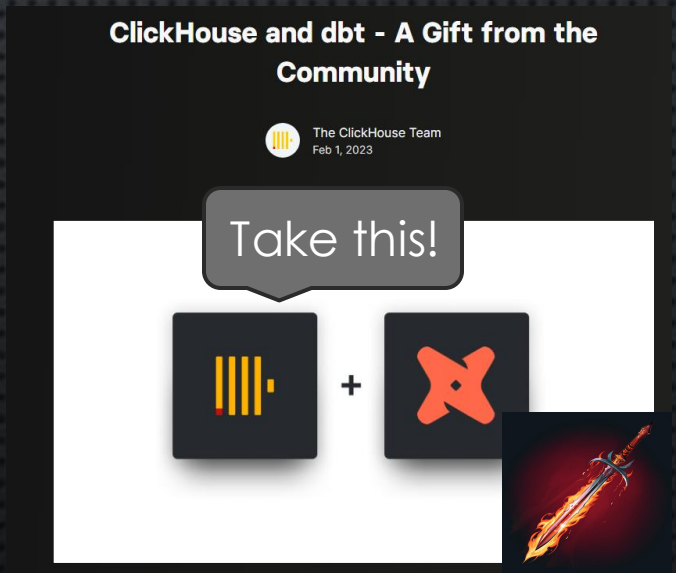
CHARACTER BACKGROUND: ASTRONOMER

Translation: **Self-service analytics data-products** with **ClickHouse** and Dagster orchestrated **dbt**

Our Adventure Map

: Agenda

- **Treasure Chest:** Self-Service Analytics Data Products
- **Arcane Lore:** Data Product? Platform?
    - **The Dragon's lair:** The challenges
- **Sword and Shield:** Our Solution
    - **Battle Plans:** Technical Architecture
- **Slaying the dragon:** Problems Overcome
    - **A dead end?:** clickhouse-dbt plugin and distributed database architecture
    - **Treasures and Traps:** Lessons Learned
- **New Horizons:** Future Roadmap

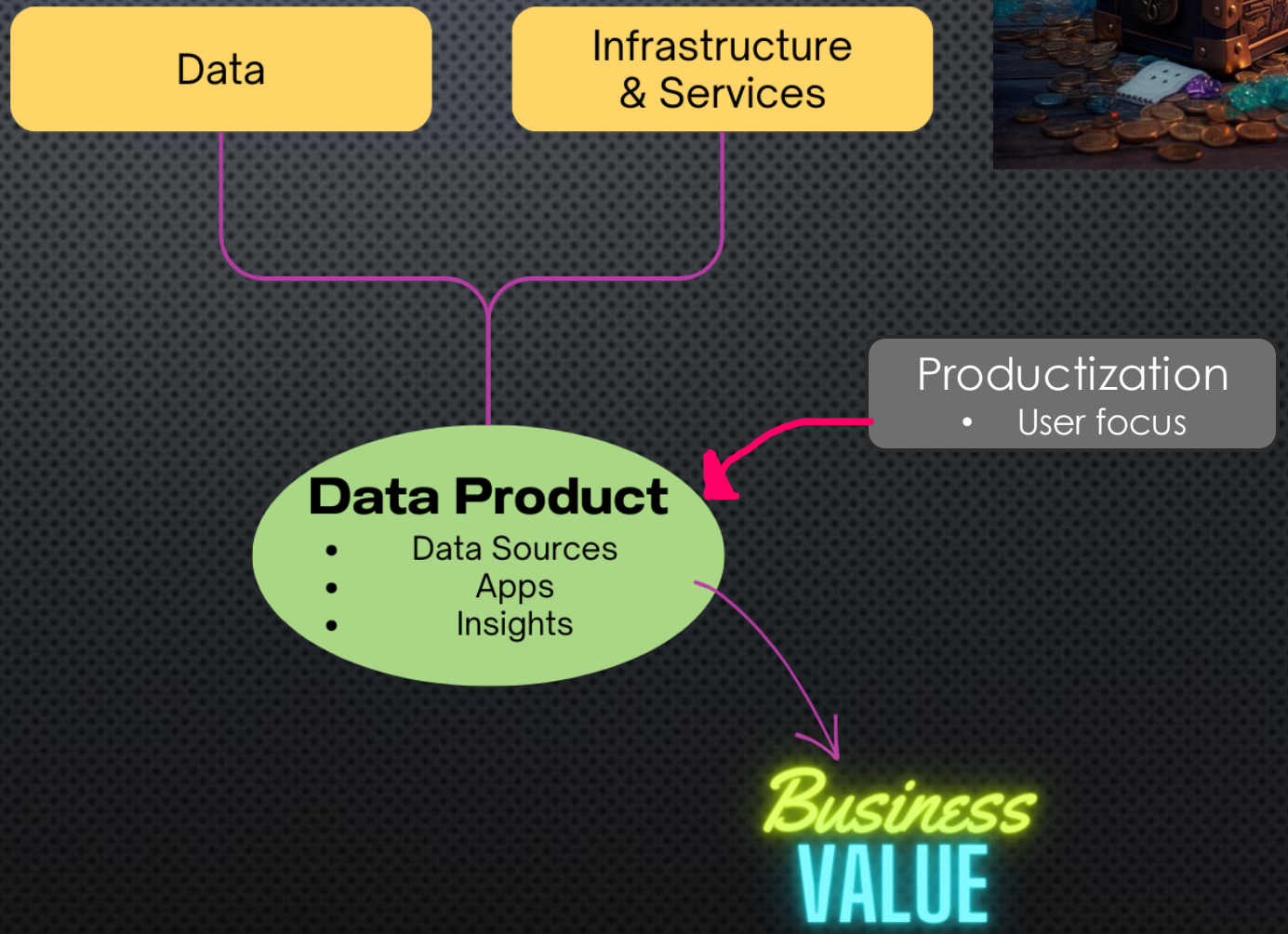


THE TREASURE CHEST

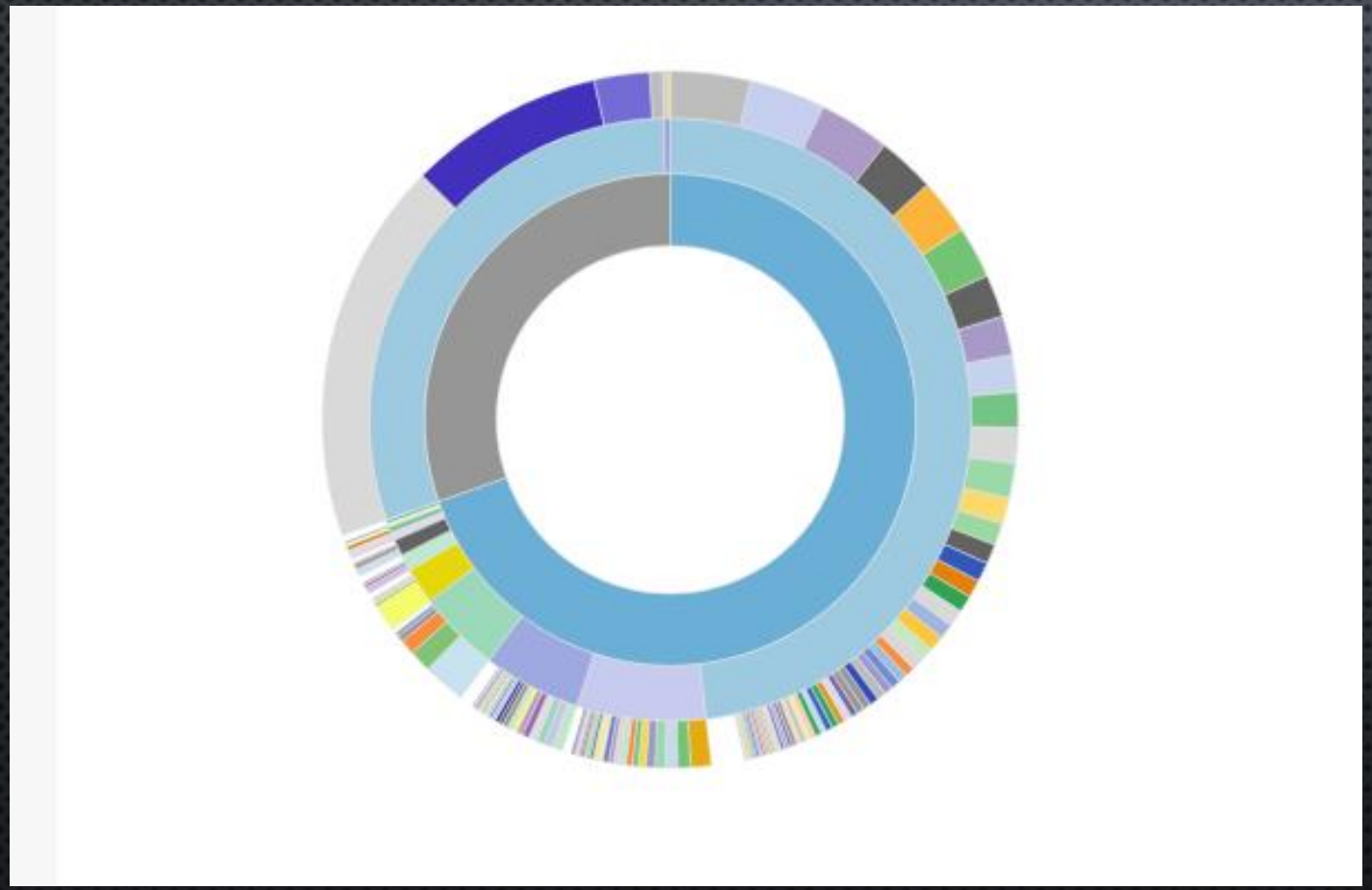
- UNLOCK VALUE OF DATA
  - HALT SOUL-CRUSHING DATA DIGGING EXPEDITIONS
- DEMOCRATIZE (AND MODERNIZE) DATA DRIVEN DECISION MAKING
- ELEVATE DATA MATURITY

ARCANE LORE: DATA PRODUCTS

- BI Dashboard
- Data API
- ML Model
- Data Visualization
- Analytics App
- "Queryable" Dataset

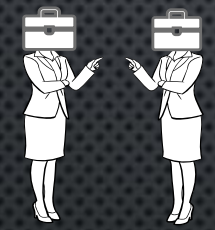


BI LAYER VISUALIZATION



WHAT IS IN A GOOD DATA PRODUCT?

DATA PRODUCT IS..	MEANING..
• DISCOVERABLE	• SEARCH, FIND, GET DOCS.
• TRUSTWORTHY	• DATA QUALITY. **OBSERVABILITY**. VALIDATION. METADATA.
• SELF-DESCRIBING	• DESCRIPTION. SCHEMA. NO SHERPAS NEEDED.
• INTEROPERABLE	• STANDARDS (OPEN). UNIFIED NAMING.
• ACCOUNTABLE	• CLEAR OWNERSHIP. RESPONSIBILITY. CONTACT.
• CONSUMED	• COMMUNICATION WITH STAKEHOLDERS. USER-FOCUS.
• ADDRESSABLE & SECURE	• ACCESS & USE, SECURELY.



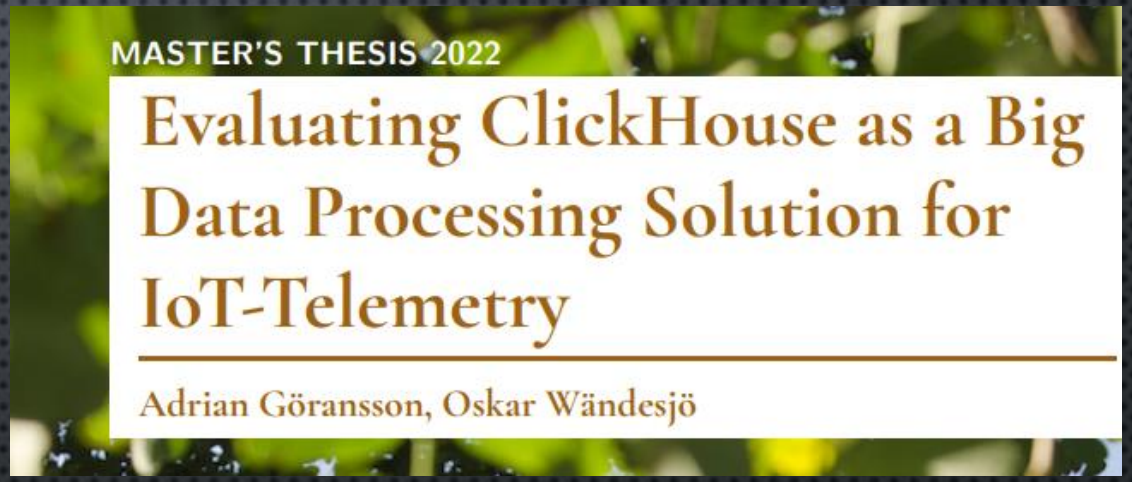
VALUEABLE

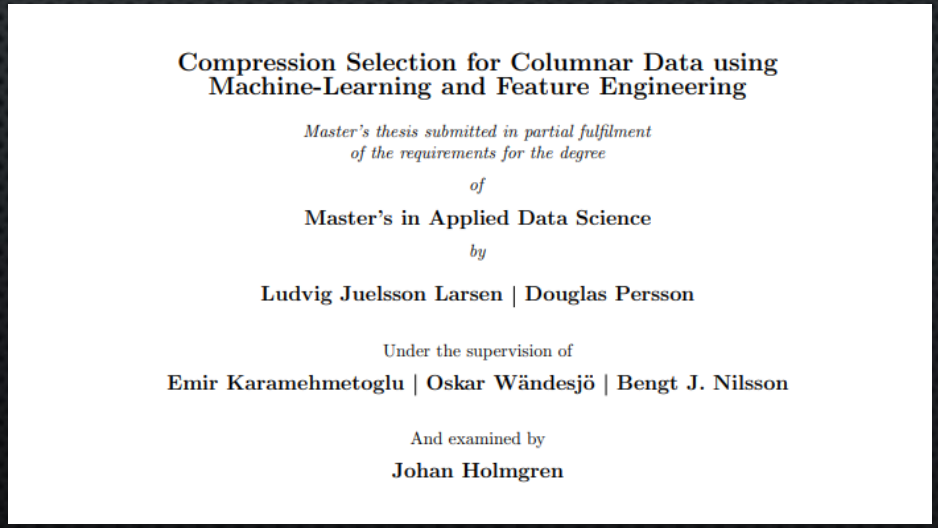
DBM CHOSE CLICKHOUSE FOR ITS ANALYTICS DATA PRODUCTS

our team

ClickHouse

- Open Source
- Active Community & Development
- Open standards (SQL)
- Laughs at scale & speed





Compression Selection for Columnar Data using Machine-Learning and Feature Engineering

Master's thesis submitted in partial fulfilment of the requirements for the degree

of

Master's in Applied Data Science

by

Ludvig Juelsson Larsen | Douglas Persson

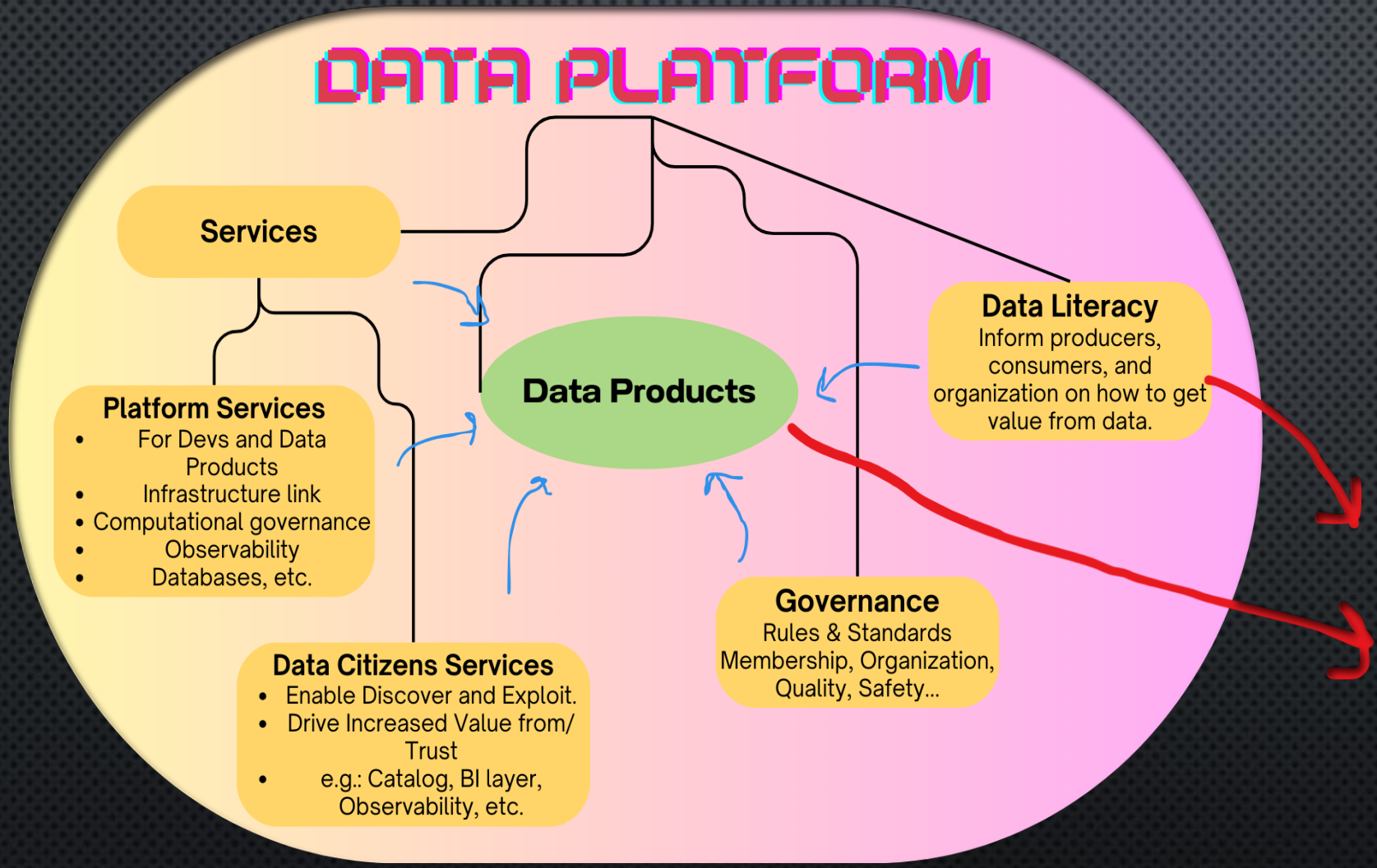
Under the supervision of

Emir Karamehmetoglu | Oskar Wändesjö | Bengt J. Nilsson

And examined by

Johan Holmgren

ARCANE LORE: DATA PLATFORM



THE DRAGON'S LAIR

- Production ClickHouse cluster with hundreds of trillions of entries
  - Sharded & Replicated.
  - Tables with tens of billions of rows.
- Need for data transformation and modeling.
- Data orchestration, observability, and reporting.
- BI-Layer needs raw & transformed data access.
- Requirements:
  - Speed.
  - Data Quality & Governance.
  - Minimize data downtime.

THE DRAGON'S LAIR

- Sharded Replicated*MergeTree Tables
  - Existing data source product have intelligent sharding key.
  - Hard to replicate for users
- Services & Users can access Any node in a load balanced way
- How to smooth over the difficult unsolved problem of distributed database design for self-service?

QUEST REQUIREMENTS

STAKEHOLDERS

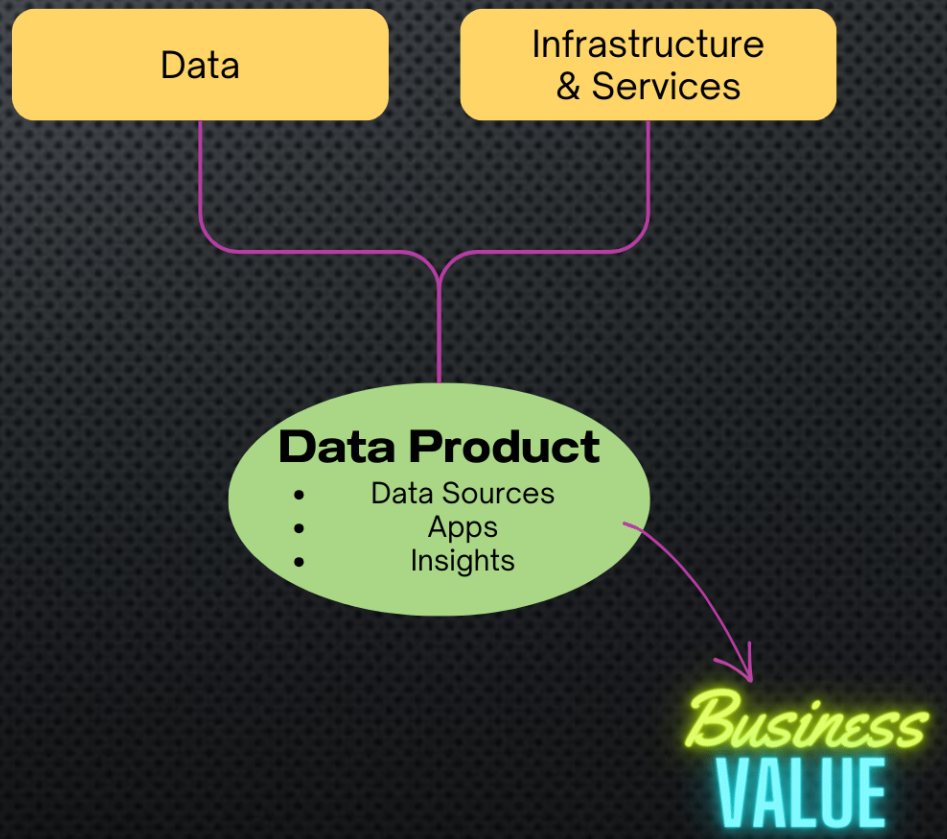
- INTERNAL FACING
- DATA ANALYSTS/SCIENTISTS
- (FUTURE) DATA PLATFORM CITIZENS
- (FUTURE) DATA STEWARDS AND PRODUCT OWNERS.

REQUIREMENTS

- **DATA TRANSFORMATIONS & DATA MODELING**
- DATAOPS WAY OF WORKING WITH ANALYSIS
- X AS CODE. CI/CD.
- **SERVE PRODUCTION MODELS AND DEV DATA EXPLORATION**
- **SELF-SERVICE**
- **DATA QUALITY**
- OBSERVABILITY
- COMPOSABILITY & **INTEROPERABLE**
- **FASTER DASHBOARDS**
- **SELF SCHEDULED**
- **ALERTS** AND MONITORING
- DOCUMENTED AND EASY TO CREATE DOCUMENTATION
- **PYTHON AND SQL.**
- HOOK UP TO DATA SOURCES OUTSIDE THE PLATFORM
- SERVES FIKA AND WHATEVER THE ANALYSTS NEED.

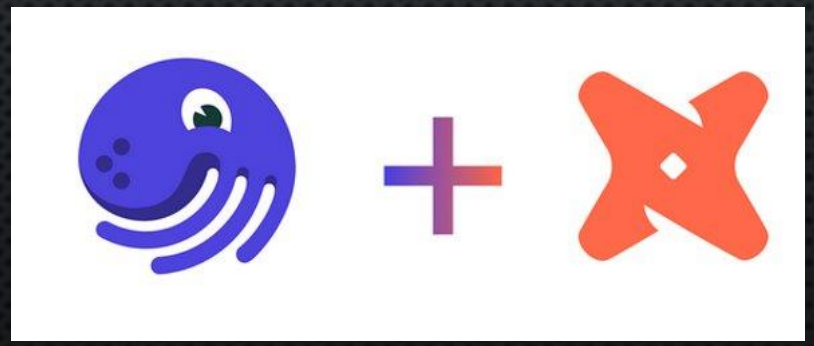
SELF-SERVICE ANALYSIS DATA PRODUCTS – JUST ESSENTIALS

- Democratize Access
- Transformation & Modeling
- Make DBT work with (our) ClickHouse
- SQL & python
- Automate & Schedule

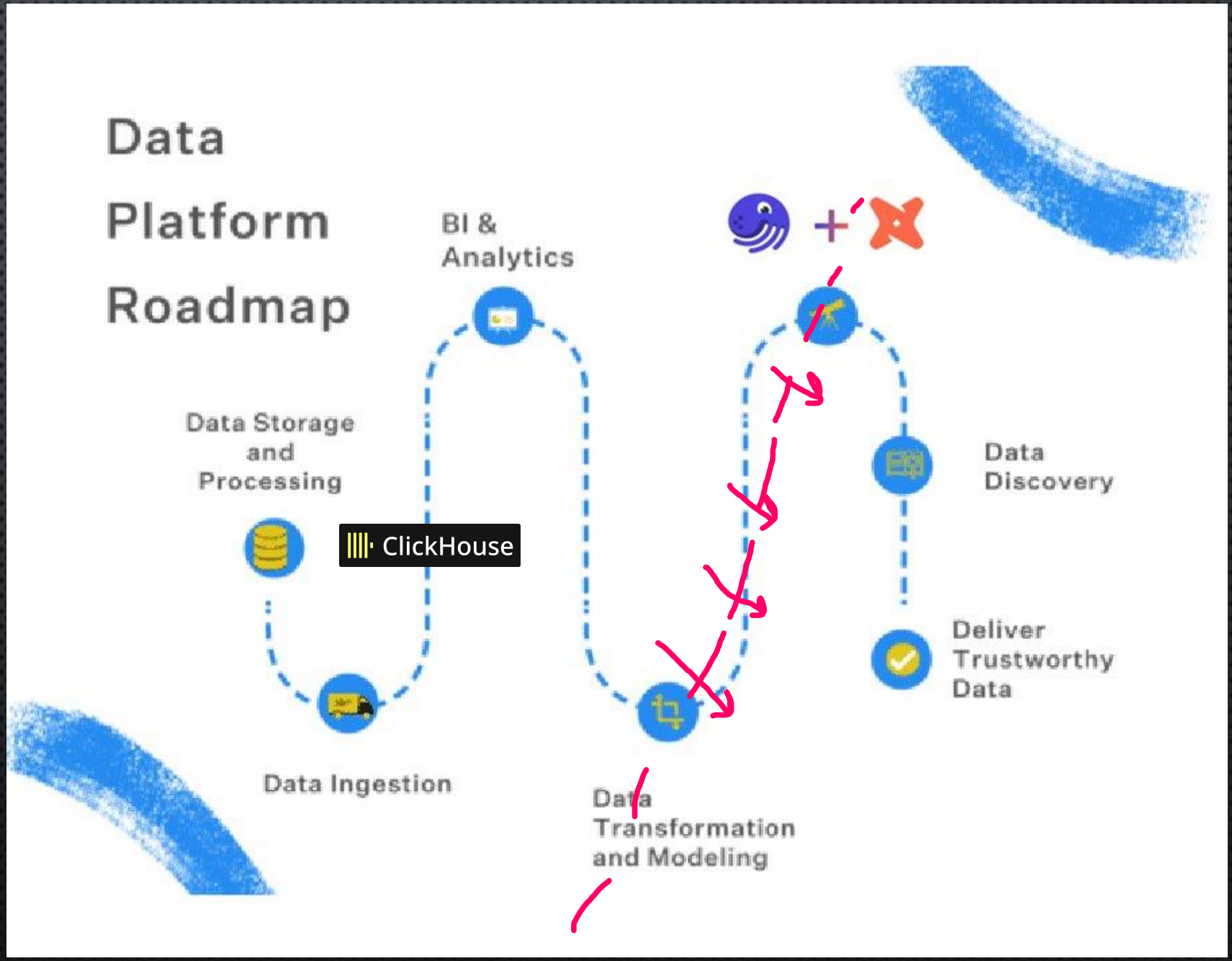


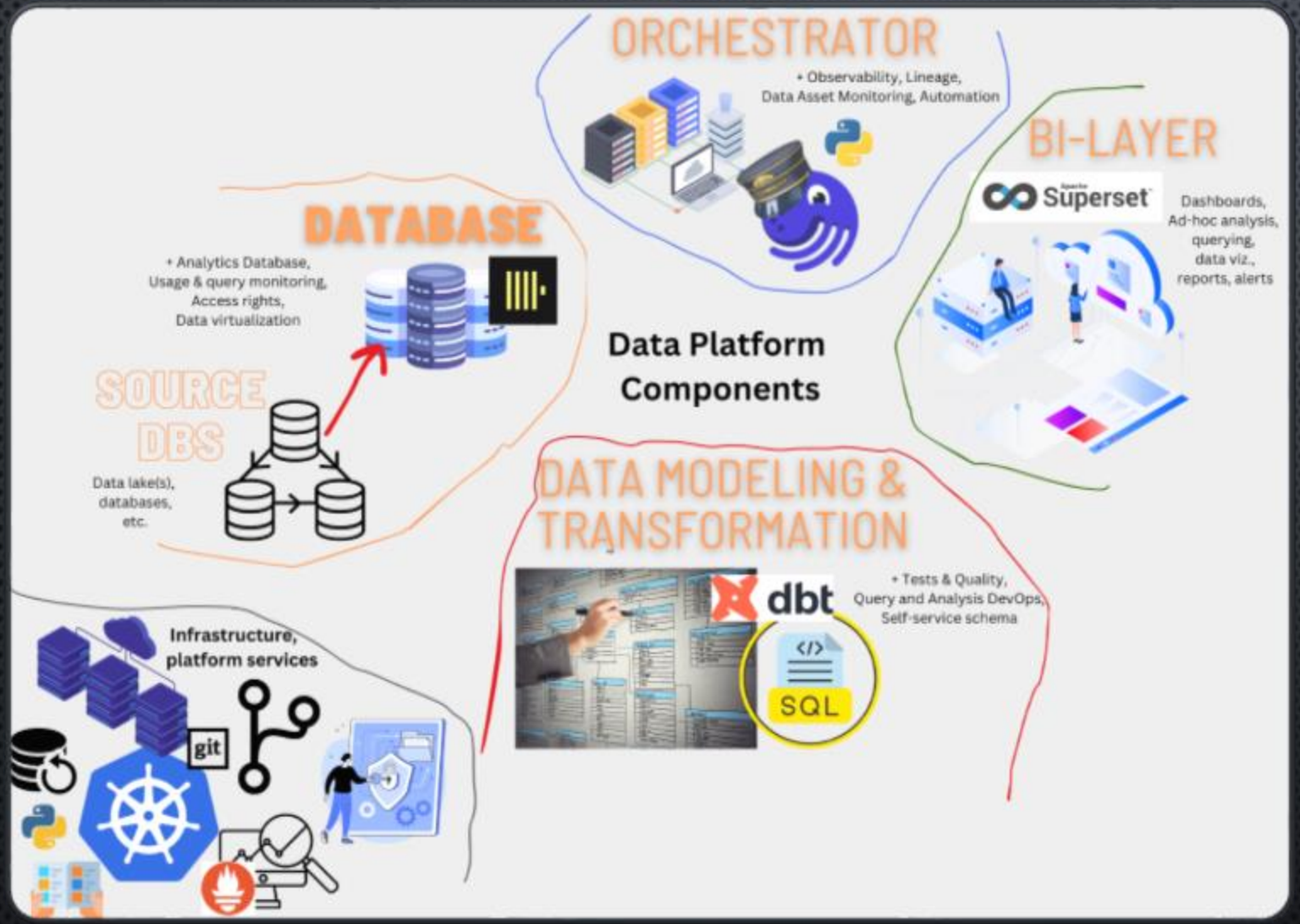
DON'T GO ALONE, TAKE THIS!

- Dagster Orchestrated Data Build Tool (DBT) Platform
- Integration with data orchestrator Dagster
- Self-service dbt + python data platform for 'analytics engineering'



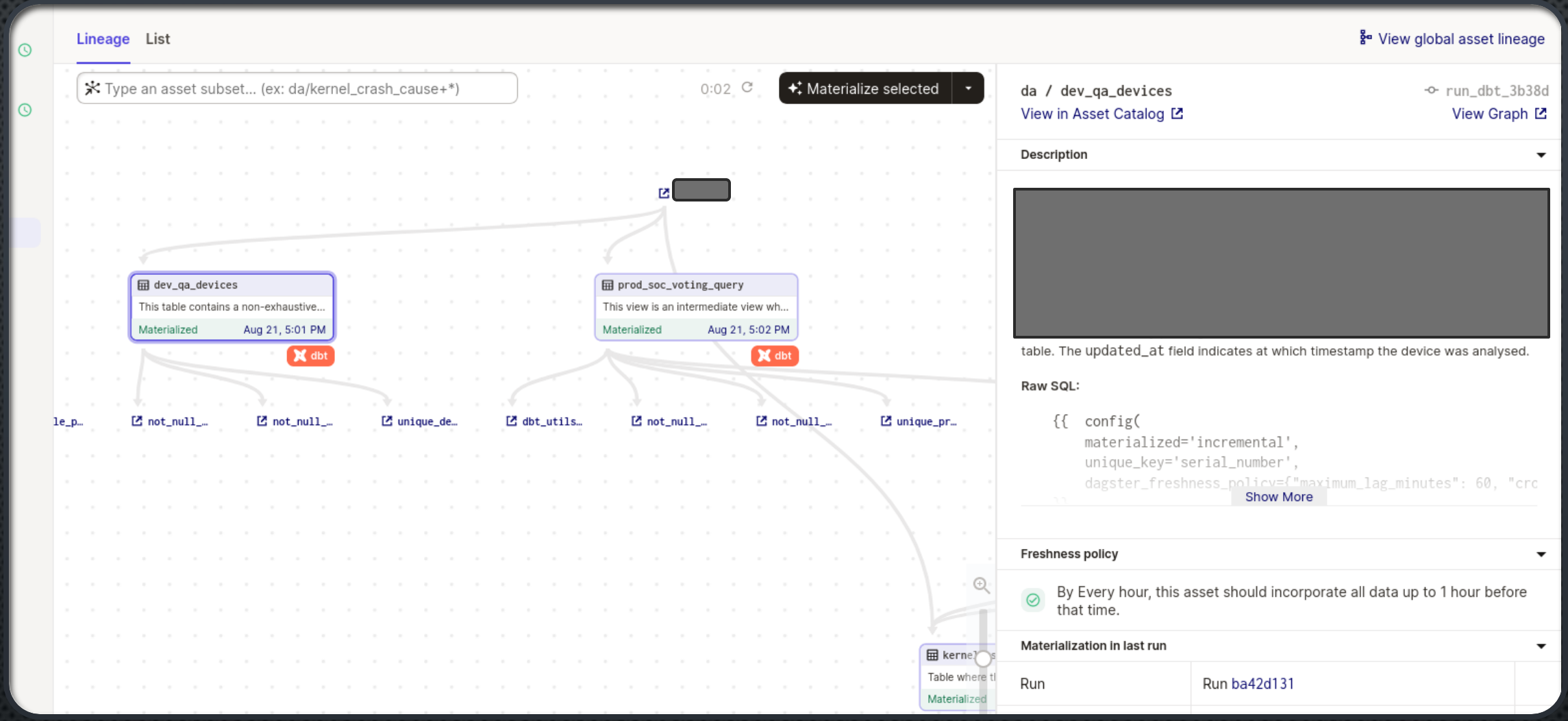
BATTLE PLANS





BATTLEPLANS

- ClickHouse: Fast analytical database
- dbt: Data transformation and modeling
- Dagster: Data orchestrator
- Superset: BI Dashboard



DBT + DAGSTER DATA TRANSFORMATION AND MODELING PLATFORM

BY ORCHESTRATING DBT WITH DAGSTER: AUTOMATED, OBSERVABLE, DOCUMENTED, AND EVEN ALLOWS CONSUMERS OF THE ANALYSIS DATA PRODUCTS TO SELF-SERVICE (MONITOR OR TRIGGER DATA REFRESH, SEE ISSUES, STATUS, TESTS/QUALITY, DOCUMENTATION)

SLAYING THE DRAGON

- ClickHouse-dbt plugin limitations
- Sharded and Distributed setup
- New 'cluster' workaround
- Trade-offs

Treasures and traps

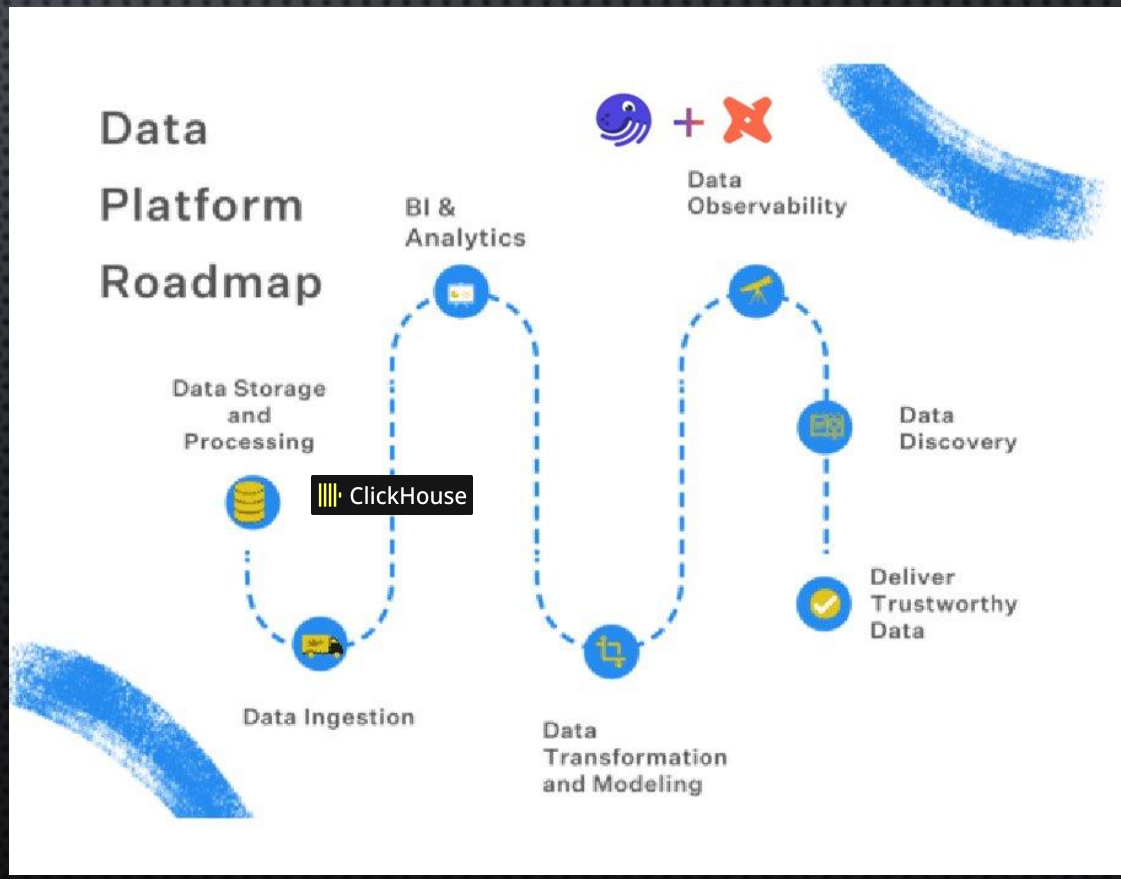
- Importance of community support, open source.
  - What we want does not exist as a product, but we can build it.
- Data and code versioning
- Observability and reporting
- Governance is hard

What was not an issue

- Data duplication
  - Compression is high level magic.
  - (In fact, it provided some user value).

FUTURE ROADMAP

- Rapid progress in the clickhouse-community plugin
  - Distributed Tables
- Areas for further support
  - Distributed Tables
  - Materialized Views.
  - DBT and clickhouse driver.
  - Python integration.
  - Time to Live (TTL)
- Next steps in our data platform journey
  - Intelligent User limits.



Have we tamed the data dragon, or has our adventure just begun?

YOUR TURN: ANY QUESTIONS?