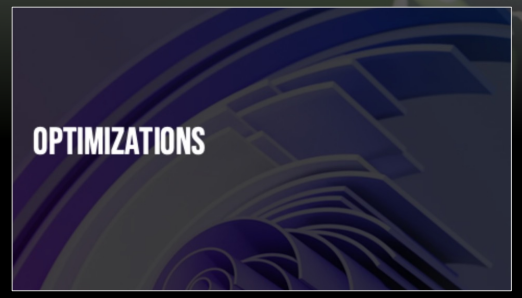
INCREFF
Incredible Efficiency

UNLOCKING THE POWER OF BIG DATA: SCALABLE BI WITH CLICKHOUSE AT INCREFF

Navaneet, Business Intelligence Lead @ Increff

TOPICS



INTRODUCTION

ABOUT INCREFF

We are a retail SaaS company simplifying complex inventory management & supply chain challenges. Our data-driven solutions empower brands to automate decision-making, bring accuracy to processes and drive sustainable retailing.

OMNI SOLUTION:

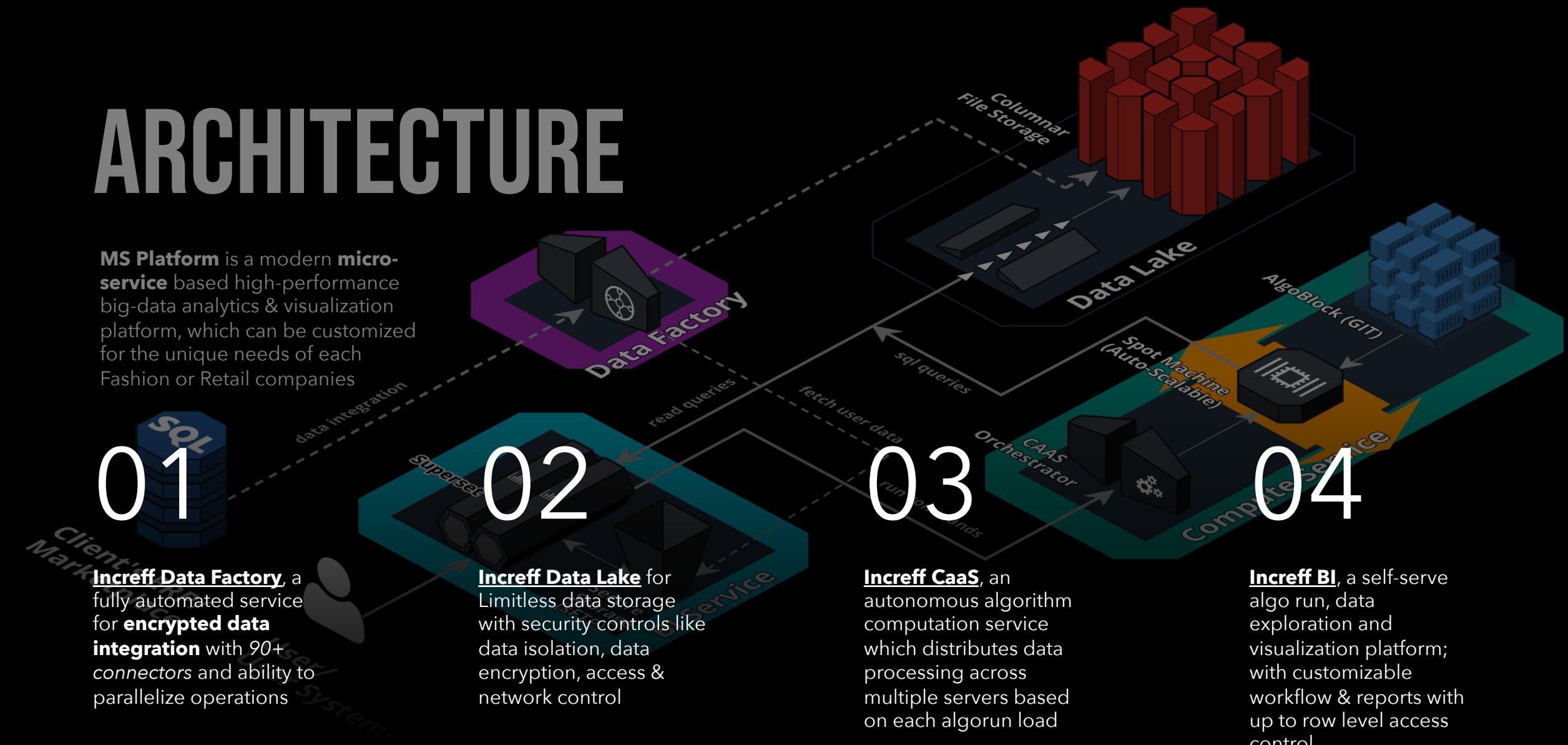
Unlock high scalability and efficiency with frictionless warehousing and omnichannel order management.

MERCHANDISING SOFTWARE:

Empower buyers and planners to maximize sales opportunities and unlock hidden profits with a data-driven approach.

ARCHITECTURE

MS Platform is a modern **micro-service** based high-performance big-data analytics & visualization platform, which can be customized for the unique needs of each Fashion or Retail companies



01

Increff Data Factory, a fully automated service for **encrypted data integration** with *90+ connectors* and ability to parallelize operations

02

Increff Data Lake for Limitless data storage with security controls like data isolation, data encryption, access & network control

03

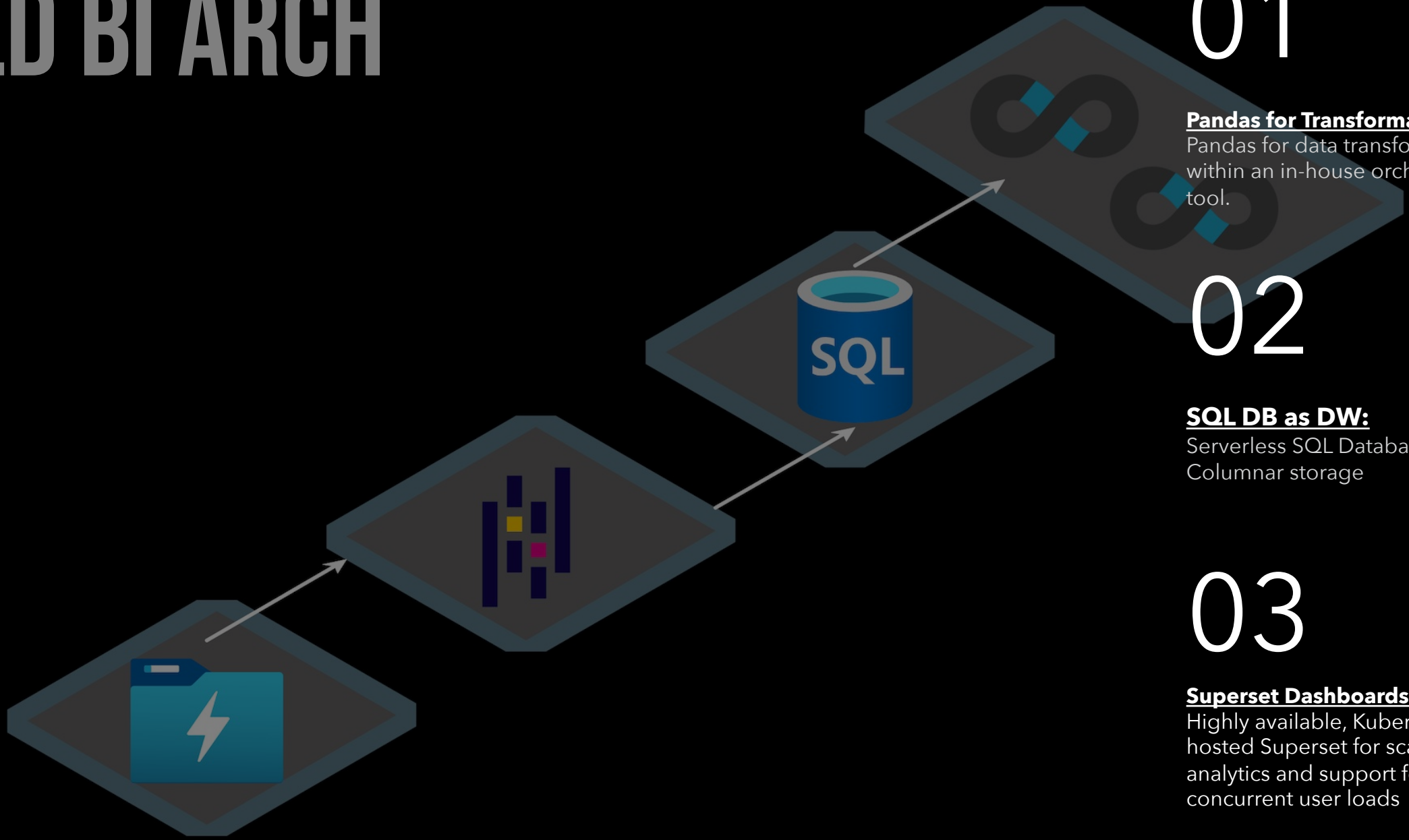
Increff CaaS, an autonomous algorithm computation service which distributes data processing across multiple servers based on each algorun load

04

Increff BI, a self-serve algo run, data exploration and visualization platform; with customizable workflow & reports with up to row level access control

CHALLENGES & SCALE

OLD BI ARCH



01

Pandas for Transformations:
Pandas for data transformations within an in-house orchestration tool.

02

SQL DB as DW:
Serverless SQL Database with Columnar storage

03

Superset Dashboards:
Highly available, Kubernetes-hosted Superset for scalable analytics and support for large concurrent user loads

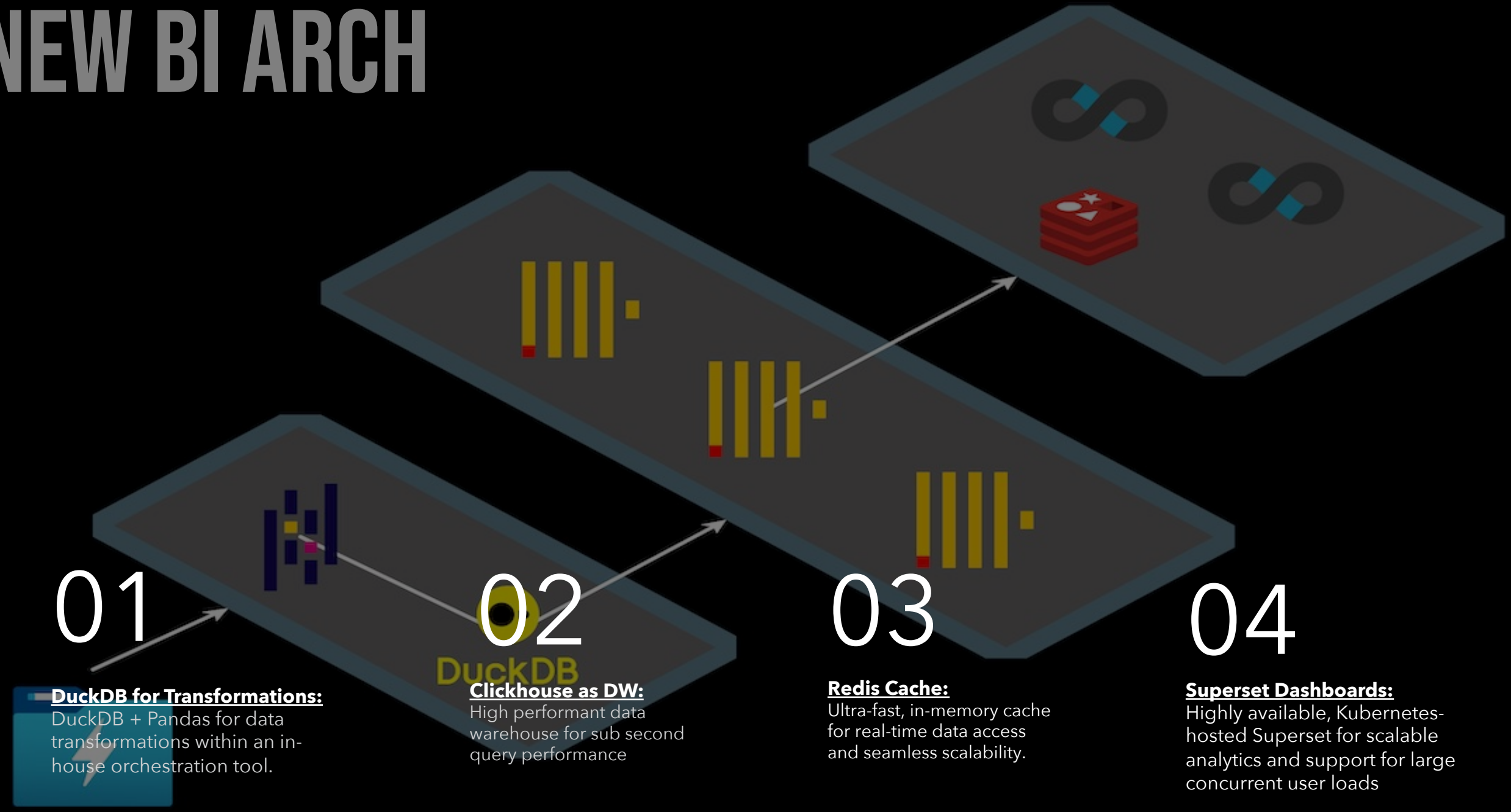
CHALLENGES

Expected Performance as a small scale DW for small BI workload

Recently we onboarded one of the India's biggest fashion retail company and we were not able to scale our existing DW.

Clickhouse was benchmarked for 3X the expected scale.

NEW BI ARCH



01

DuckDB for Transformations:
DuckDB + Pandas for data transformations within an in-house orchestration tool.

02

Clickhouse as DW:
High performant data warehouse for sub second query performance

03

Redis Cache:
Ultra-fast, in-memory cache for real-time data access and seamless scalability.

04

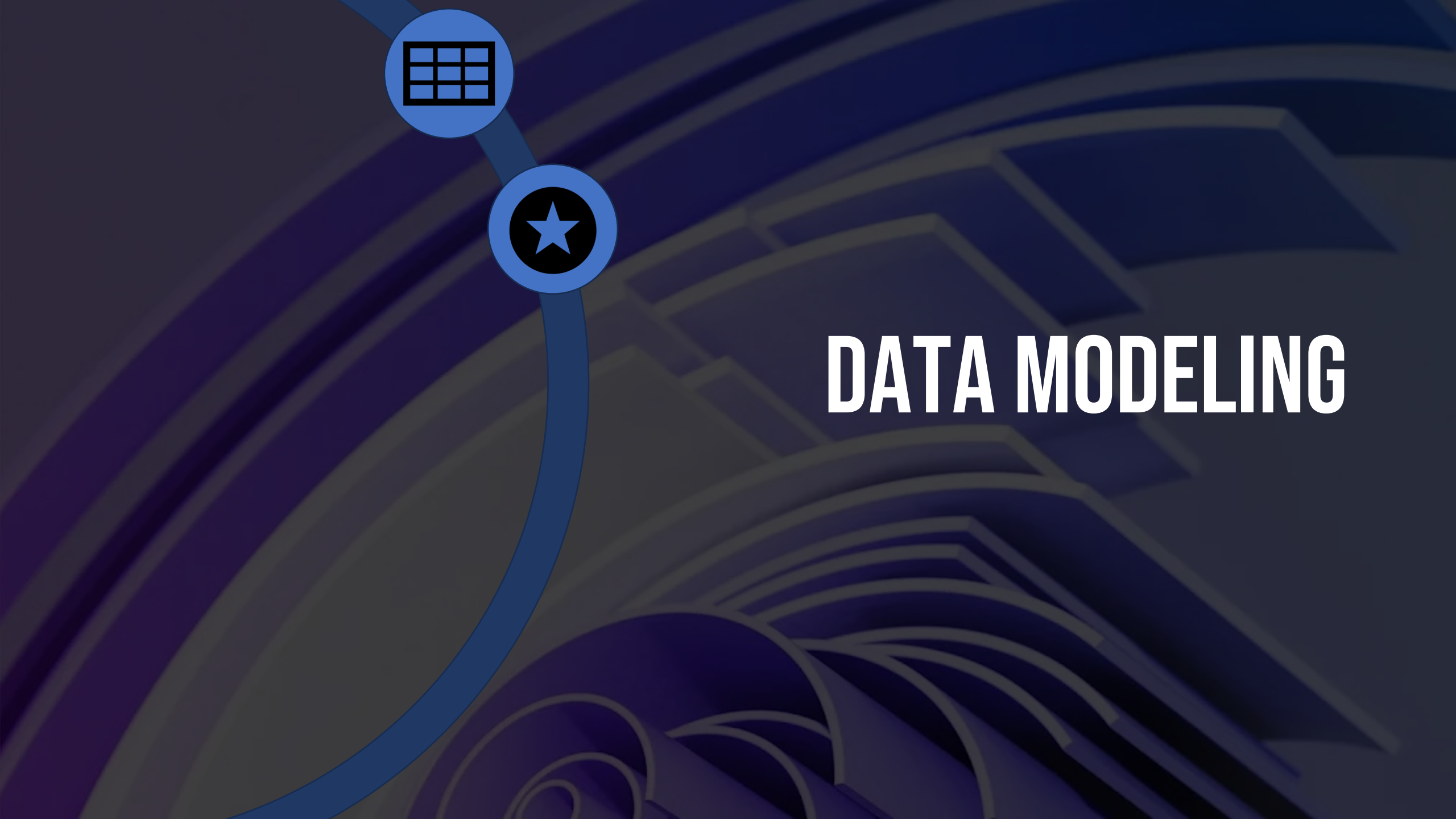
Superset Dashboards:
Highly available, Kubernetes-hosted Superset for scalable analytics and support for large concurrent user loads

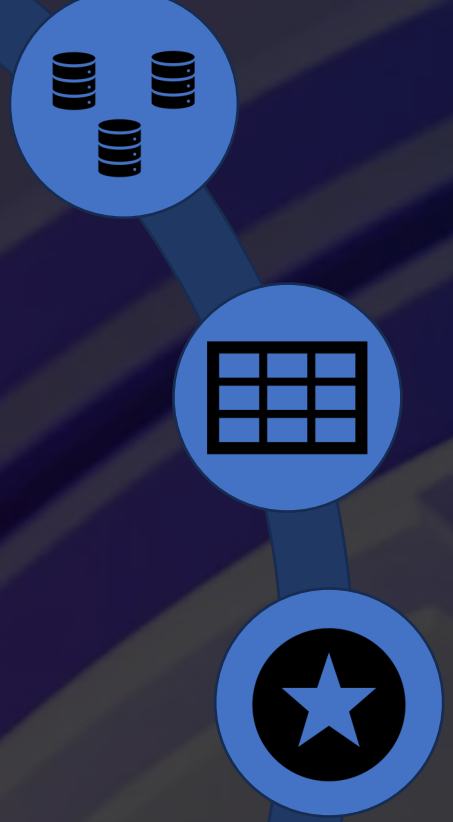
SCALABILITY

- Scale - 200M
- P90 < 1s
- P95 < 3s
- P99 < 5s

- Scale - 1B
- P90 < 1s
- P95 < 3s
- P99 < 5s

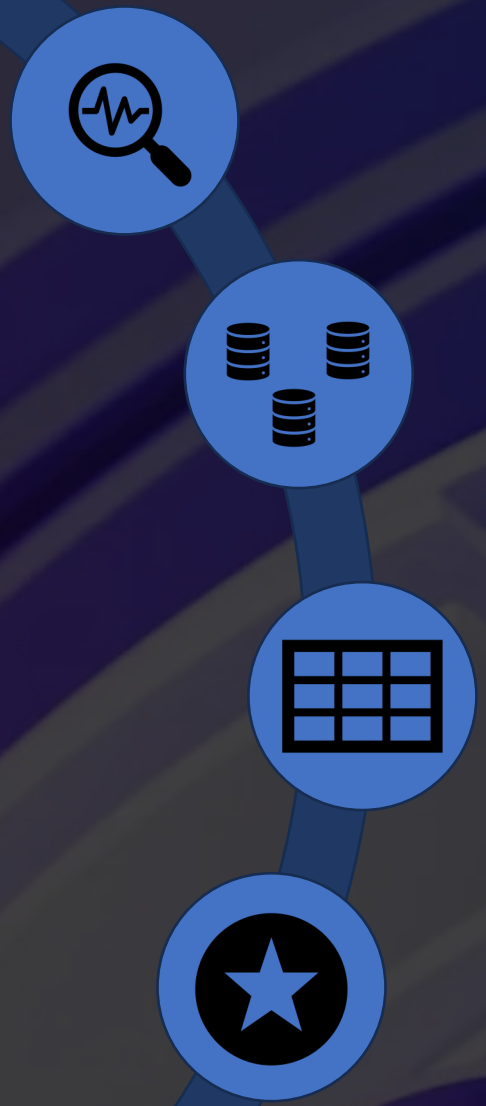
OPTIMIZATIONS


DATA MODELING



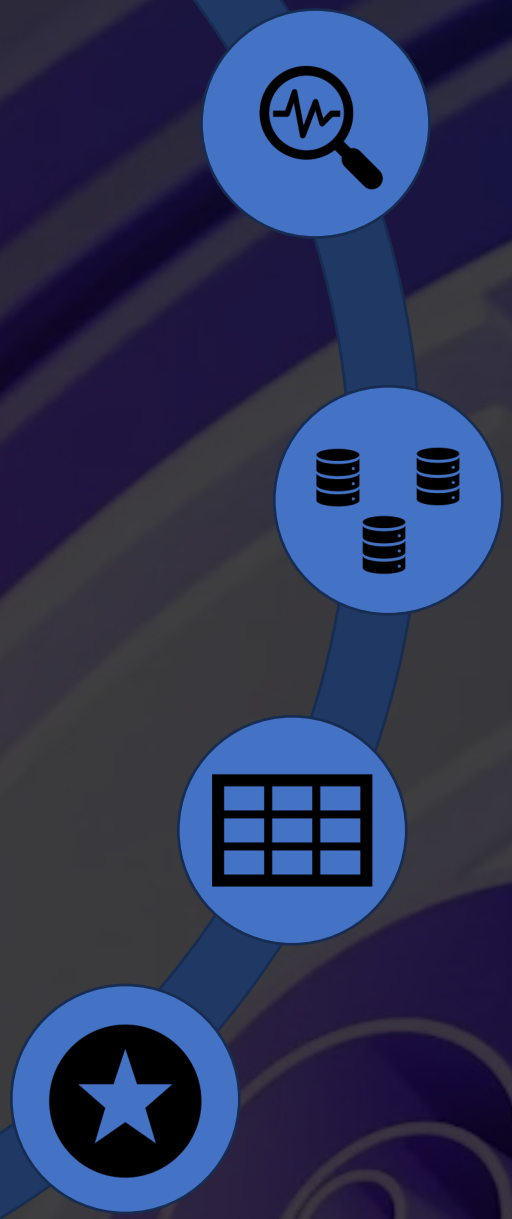
STAR SCHEMA

- A constellation schema utilizes shared dimensions across fact tables with different granularities (e.g., day, week, month).
- Since most dashboards required data at the month and week levels, the Fact Constellation approach reduced the load on the ClickHouse server.



ONE BIG TABLE (OBT)

- For custom scenarios that demand varying levels of data granularity.
- To maintain simplicity in the core star schema.

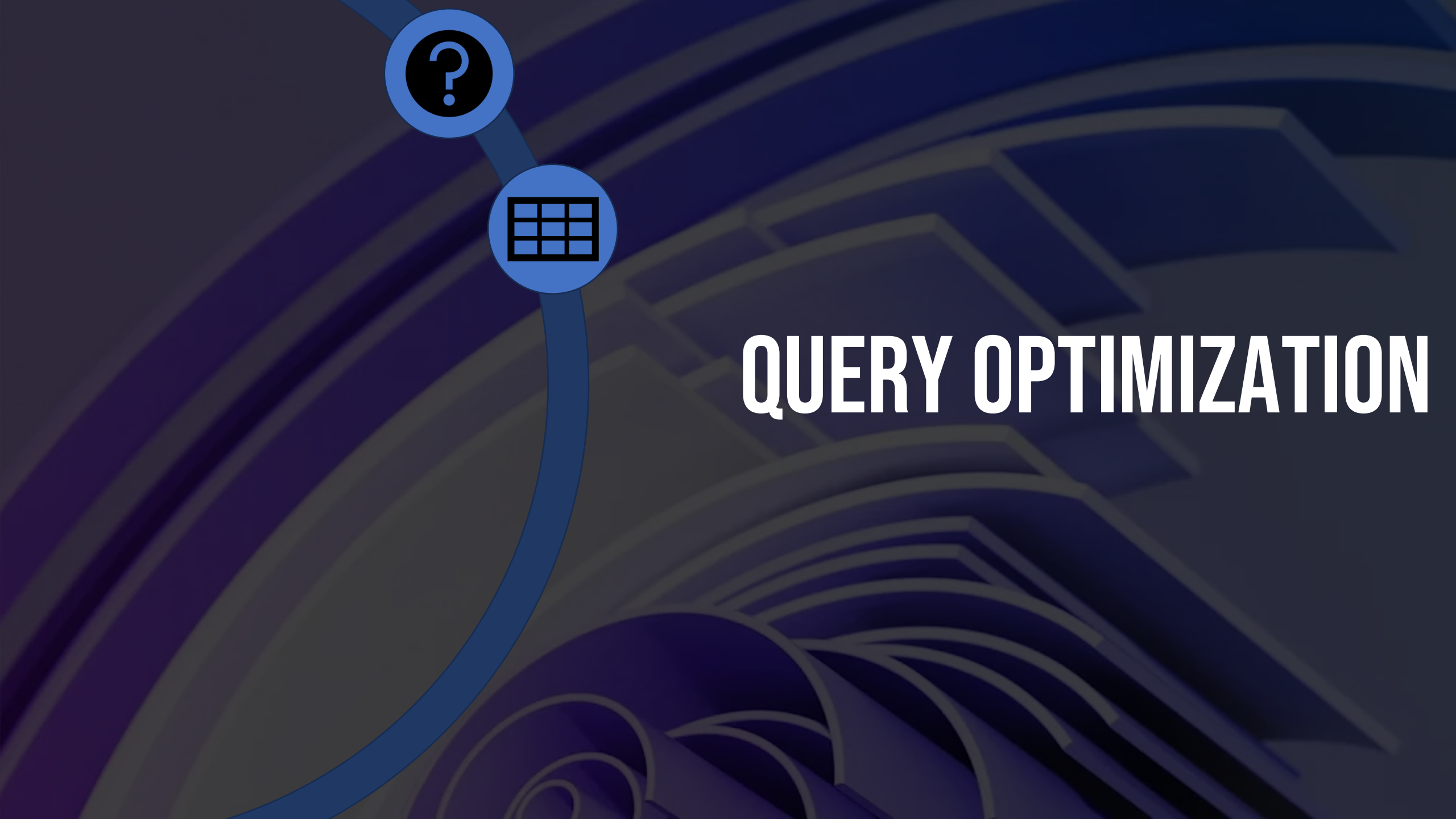


PARTITION AND PRIMARY KEYS

- Defining a primary key that reflected our query patterns.
- Partitioning tables by a column that is frequently used in filters or range queries. In our case, our fact tables are partitioned at date granularity.

RIGHT DATA TYPES

- Data types significantly influence compression. Choosing the appropriate data type enhances compression, leading to faster read operations.
- For example, we applied the LowCardinality data type to all categorical columns in our dimension tables, which boosted read query performance through dictionary encoding.


QUERY OPTIMIZATION

JINJA IN APACHE SUPERSET

- Leveraging Jinja in Apache Superset to minimize the use of joins, with join conditions applied only when absolutely necessary.

PREWHERE

- Utilizing PREWHERE greatly improved the performance of specific queries by optimizing the default execution plan.

FUTURE OPTIMISATIONS:

While the current architecture has been tested to support a 3X scale, there are several optimizations that we plan to explore and benchmark for future scalability needs, including:

- Projections
- Materialized Views
- Using Dictionaries for joins
- Fully denormalizing to a One Big Table (OBT)

FUTURE ROADMAP

FUTURE ROADMAP

- Prompt based data extraction
- Prompt based data visualization & dashboard creation
- Open Source Clickhouse implementation
- UI based file browser to manually upload or automatically fetch from external systems


THANK YOU