IntelliPaat

Dell - Data Engineering Training

Uday Kumar – Data Platform Architect

Day 4

Data Engineering Training

Agenda

1. Dell Object Storage Support
2. What are data products?
3. Data product's Components
4. Value proposition of data products
5. Lifecycle of a data product
6. Introduction to Starburst
7. Overview of Starburst
8. Key features and capabilities
9. Use cases for Starburst in building Data Products

DELL OBJECT STORAGE SUPPORT

Dell ECS and **ObjectScale** are high performance object storage systems that are compatible with **Amazon S3** and any catalog using one of the following connectors:

- Starburst Delta Lake connector
- Starburst Hive connector
- Starburst Iceberg connector

The requirements of the **connector apply when using Dell ECS or ObjectScale** as a storage backend. This specifically includes the configured metastore and the network access between the cluster and the storage.

Data Products

INTRODUCTION TO DATA PRODUCTS

DJ Patil, former **United States Chief Data Scientist**, defined a data product as “a product that facilitates an end goal through the use of data” (from his book Data Jujitsu: The Art of Turning Data into Product, 2012).

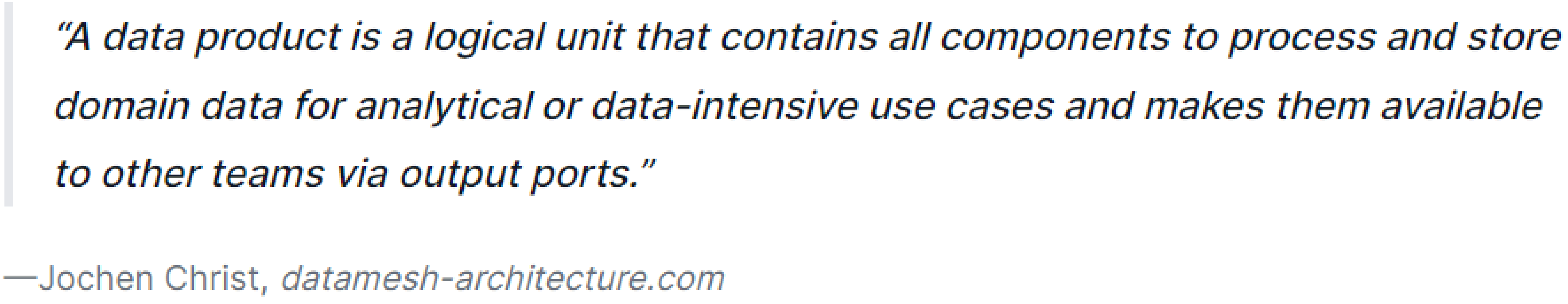
Digital product or feature can be considered a “data product” if it uses data to facilitate a goal. For example, the home page of a digital newspaper can be a data product if the news items featured in the home page I see are dynamically selected based on my previous navigation data.

Data products are groups them by type: **raw data, derived data, algorithms, decision support and automated decision-making**.

INTRODUCTION TO DATA PRODUCTS

Data Products are the foundational **building block of an enterprise Data Mesh.** But what exactly is a Data Product, how do they work, how can they be identified, and how can they be built quickly?

A data product is a **logical unit that contains all components** to process domain data and provide data sets via output ports for analytical use.



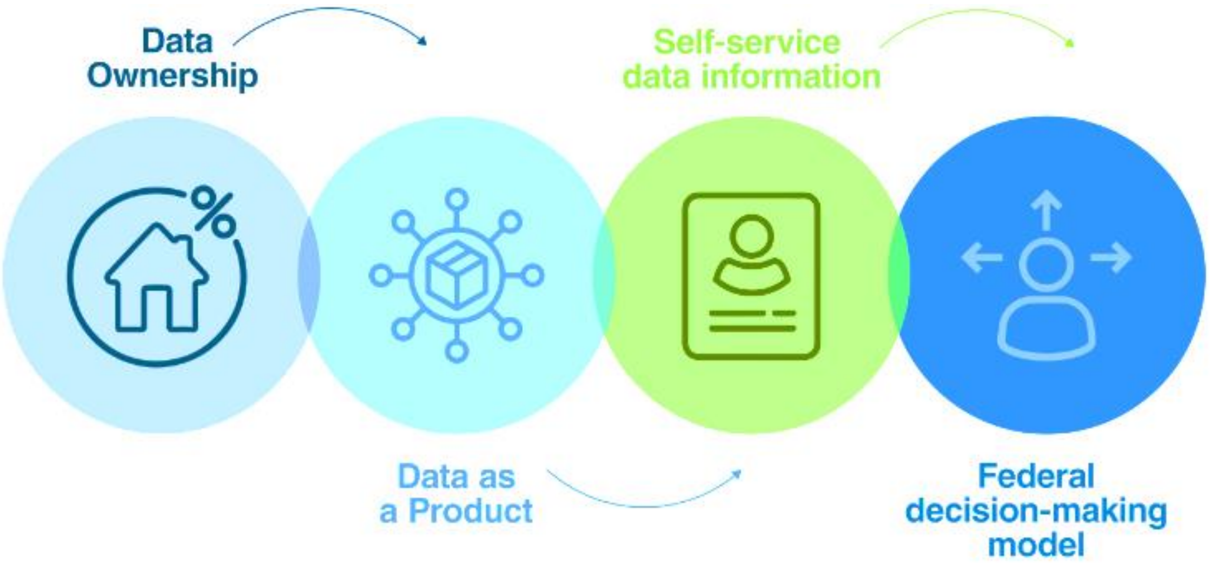
> *"A data product is a logical unit that contains all components to process and store domain data for analytical or data-intensive use cases and makes them available to other teams via output ports."*

—Jochen Christ, *datamesh-architecture.com*

DATA MESH

Data Mesh is a paradigm shift in big analytical data management that addresses some of the limitations of the past paradigms, data warehousing and data lake. **Data Mesh is founded in four principles: "domain-driven ownership of data", "data as a product", "self-serve data platform" and a "federated computational governance".**

Data mesh is not a **data storage technology like an enterprise data lake or data warehousing services**. Instead, it's **a journey toward building an ecosystem** where teams can access the data they need at the speed of business and use the insights to respond rapidly to changing market demands



DATA PRODUCTS - EXAMPLE

A company dashboard to visualise the main KPIs of your business

A data warehouse

Recommended restaurants nearby

“faster route now available” notification on Google Maps

A self-driving car

DATA PRODUCT LIFECYLE

Identify the problem and define the objective

Conceptualize the idea

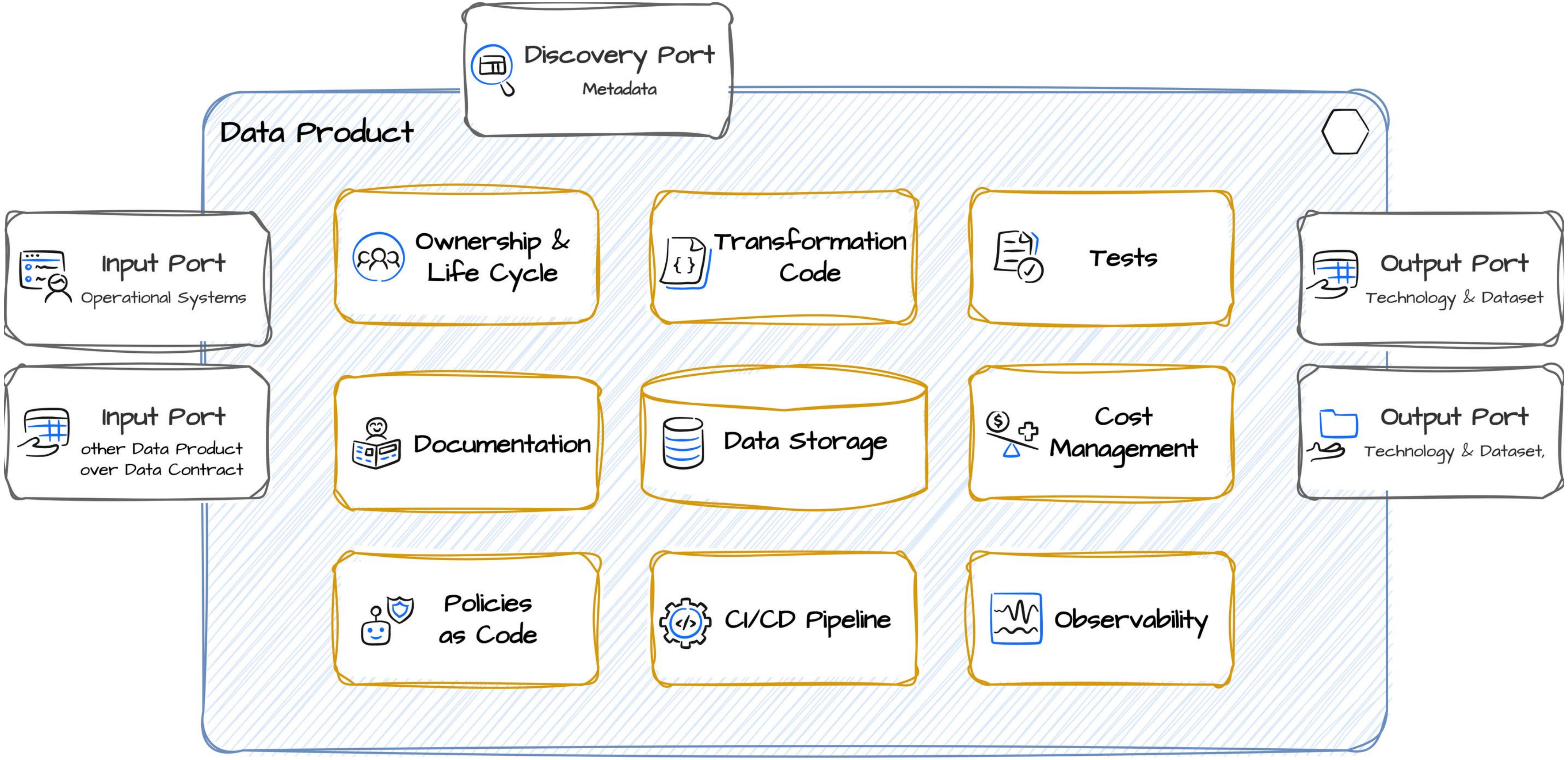
Conduct market research

Gather user data

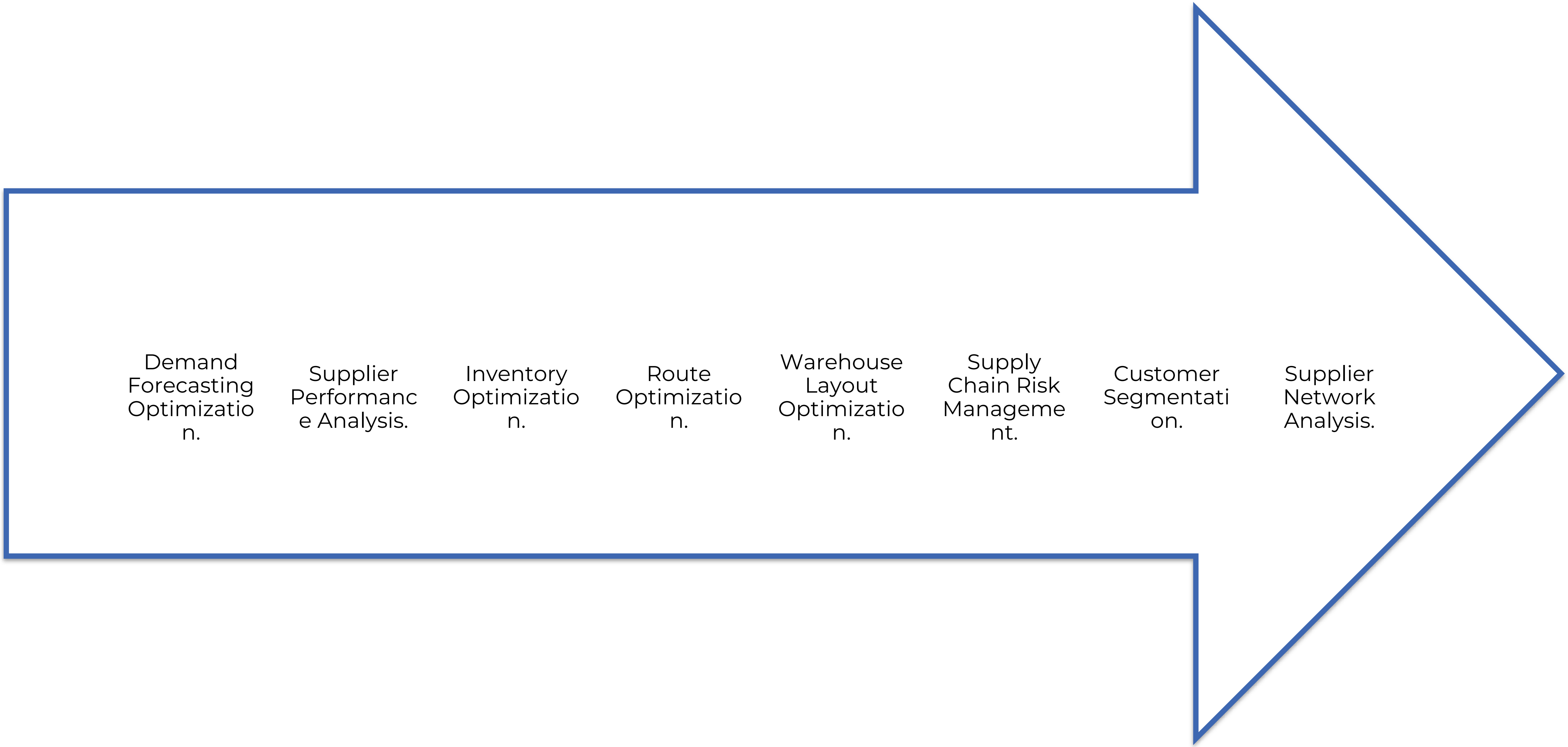
Decide on the architecture and framework

Design the data product

DATA PRODUCT ARCHITECTURE



FEW DATA PRODUCTS IN SUPPLY CHAIN MANAGEMENT



Demand Forecasting Optimization.

Supplier Performance Analysis.

Inventory Optimization.

Route Optimization.

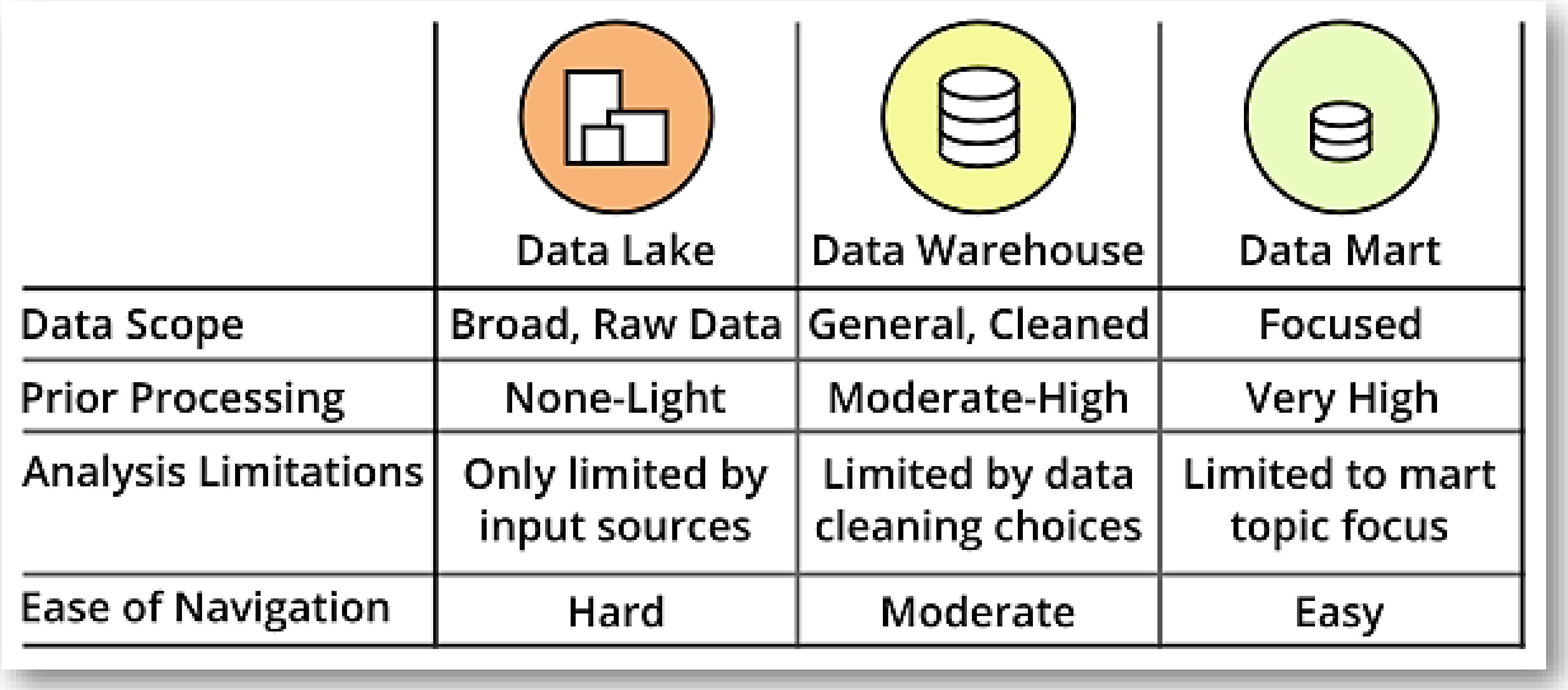
Warehouse Layout Optimization.

Supply Chain Risk Management.

Customer Segmentation.

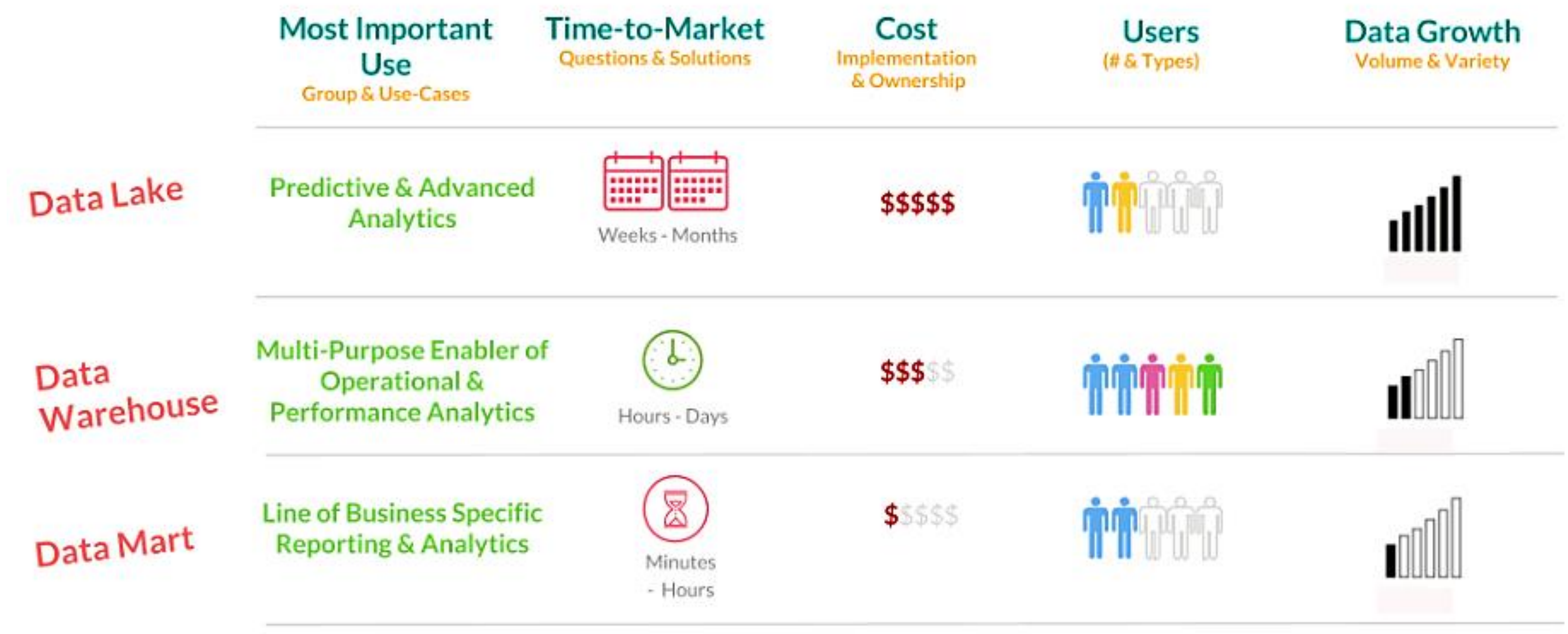
Supplier Network Analysis.

COMPARISON DATA LAKE, DATA WAREHOUSE AND DATA MART



	Data Lake	Data Warehouse	Data Mart
Data Scope	Broad, Raw Data	General, Cleaned	Focused
Prior Processing	None-Light	Moderate-High	Very High
Analysis Limitations	Only limited by input sources	Limited by data cleaning choices	Limited to mart topic focus
Ease of Navigation	Hard	Moderate	Easy

COMPARISON DATA LAKE, DATA WAREHOUSE AND DATA MART



	Most Important Use Group & Use-Cases	Time-to-Market Questions & Solutions	Cost Implementation & Ownership	Users (# & Types)	Data Growth Volume & Variety
Data Lake	Predictive & Advanced Analytics	Weeks - Months	$$$$$		
Data Warehouse	Multi-Purpose Enabler of Operational & Performance Analytics	Hours - Days	$$$$$		
Data Mart	Line of Business Specific Reporting & Analytics	Minutes - Hours	$$$$$		

DATA PRODUCTS – AMAZON SMART WAREHOUSE

Starburst

STARBURST OVERVIEW

The modern data lake

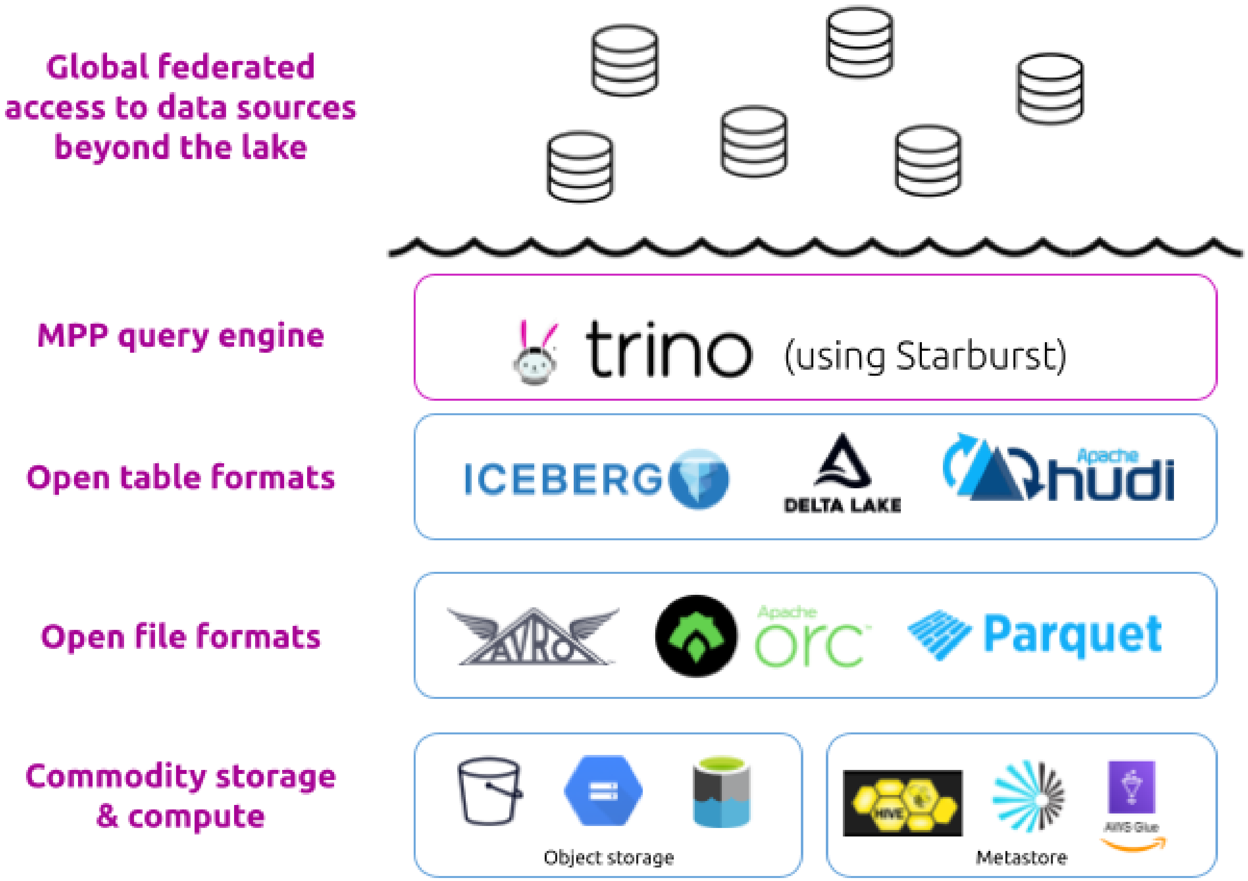
A modern data lake, also known as a **data lakehouse, is a hybrid data architecture combining the features and benefits of both data lakes and data warehouses**. Modern data lakes address the limitations and challenges associated with traditional data lakes and data warehouses, providing a more comprehensive and unified solution for data storage, processing, and analytics.

What is a modern data lake?

You can think of the **modern data lake as an extension of traditional data lakes** that uses open table formats like **Iceberg, Delta Lake, and Hudi** alongside open file formats to achieve superior features. As a data lake analytics platform, Starburst makes it easy to build and manage your modern data lake using the configuration of your choice.

STARBURST OVERVIEW

The Modern Data Lake



Global federated access to data sources beyond the lake

MPP query engine

Open table formats

Open file formats

Commodity storage & compute

Starburst

Data Lake Analytics Platform

The easiest way to *build and manage* your Modern Data Lake

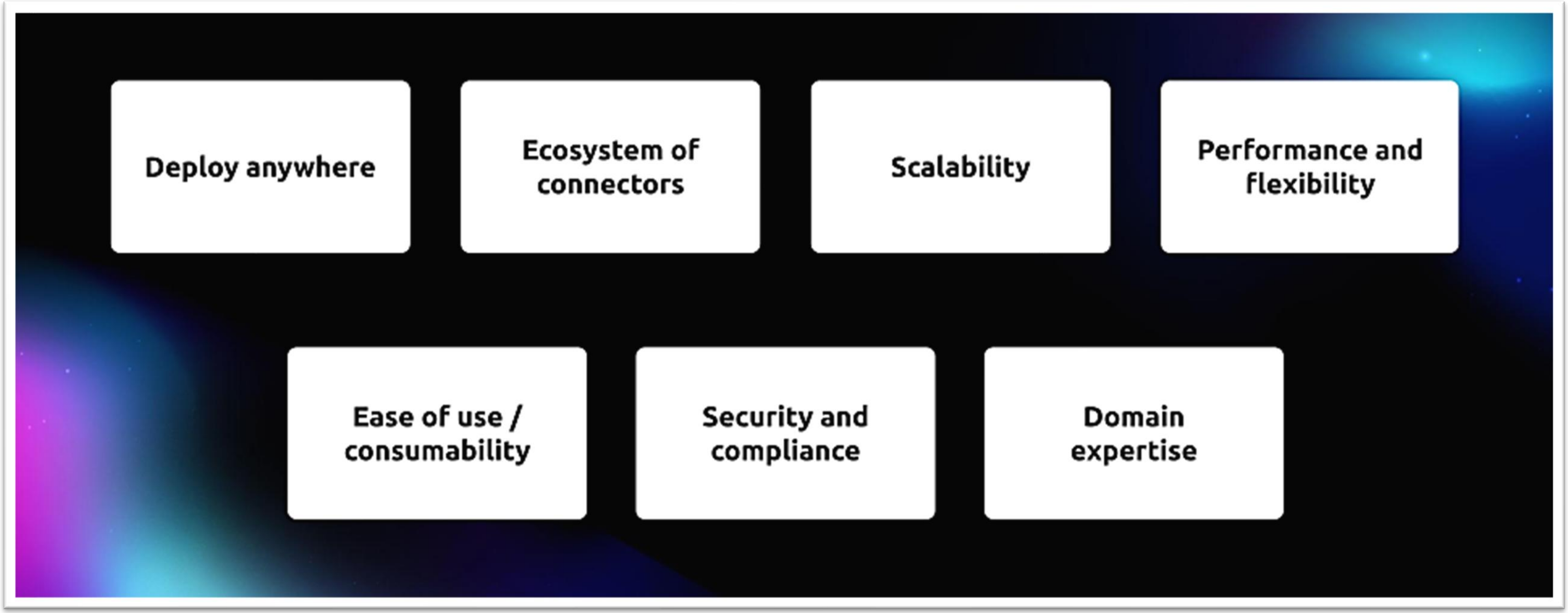
90% Faster time to insight

53% Lower TCO

100% Future-proof architecture

Starburst

STARBURST OVERVIEW



Deploy anywhere

Ecosystem of connectors

Scalability

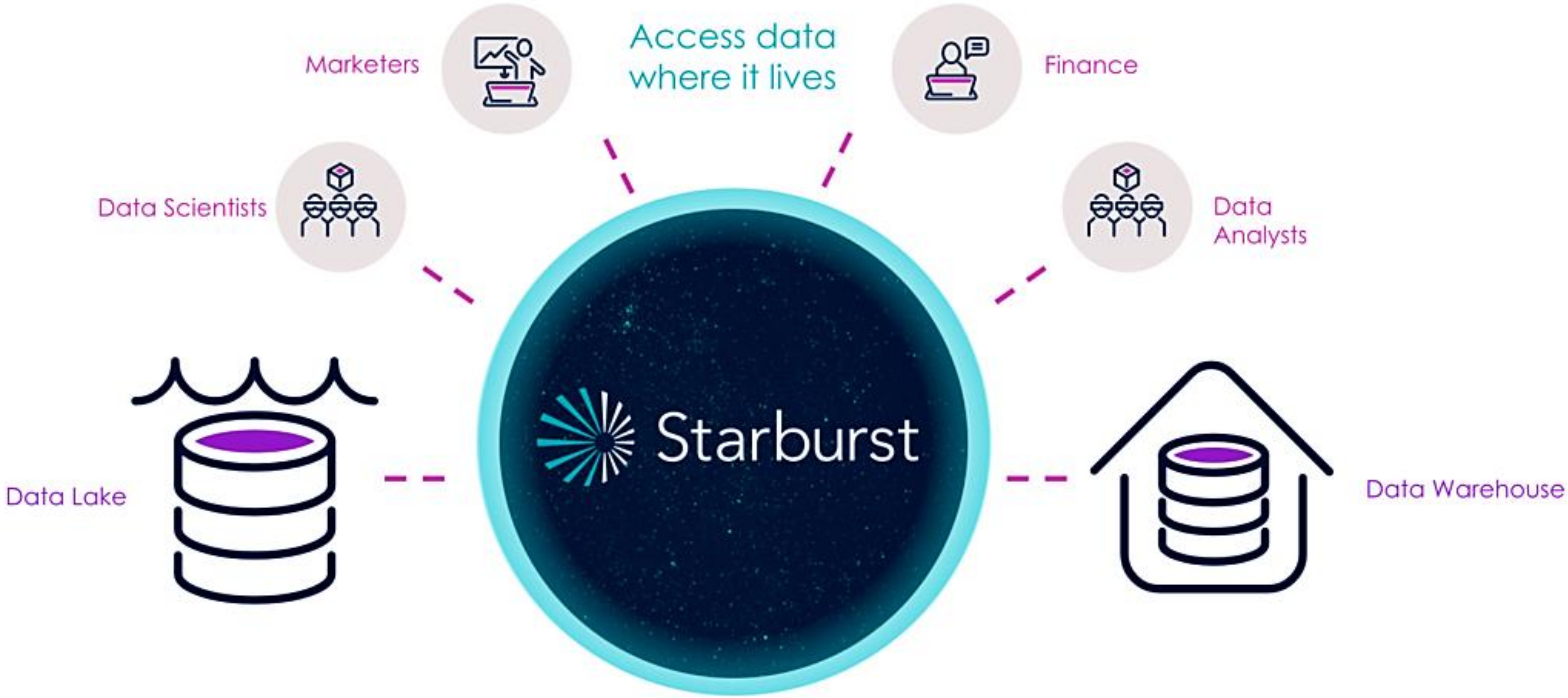
Performance and flexibility

Ease of use / consumability

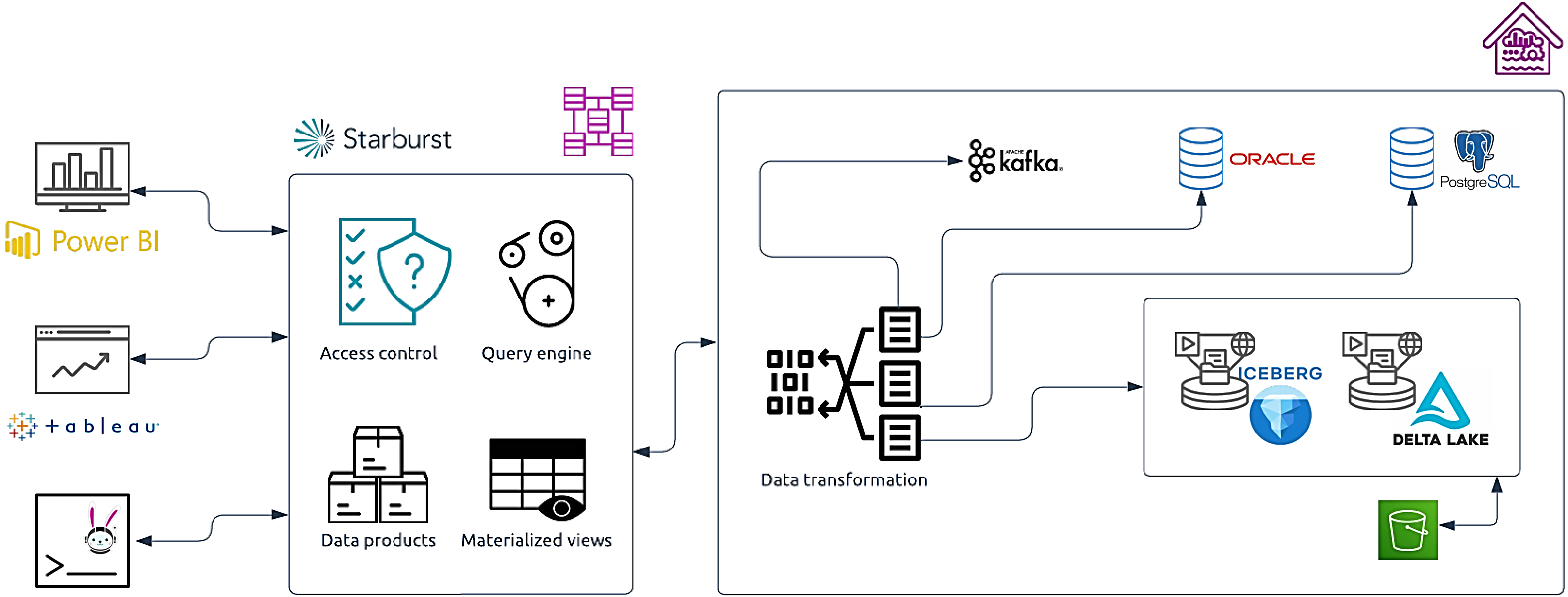
Security and compliance

Domain expertise

STARBURST OVERVIEW



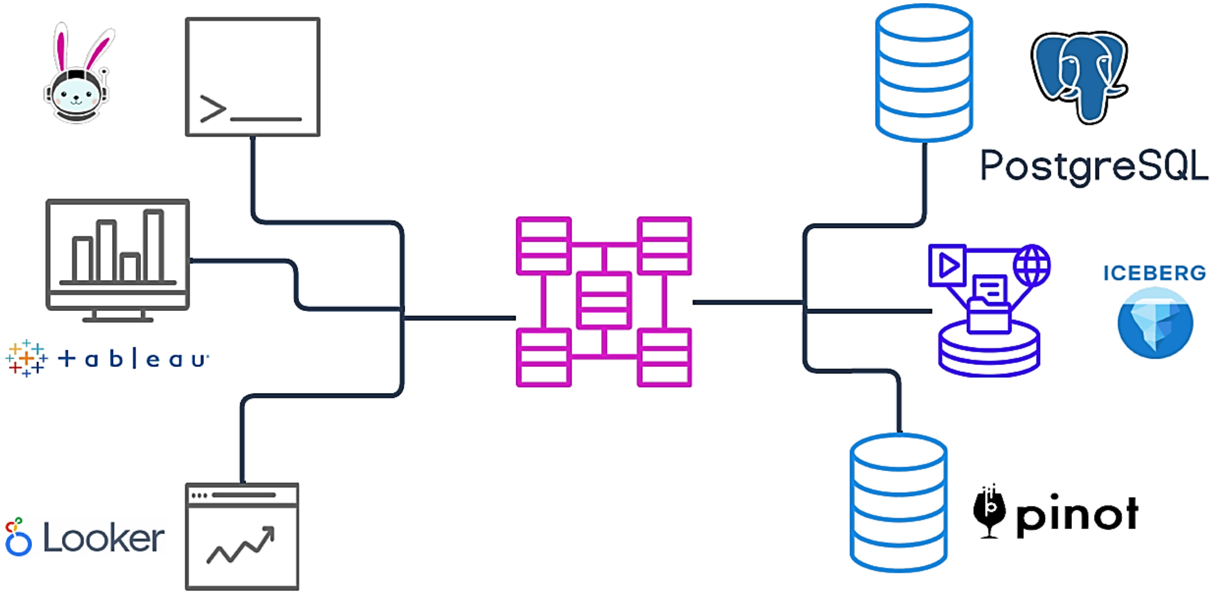
STARBURST OVERVIEW – MODERN DATA LAKE



Data Lakehouse Architecture

STARBURST ARCHITECTURE

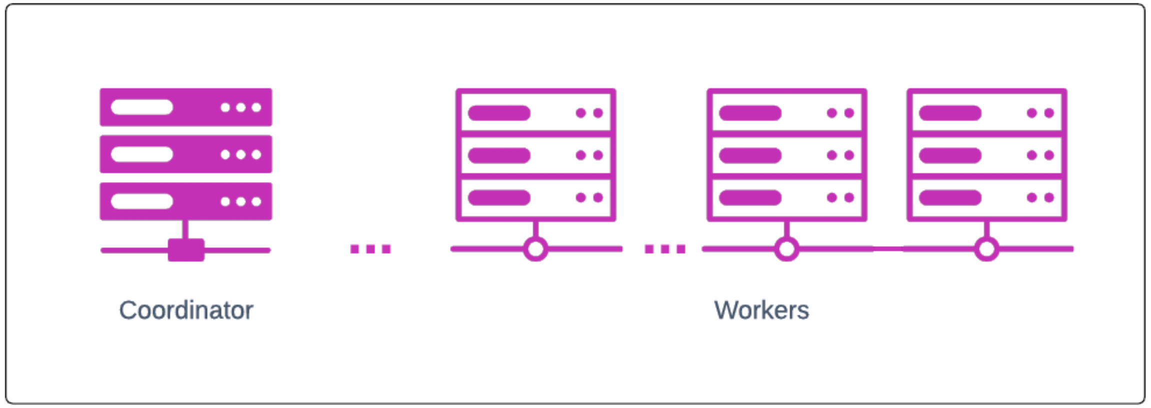
Starburst Enterprise and Starburst Galaxy are **massively parallel processing (MPP)** compute clusters running the distributed SQL query engine



A Trino cluster has two node types:

Coordinator - a single server that handles incoming queries, and provides query parsing and analysis, scheduling and planning. Distributes processing to worker nodes.

Workers - servers that execute tasks as directed by the coordinator, including retrieving data from the data source and processing data.



STARBURST ARCHITECTURE

Connectors are what allow Starburst products to separate compute from storage. The configuration necessary to access a data source is called a **catalog**. Each catalog is configured with the connector for that particular data source.

A **connector** is called when a **catalog** that is configured to use the connector is used in a query. Data source connections are established based on catalog configuration.

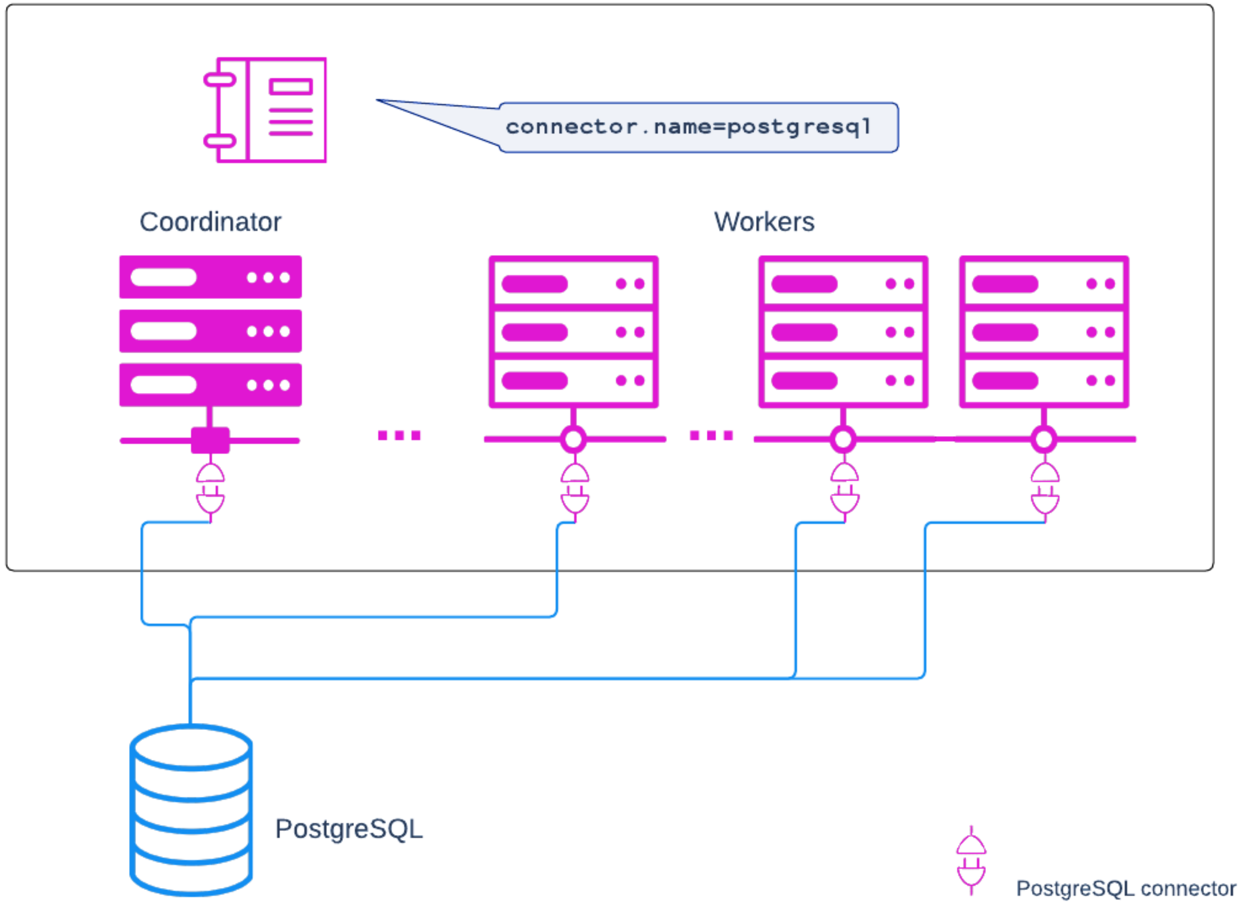
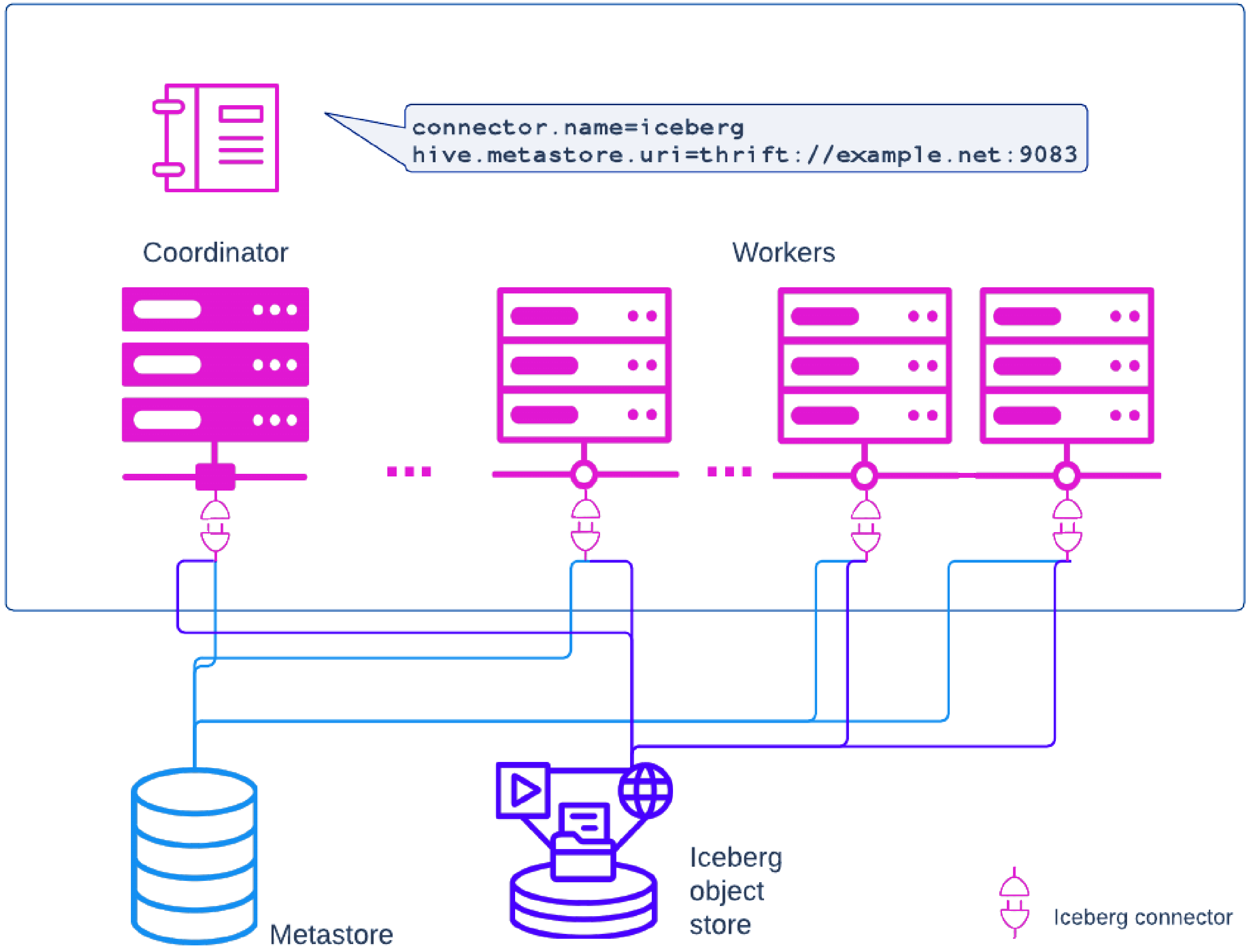


Figure - Shows how this works with PostgreSQL,

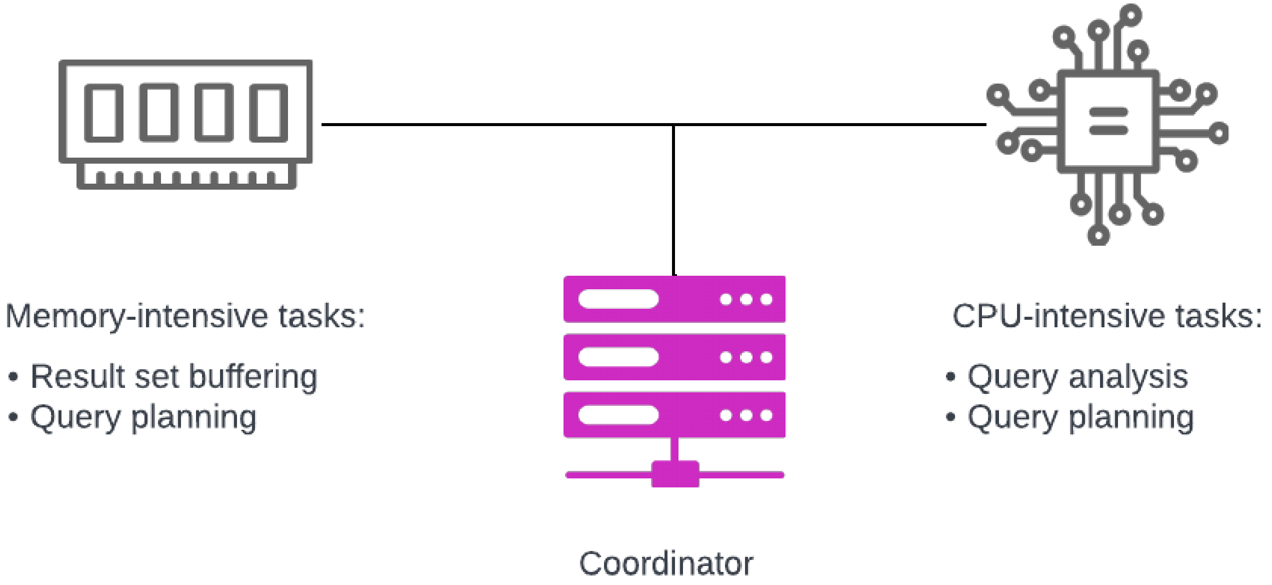
STARBURST ARCHITECTURE



```
connector.name=iceberg
hive.metastore.uri=thrift://example.net:9083
```

STARBURST ARCHITECTURE - COORDINATOR

- Query parsing and analysis
- Query planning and optimization
- Communications with clients



Memory-intensive tasks:

- Result set buffering
- Query planning

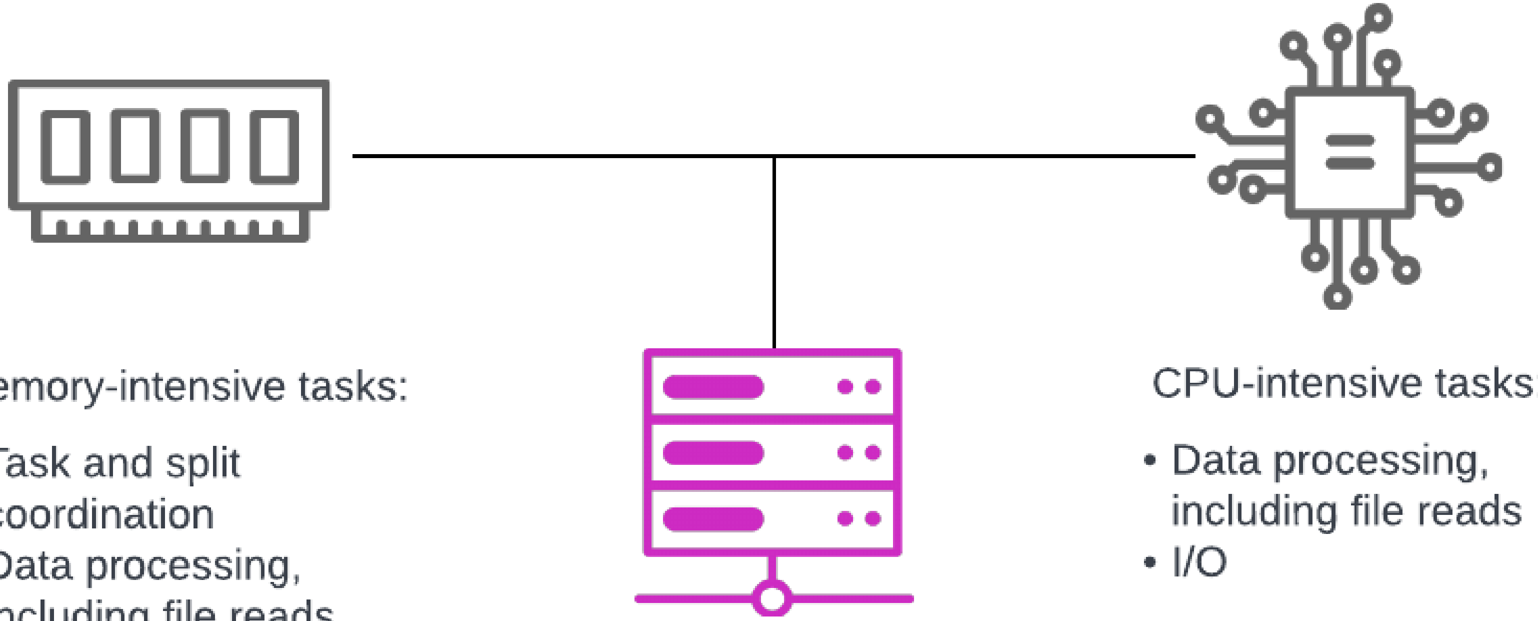
CPU-intensive tasks:

- Query analysis
- Query planning

Coordinator

STARBURST ARCHITECTURE – WORKER

- Executes Query
- Cost Optimized Query Plan



Memory-intensive tasks:

- Task and split coordination
- Data processing, including file reads
- I/O

Worker

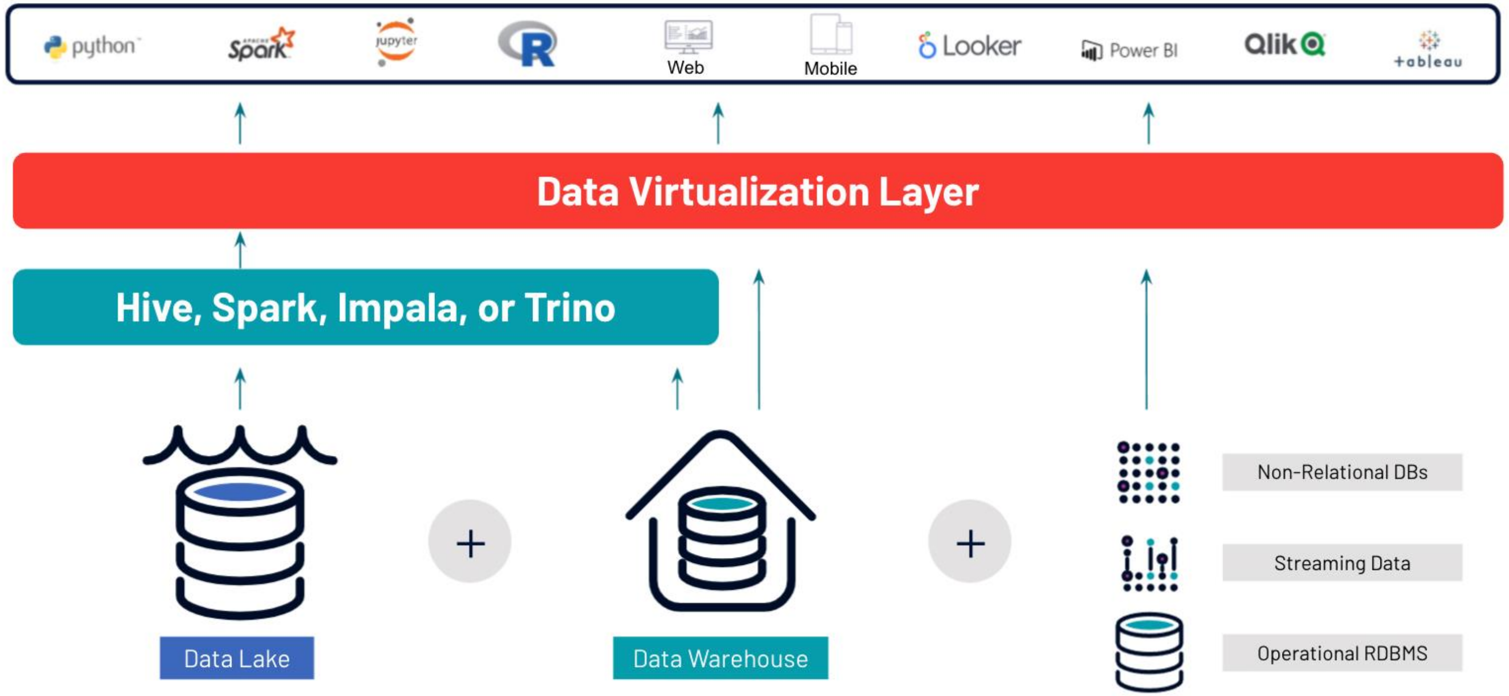
CPU-intensive tasks:

- Data processing, including file reads
- I/O

STARBURST OVERVIEW

Challenges with legacy data virtualization:

- Creates more vendor lock in.
- Must leverage MPP engines (Spark, Hive, Impala, Trino) for high performance data lake queries.
- Federation servers create performance and concurrency bottlenecks
- Requires integration for fast parallel execution against data lakes

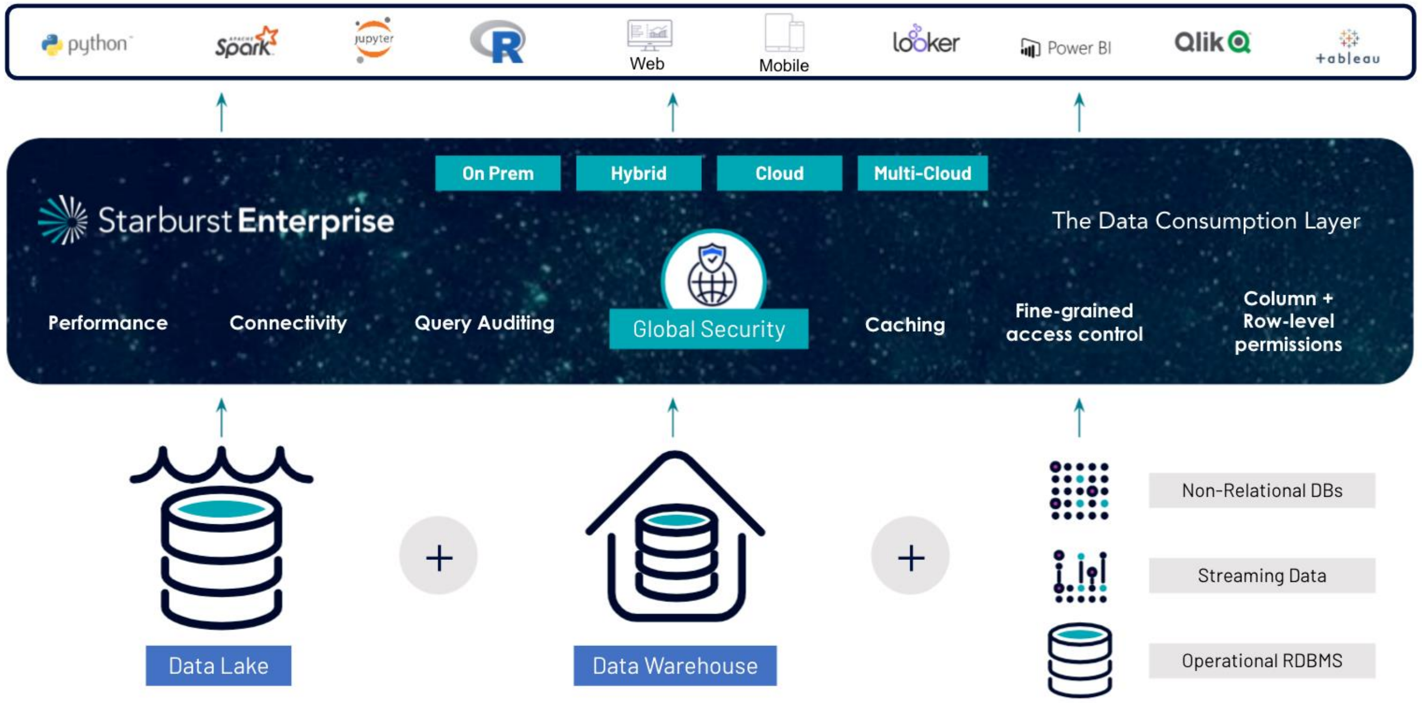


WHY STARBURST OVERVIEW

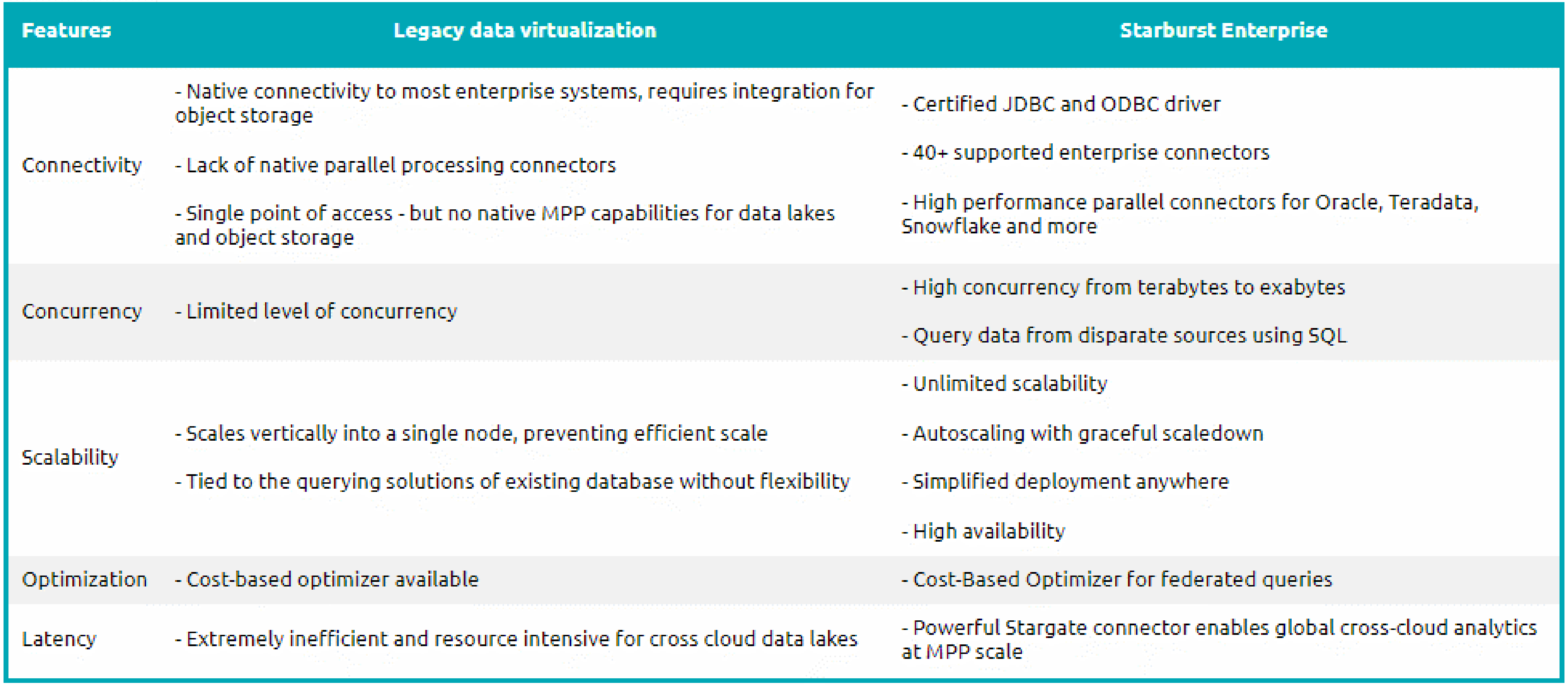
Advantages with Starburst:

- 10 – 100x faster query performance over other MPP engines
- 1/3rd the compute resources vs Hive, Spark & Impala
- Zero reliance on source data systems to perform joins but have the flexibility to pushdown where it makes sense to optimize performance.
- ANSI SQL standard no matter where the data originates
- Proven at 1000+ node and 100+PB scale
- Performant ground to cloud, multi-cloud, and multi-region analytics on data lakes with Starburst Stargate
- No vendor lock-in to underlying data sources. Provides storage optionality

Starburst Architecture



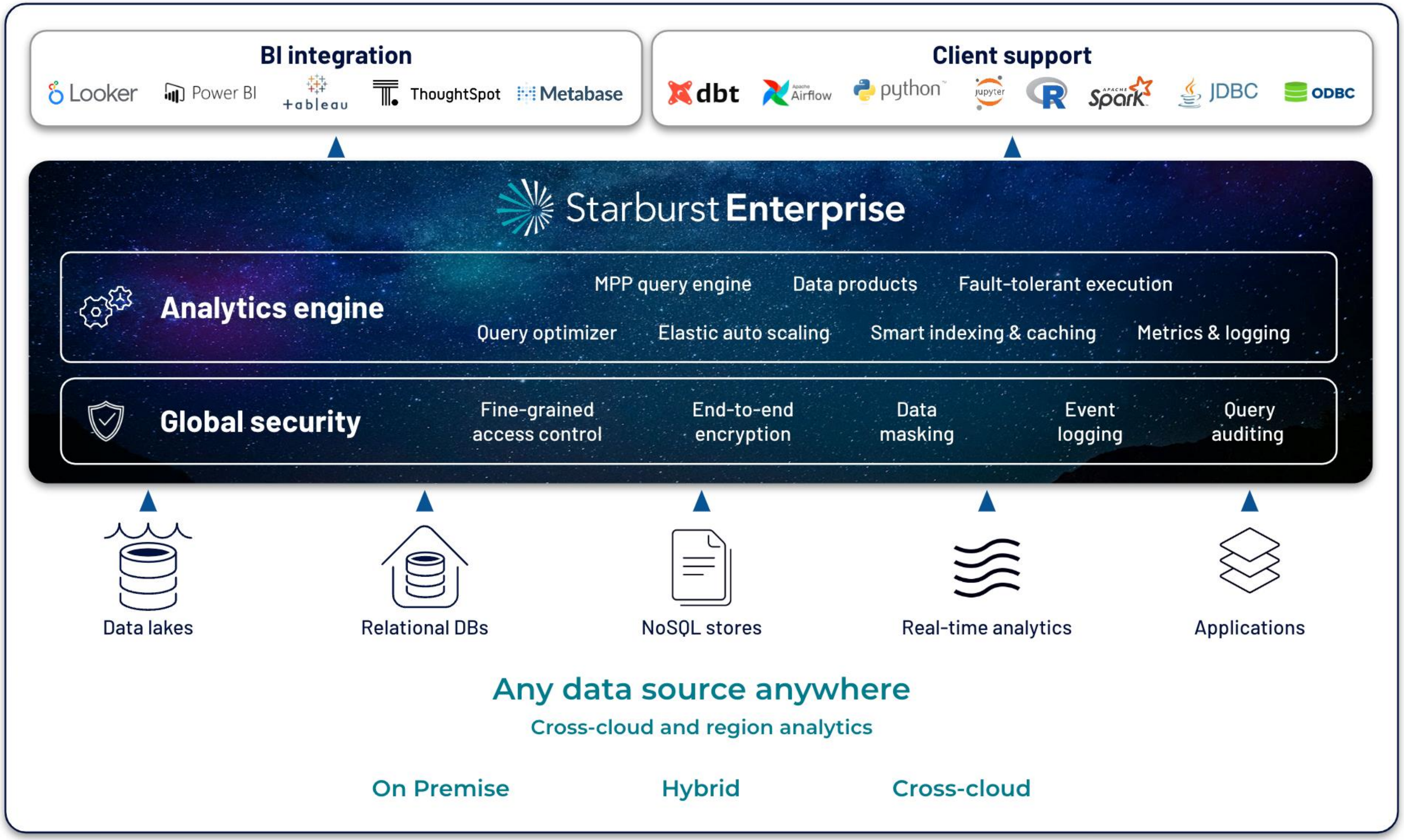
COMPARISON



Features	Legacy data virtualization	Starburst Enterprise
Connectivity	- Native connectivity to most enterprise systems, requires integration for object storage - Lack of native parallel processing connectors - Single point of access - but no native MPP capabilities for data lakes and object storage	- Certified JDBC and ODBC driver - 40+ supported enterprise connectors - High performance parallel connectors for Oracle, Teradata, Snowflake and more
Concurrency	- Limited level of concurrency	- High concurrency from terabytes to exabytes - Query data from disparate sources using SQL
Scalability	- Scales vertically into a single node, preventing efficient scale - Tied to the querying solutions of existing database without flexibility	- Unlimited scalability - Autoscaling with graceful scaledown - Simplified deployment anywhere - High availability
Optimization	- Cost-based optimizer available	- Cost-Based Optimizer for federated queries
Latency	- Extremely inefficient and resource intensive for cross cloud data lakes	- Powerful Stargate connector enables global cross-cloud analytics at MPP scale

STARBURST ENTERPRISE OVERVIEW

Starburst Enterprise is a **fully supported, production-tested and enterprise-grade distribution of open source Trino (formerly Presto® SQL)**. It improves performance and security while making it easy to deploy, connect, and manage your Trino environment.



BI integration

Looker, Power BI, tableau, ThoughtSpot, Metabase

Client support

dbt, Apache Airflow, python, jupyter, R, Apache Spark, JDBC, ODBC

Starburst Enterprise

Analytics engine

MPP query engine, Data products, Fault-tolerant execution, Query optimizer, Elastic auto scaling, Smart indexing & caching, Metrics & logging

Global security

Fine-grained access control, End-to-end encryption, Data masking, Event logging, Query auditing

Data lakes, Relational DBs, NoSQL stores, Real-time analytics, Applications

Any data source anywhere

Cross-cloud and region analytics

On Premise, Hybrid, Cross-cloud

Contact Us

☎ 080-4524-9465

✉ support@intellipaat.com