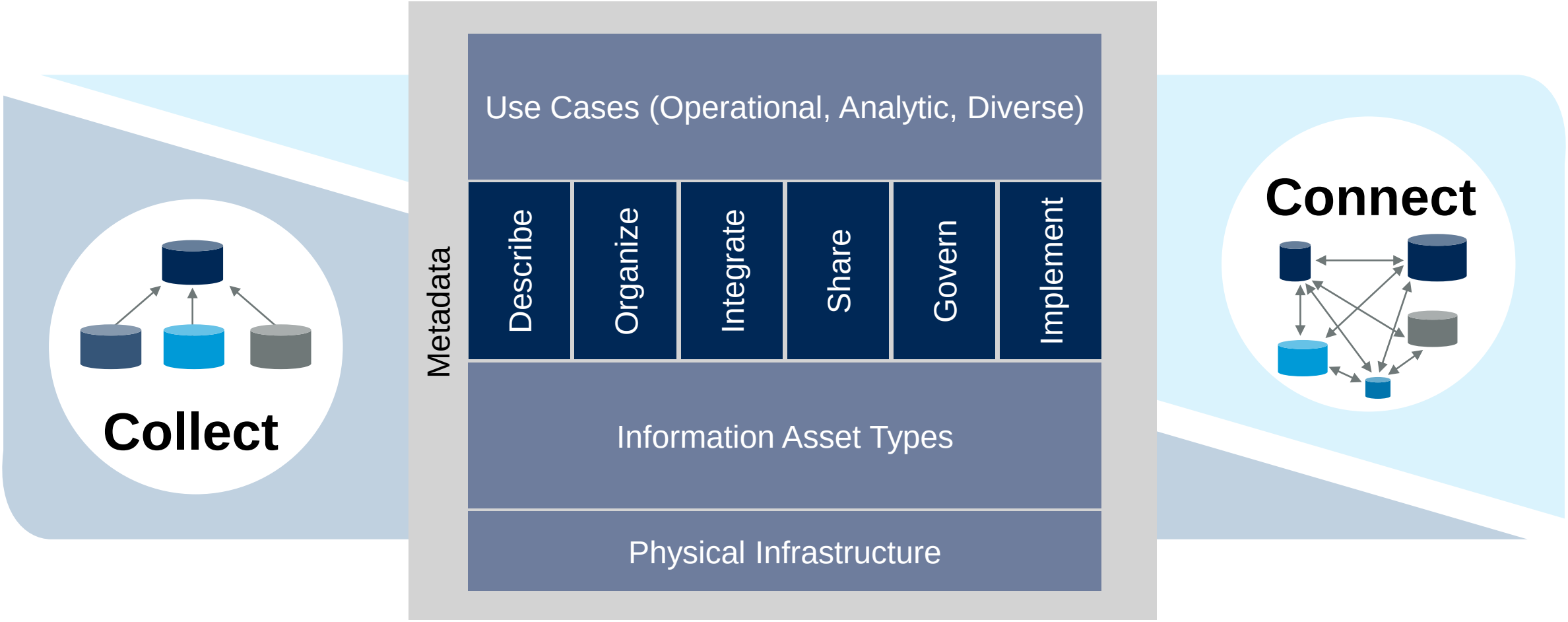
Gartner Data & Analytics Summit
18 – 21 March 2019 / Orlando, Florida

Data Hubs, Lakes and Warehouses: Choosing the Core of Your Data and Analytics Platform

Rick Greenwald

Gartner

When to Collect — And Where? When to Connect — And How?



Gartner

Key Issues

1. What are the differences between hubs, lakes and warehouses?
2. How do you balance the trade-offs between these options?
3. What are the technology options and how are they integrated?

Gartner

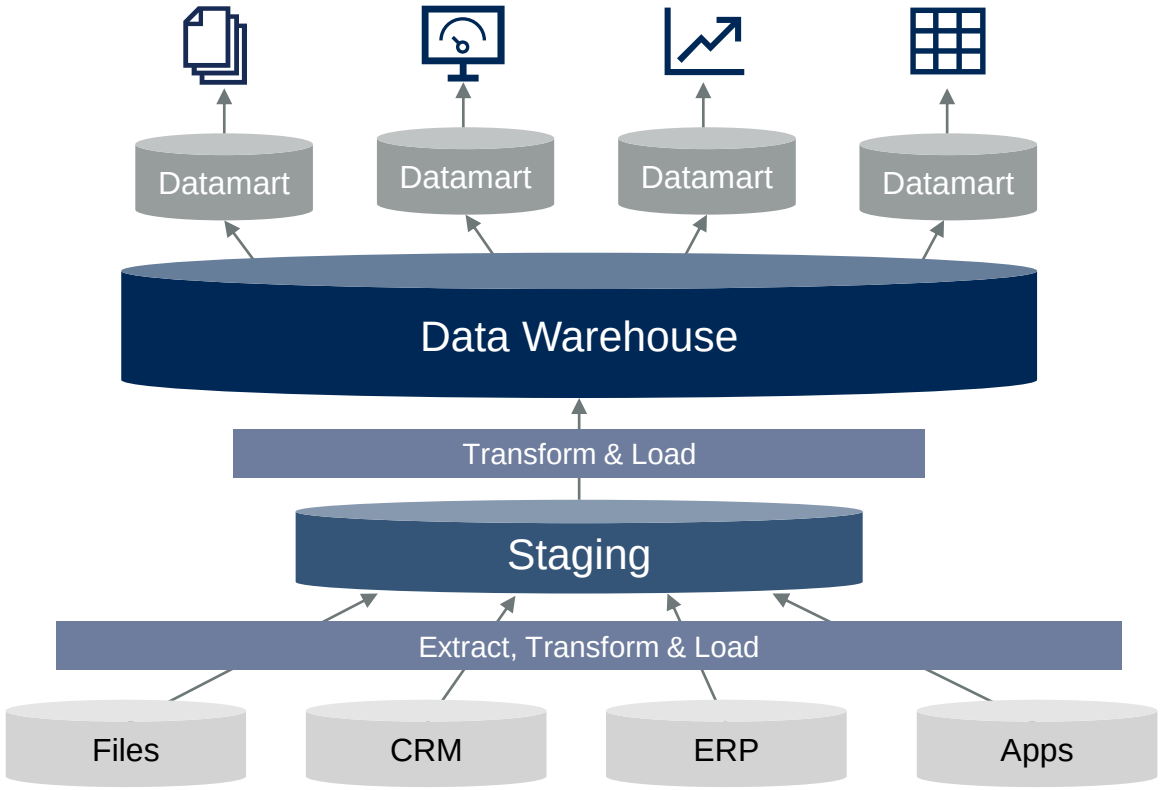
Key Issues

1. What are the differences between hubs, lakes and warehouses?
2. How do you balance the trade-offs between these options?
3. What are the technology options and how are they integrated?

Gartner

The Data Warehouse, Circa 1995

- Provides 80% of analytics using the same 20% of available data
- Optimized for repeatable processes
- Supports hundreds of enterprise consumers

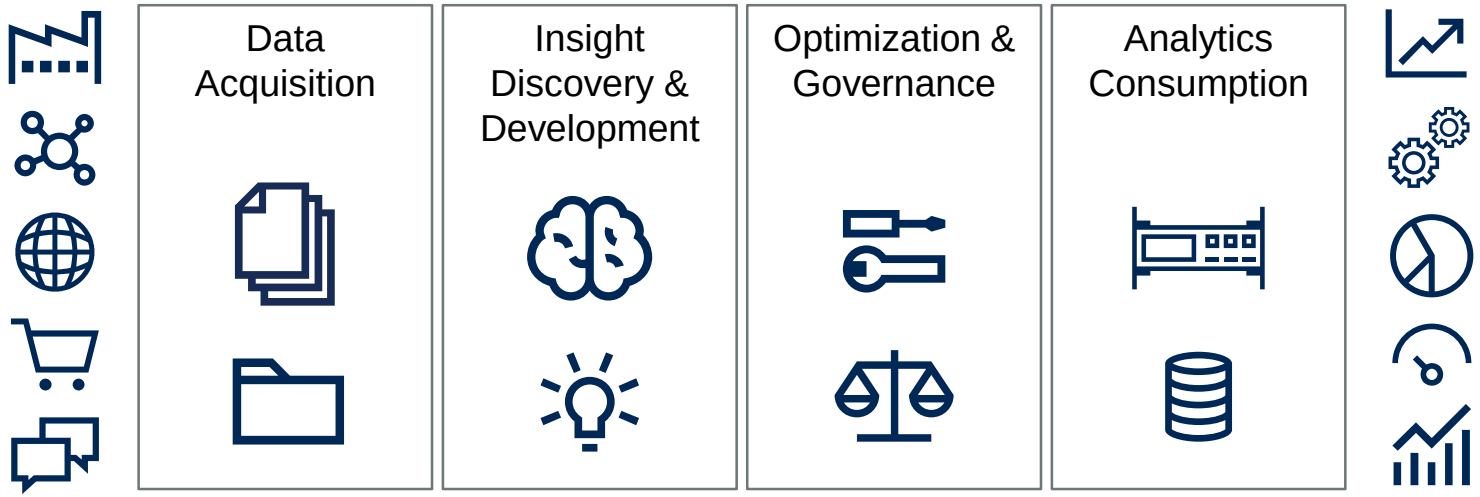


How can we ask enterprisewide questions requiring historical perspective?

Gartner

Data Lakes for Analytics Discovery

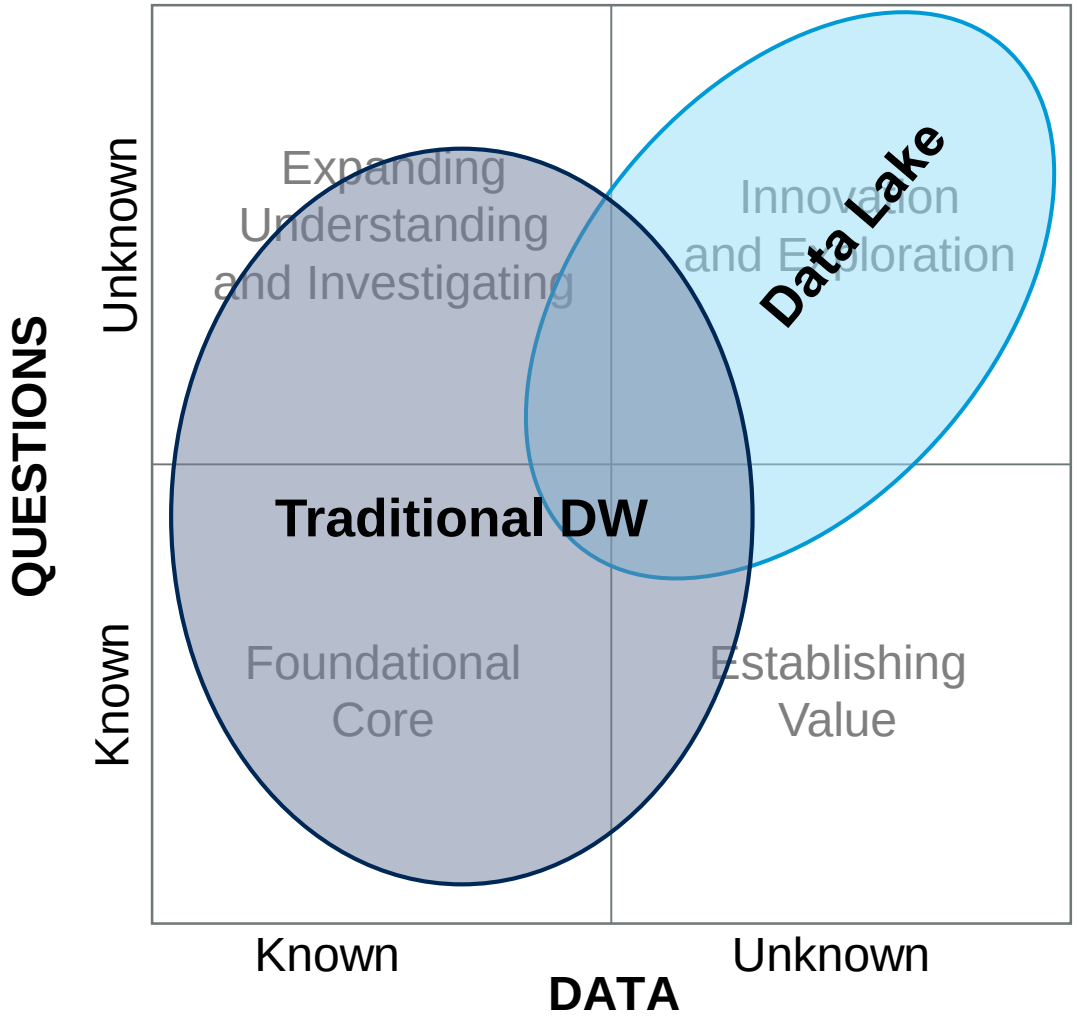
- Outgrowth of the DW staging area
- Stores raw data for exploration, analysis
- Not for everyone and every use case



How can we figure out what we don't know?

Gartner

How Do Lakes and Warehouses Relate?



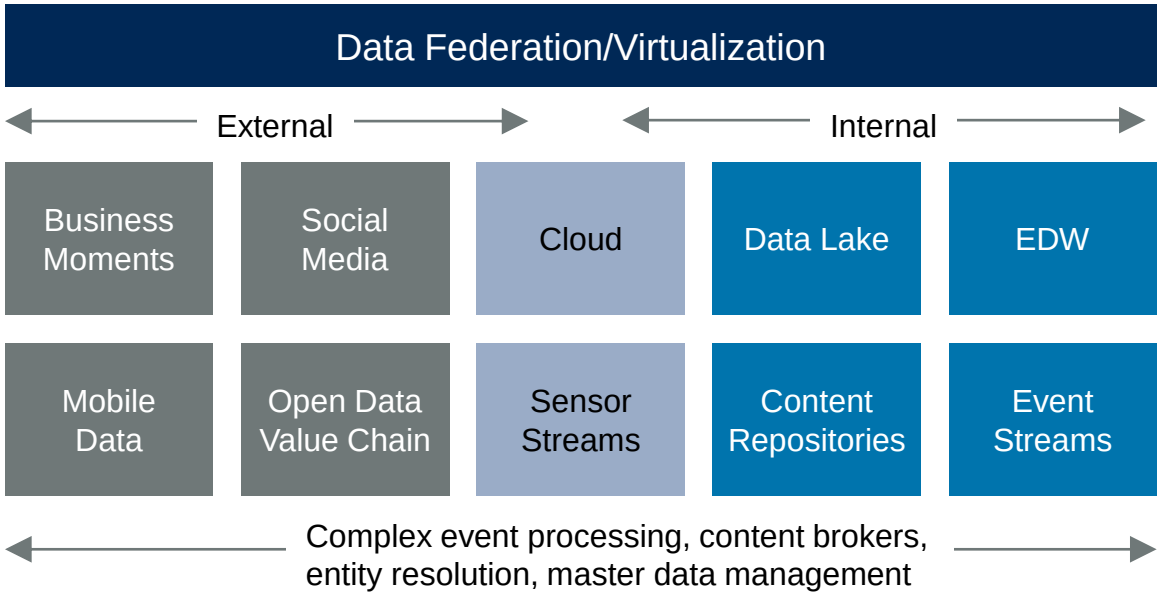
Context-Independent
Data Warehouse

Logical data warehouse overlays the **whole** chart

Gartner

Workload and Data Expansion With the Logical Data Warehouse

- Need to support the remaining 20% of analytics
- Diverse users with diverse skills and tools

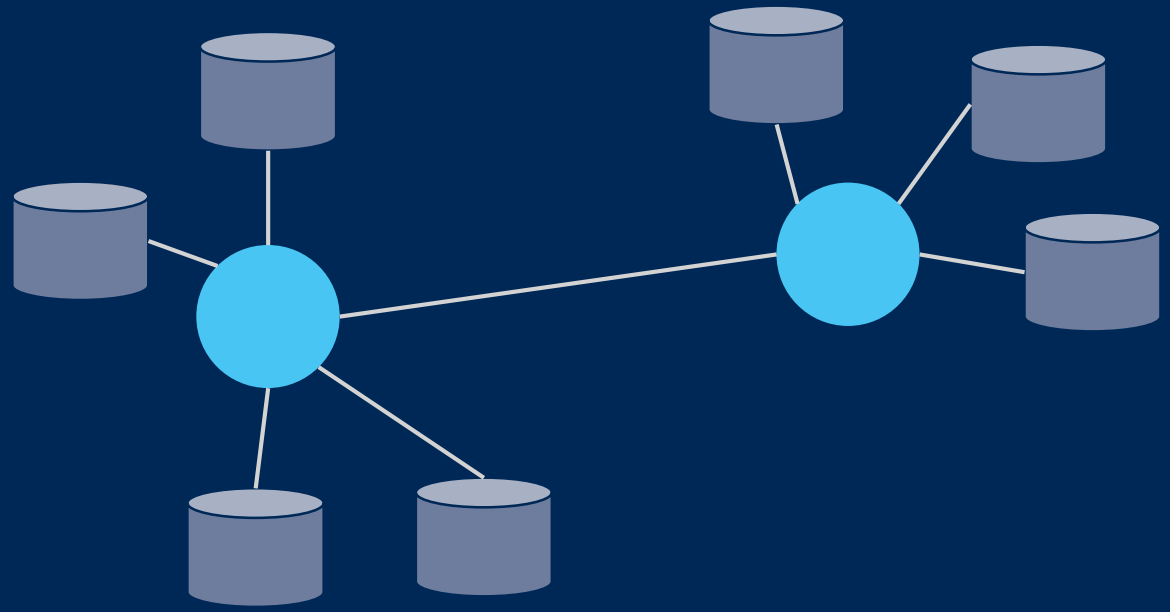


How can we expand our data management and analysis to more data types for different contexts?

Gartner

Data Hubs for Mediating Governance, Sharing and Integration

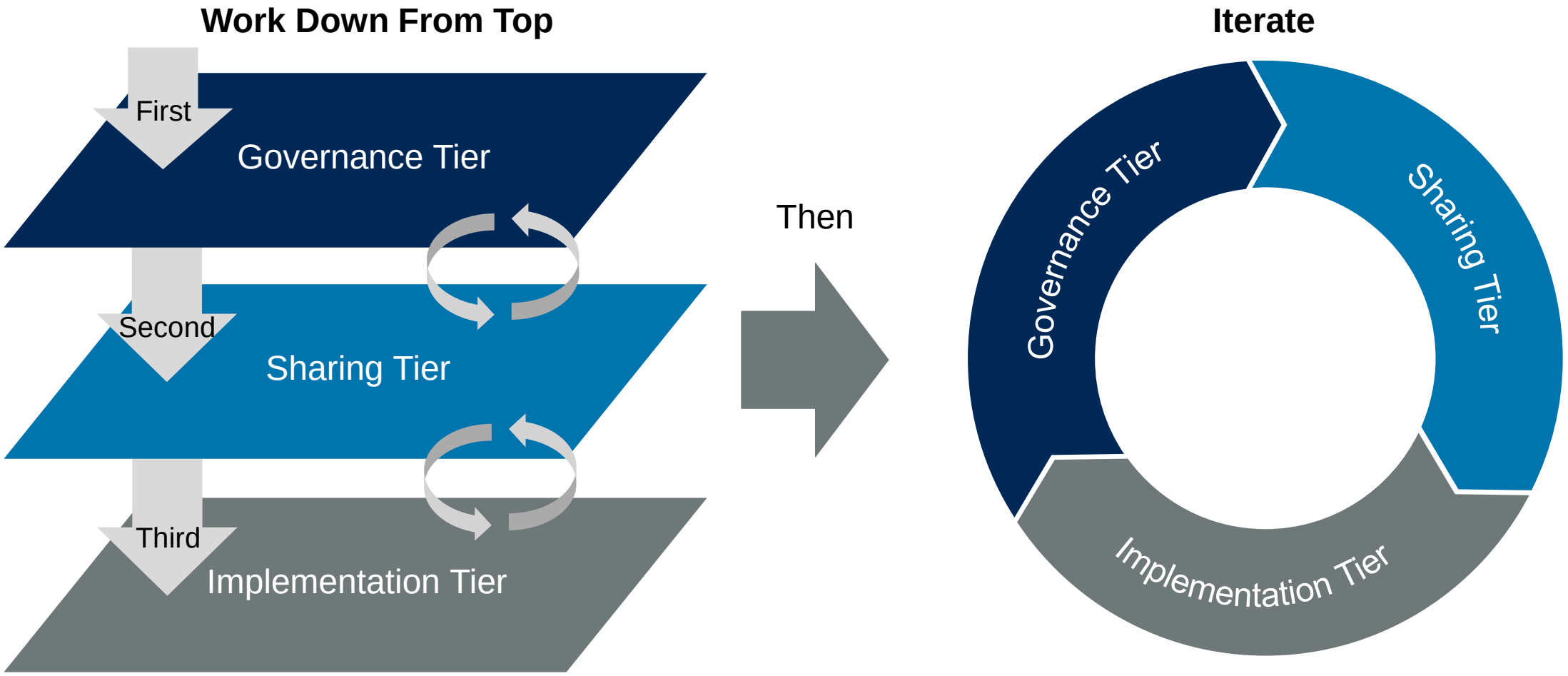
- Use cases:
  - Mediation and sharing of datasets:
    - Metadata focused
  - Distributed governance/ policy enforcement
  - Operationally focused but can be a trusted analytical data source



Determines effective mediation of semantics, and efficient data integration strategies, across applications, IoT networks, enterprises and ecosystems

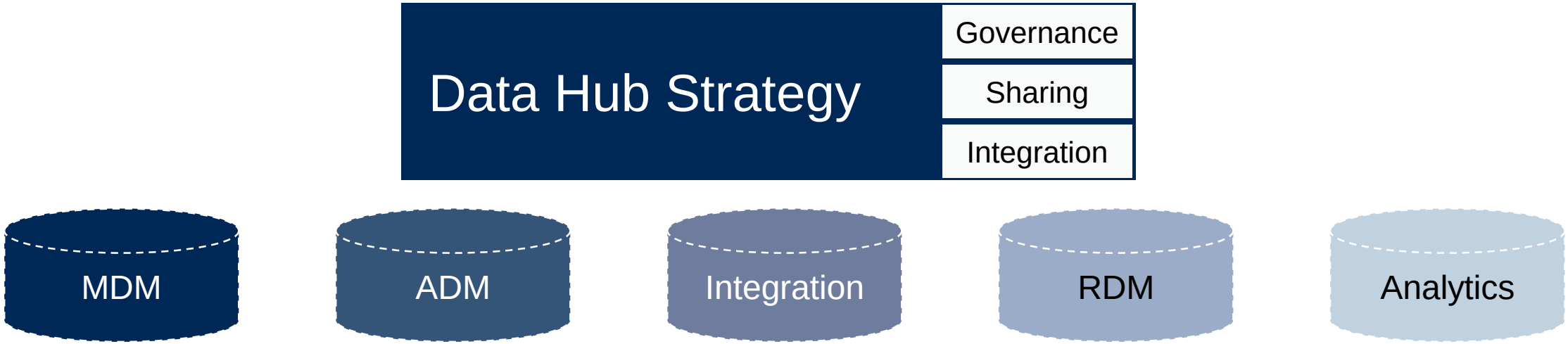
Gartner

The Elements of a Data Hub Strategy



Gartner

Types of Data Hubs



Key Issues

1. What are the differences between hubs, lakes and warehouses?
2. How do you balance the trade-offs between these options?
3. What are the technology options and how are they integrated?

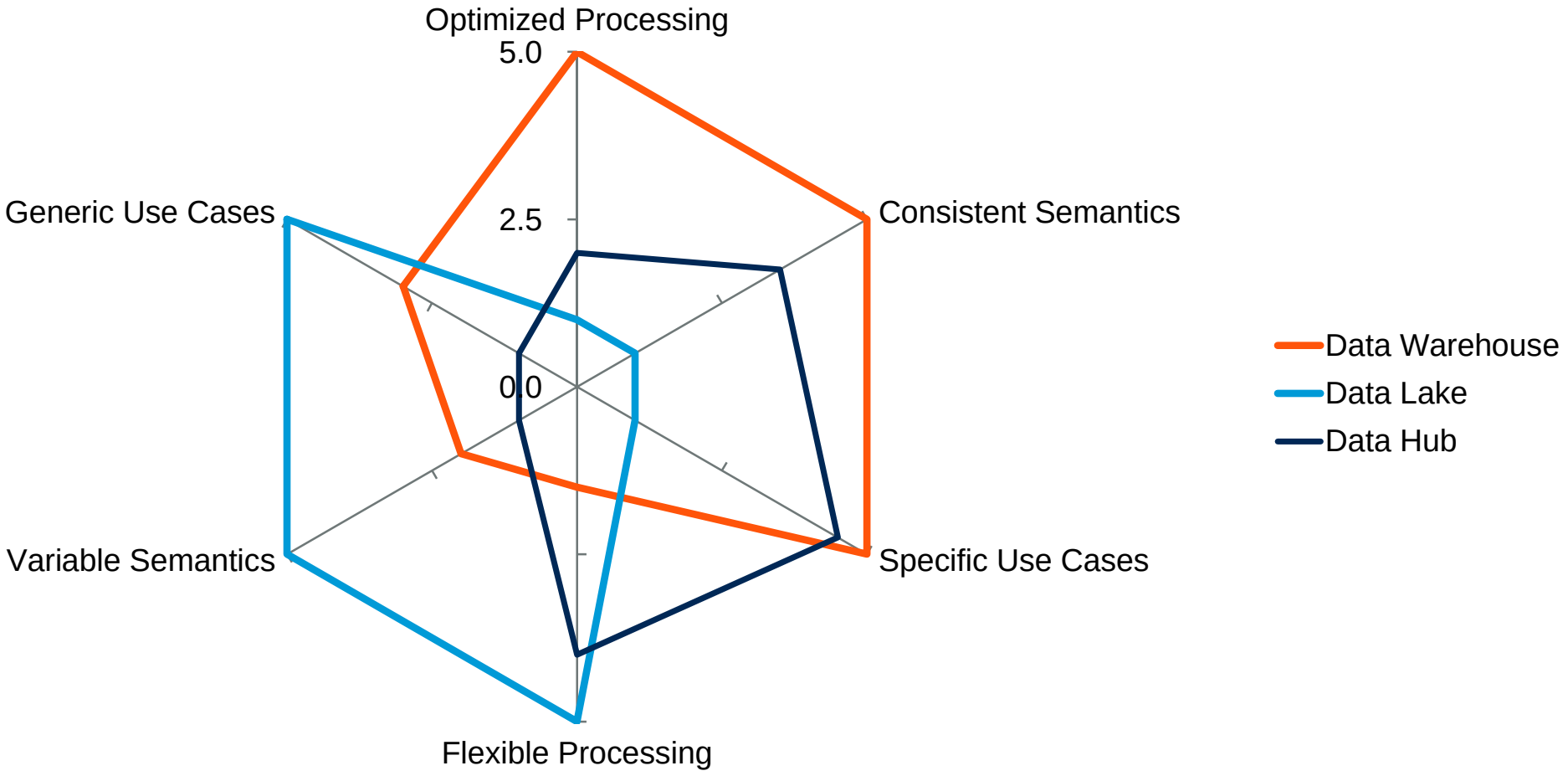
Gartner

How Do Hubs, Lakes and Warehouses Differ?

- What data processing options are available?
  - Flexible or optimized?
- What semantic capabilities are offered?
  - Variable or consistent?
- What types of use cases can I address?
  - Generic or specific?

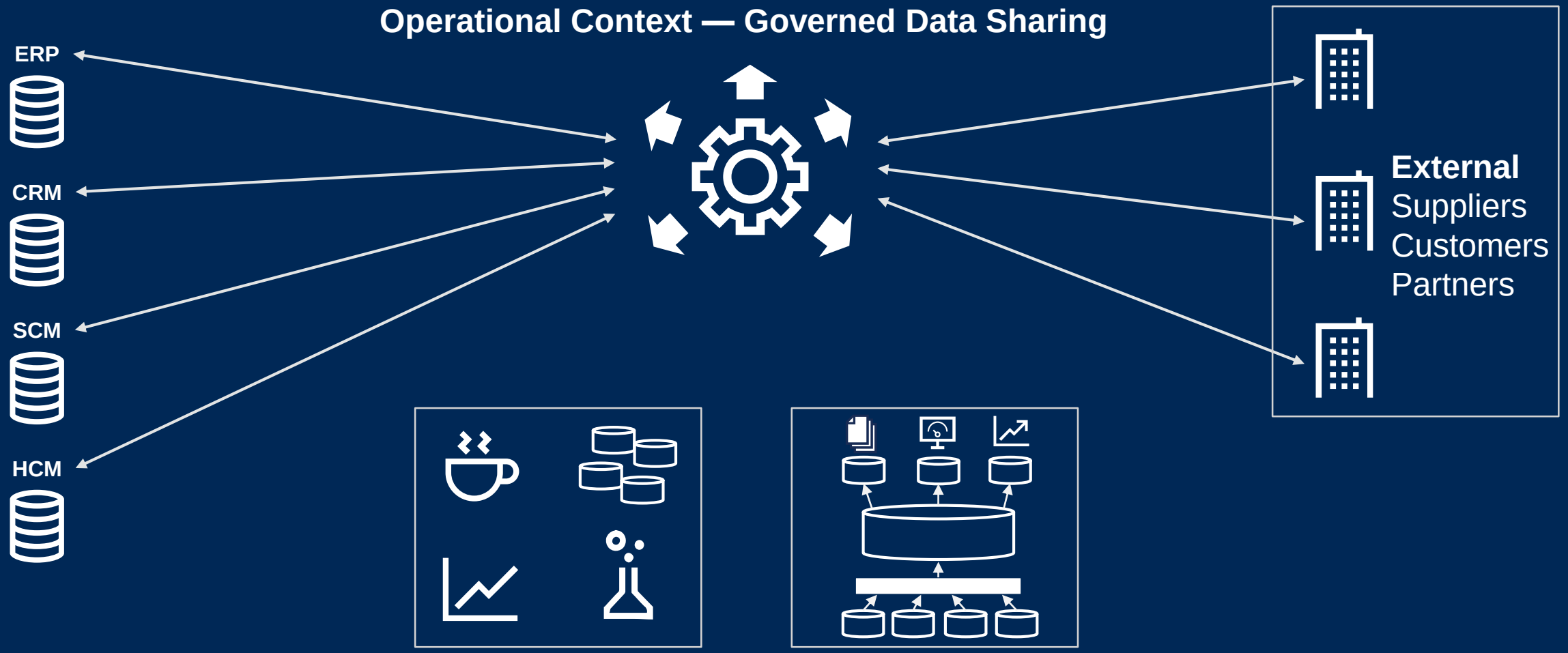
Gartner

How Do Hubs, Lakes and Warehouses Differ?



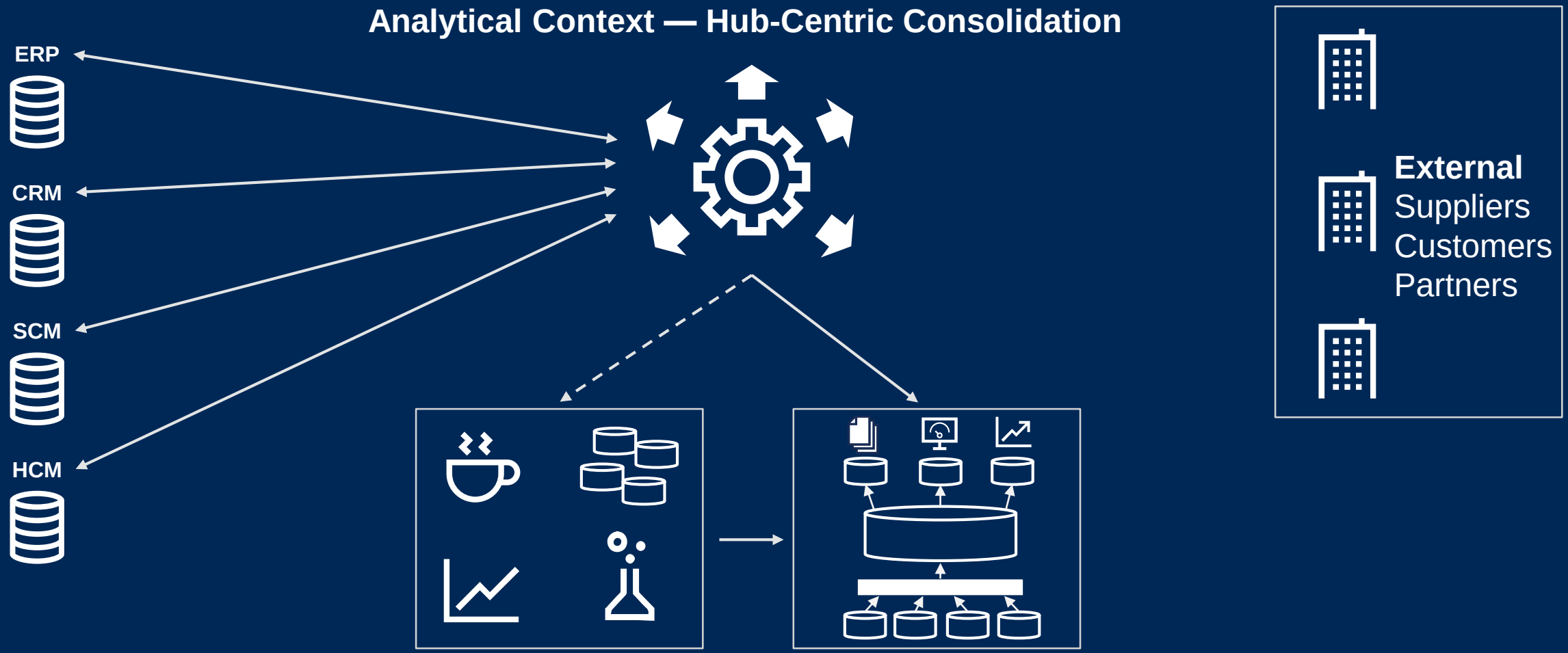
Gartner

Hubs, Lakes and Warehouses Aren't Exclusive Choices



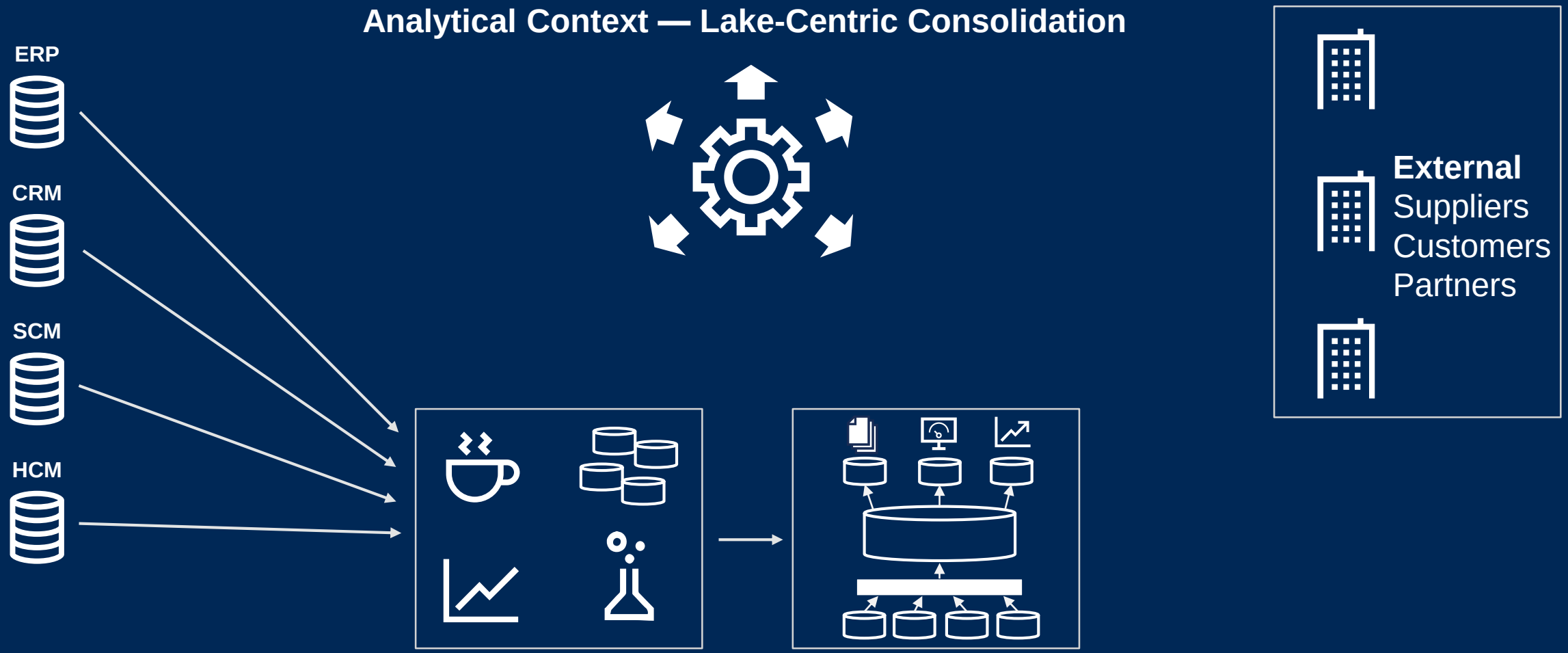
Gartner

Hubs, Lakes and Warehouses Aren't Exclusive Choices



Gartner

Hubs, Lakes and Warehouses Aren't Exclusive Choices



Gartner

Do You Need All of These Options?

Strategic Planning Assumption

By 2021, enterprises using a cohesive strategy incorporating data hubs, lakes and warehouses will support 30% more use cases than competitors.

Gartner

Key Issues

1. What are the differences between hubs, lakes and warehouses?
2. How do you balance the trade-offs between these options?
3. What are the technology options and how are they integrated?

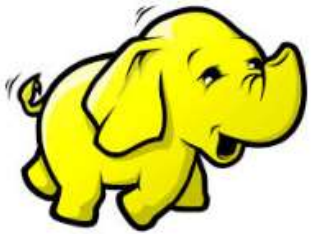
Gartner

Data Warehousing Choices Proliferate

- Continued adoption of cloud offerings:
  - Alibaba Cloud, Amazon Web Services, Google Cloud Platform, IBM, Microsoft, Oracle, Qubole, Snowflake
- Hybrid data warehousing becoming viable as incumbents lead shift:
  - IBM, Micro Focus, Microsoft, Oracle, Teradata
- Insurgent vendors filling specialized roles:
  - Cloudera-Hortonworks, MapR Technologies, MarkLogic, MemSQL, Neo4j, Treasure Data

Gartner

Data Lake Implementation Technologies

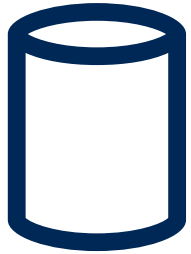


Apache Hadoop distributions:

- Simplified data ingestion and storage with several processing options
- Data lake management ecosystem emerging
- Complex deployment and management

Cloud-based block and object stores:

- Simplified data ingestion and storage
- Bring your own processing
- Nascent management and security ecosystem



Database management systems:

- Optimal for certain data types and formats
- Data processing options expanding beyond SQL
- Scaling and cost may be challenges

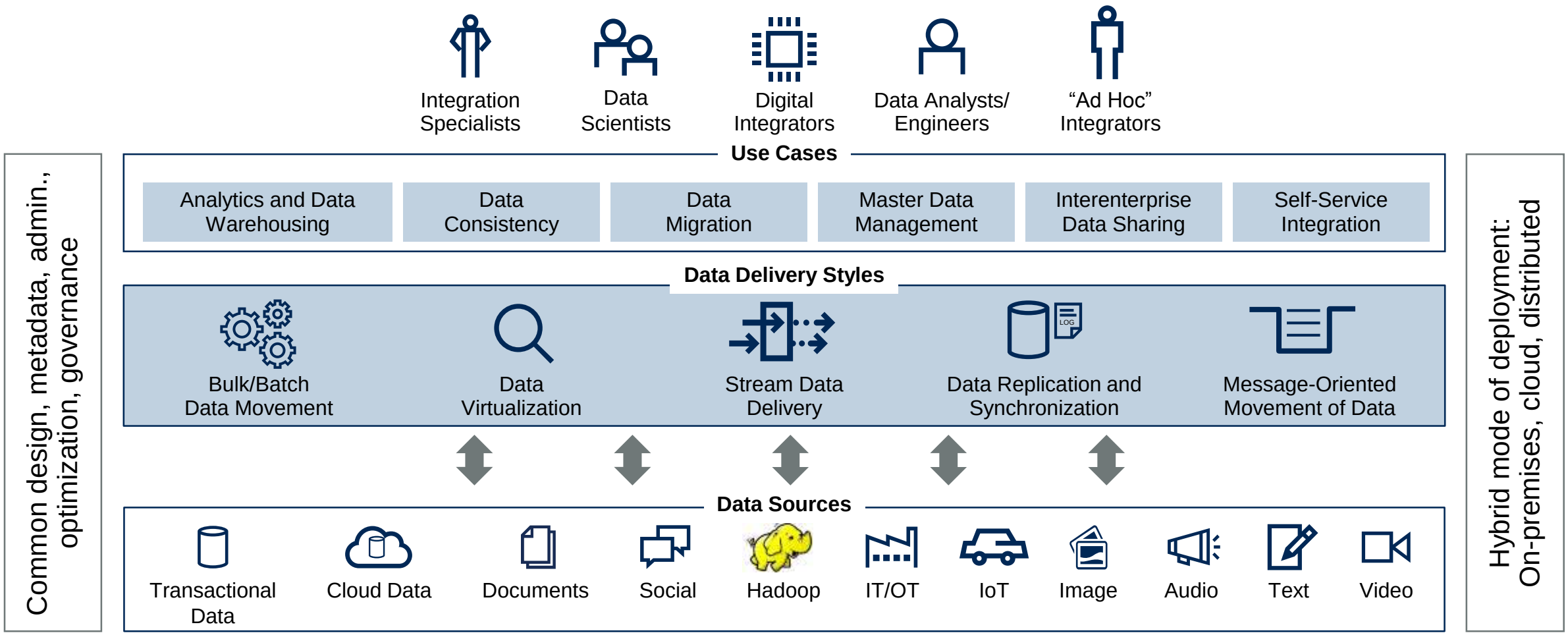
Gartner

Data Hub Technologies and Tools

- Data integration tools (ETL, replication, data virtualization).
- Application integration middleware (ESB, MOM, iPaaS, API management).
- Persistence technologies (DBMS, Hadoop, cloud-based data stores).
- Governance (data quality tools, data privacy technology, MDM solutions).
- Metadata management platforms.
- All the above, packaged as a “hub product”?

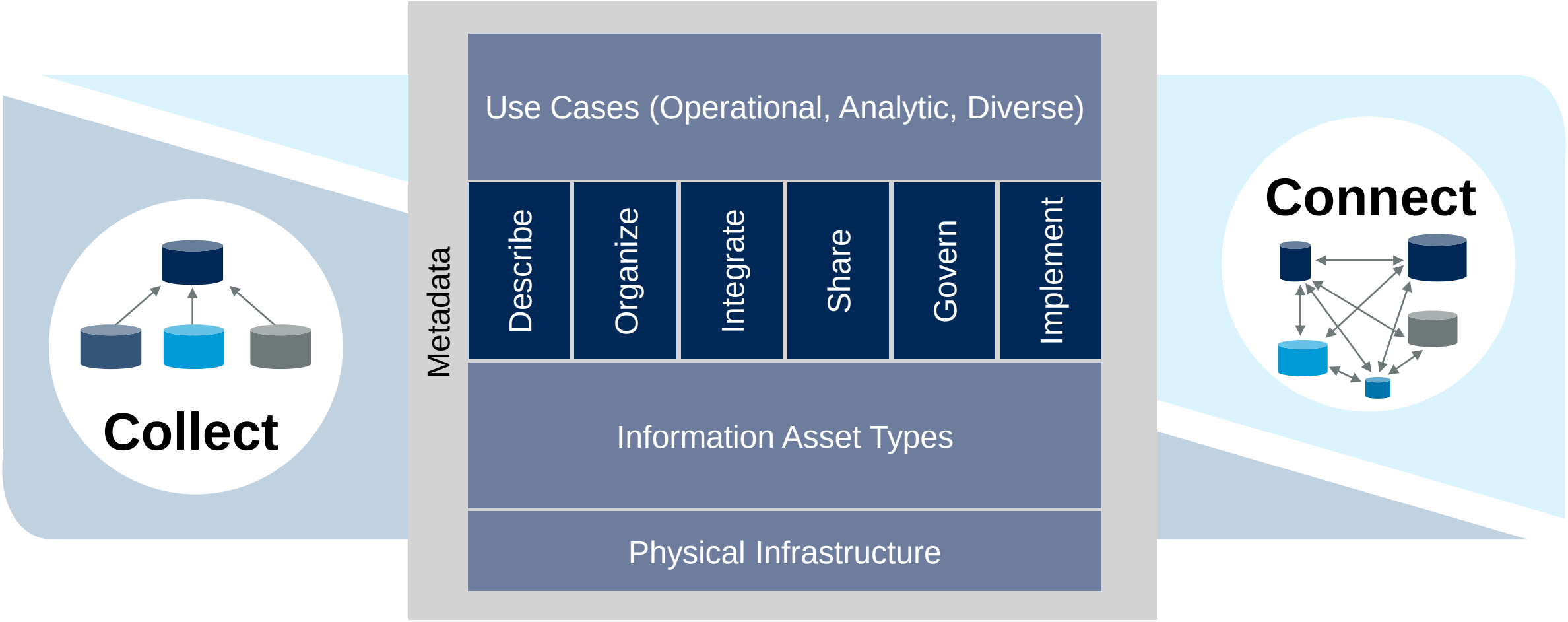
Gartner

A Range of Integration Styles to Support a Range of Patterns and Connection Types



Gartner

Apply the Right Combination of Lakes, Warehouses and Hubs to Best Enable Data Sharing and Analytics



Gartner

Recommendations

- ✓ Build the core of your digital platform based on the types of use cases, processing flexibility and semantic enablement your users require.
- ✓ Apply the data hub architecture to better balance the ability to collect data with connecting data producers and consumers as needed.
- ✓ Use data lakes for analytics exploration and data warehouses for optimization and broad consumption.
- ✓ Prepare for continuous platform evolution as business needs change.

Gartner

Recommended Gartner Research

- **Use a Data Hub Strategy to Meet Your Data and Analytics Governance and Sharing Requirements**
  Andrew White and Ted Friedman (G00295309)
- **Implementing the Data Hub: Architecture and Technology Choices**
  Ted Friedman and Andrew White (G00297674)
- **Best Practices for Designing Your Data Lake**
  Nick Heudecker (G00315546)
- **Data Management Solutions for Analytics: Current and Future States, 2017**
  Rick Greenwald and Adam Ronthal (G00336273)

For information, please contact your Gartner representative.

Gartner