Gartner Data & Analytics Summit
Summit 2018

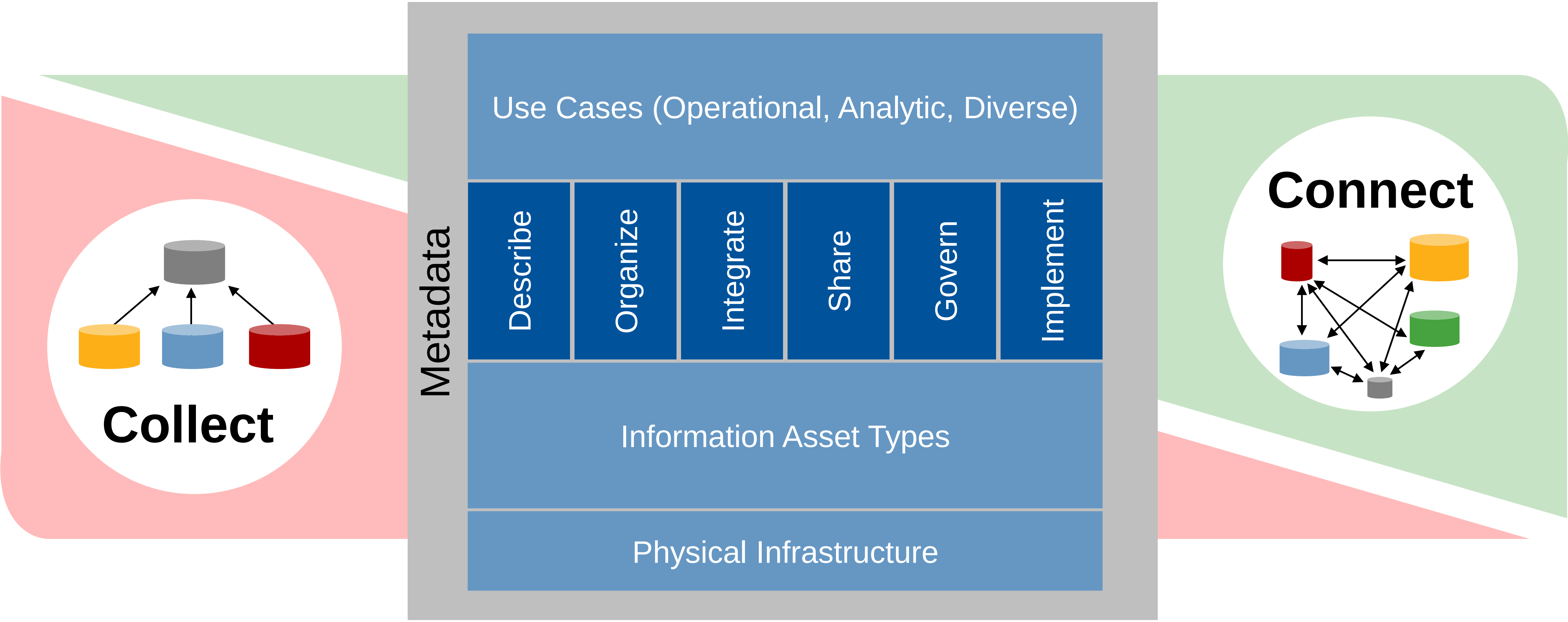
22 – 23 May 2018 / São Paulo, Brazil

Gartner
Summits

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

Nick Heudecker

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



Collect

Metadata

Use Cases (Operational, Analytic, Diverse)

Describe

Organize

Integrate

Share

Govern

Implement

Information Asset Types

Physical Infrastructure

Connect

Gartner

Key Issues

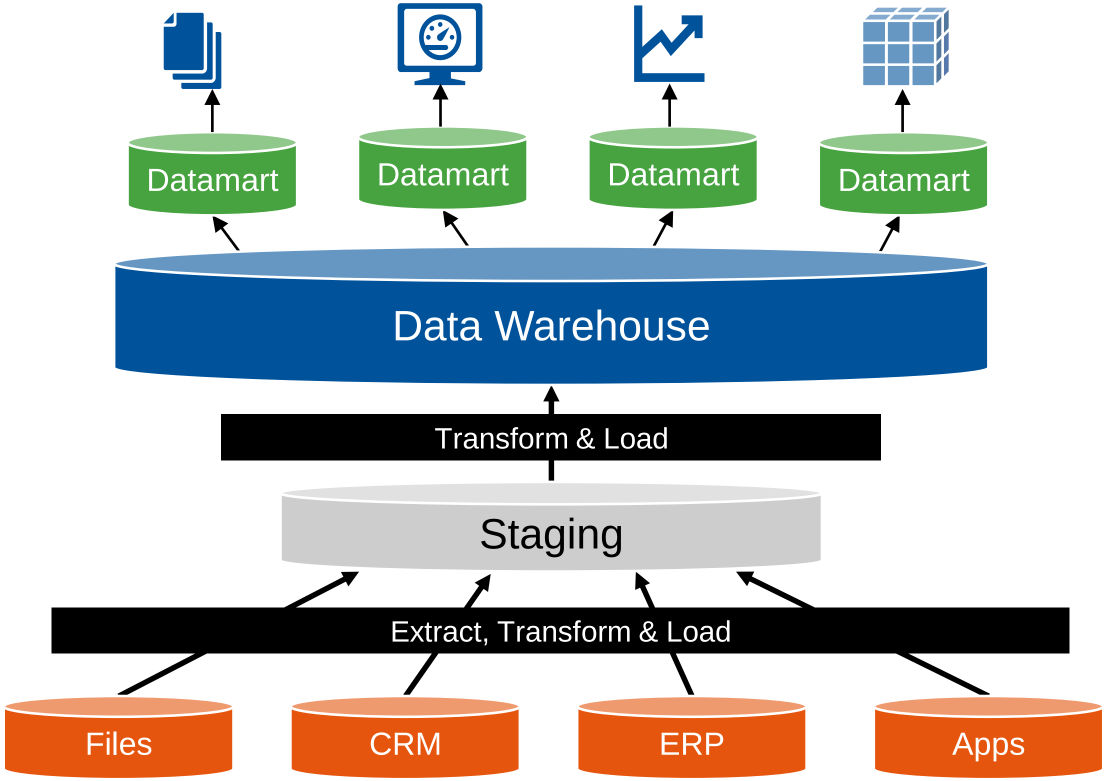
1. What are the differences between hubs, lakes and warehouses?
2. How do you balance the trade-offs between these options?
3. What are the technology options and how are they integrated?

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?

The Data Warehouse, Circa 1995

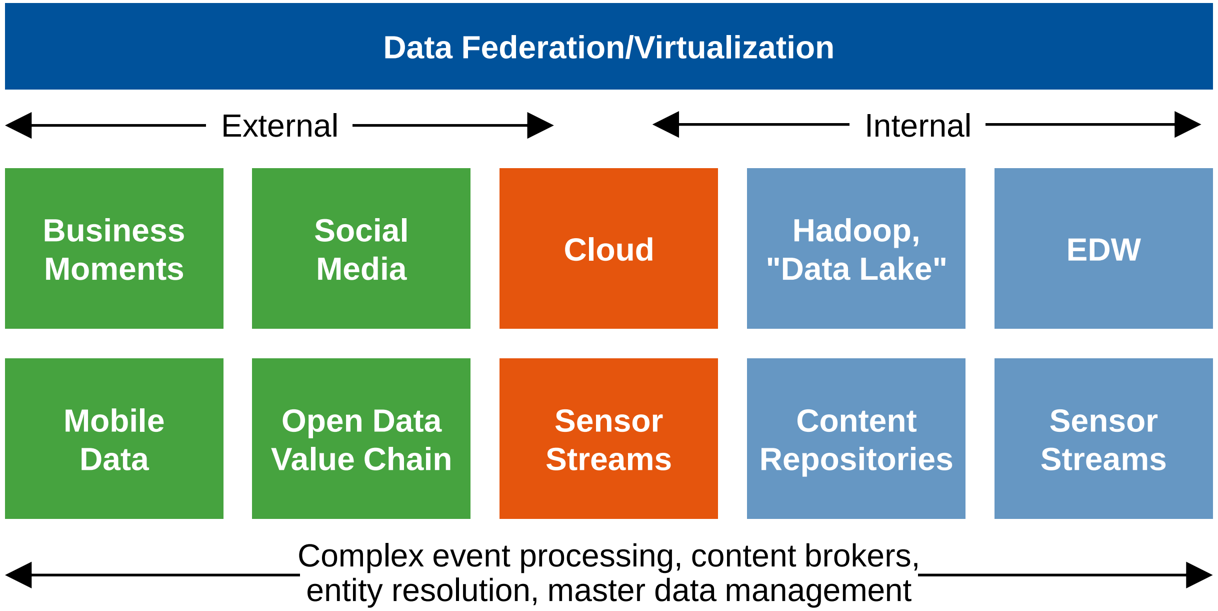
- Provides 80% of analytics with 20% of the data
- Optimized for repeatable processes
- Supports hundreds of enterprise consumers



How can we ask enterprisewide questions requiring historical perspective?

Workload and Data Expansion With the Logical Data Warehouse

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



Data Federation/Virtualization

External

Internal

Business Moments

Social Media

Cloud

Hadoop, "Data Lake"

EDW

Mobile Data

Open Data Value Chain

Sensor Streams

Content Repositories

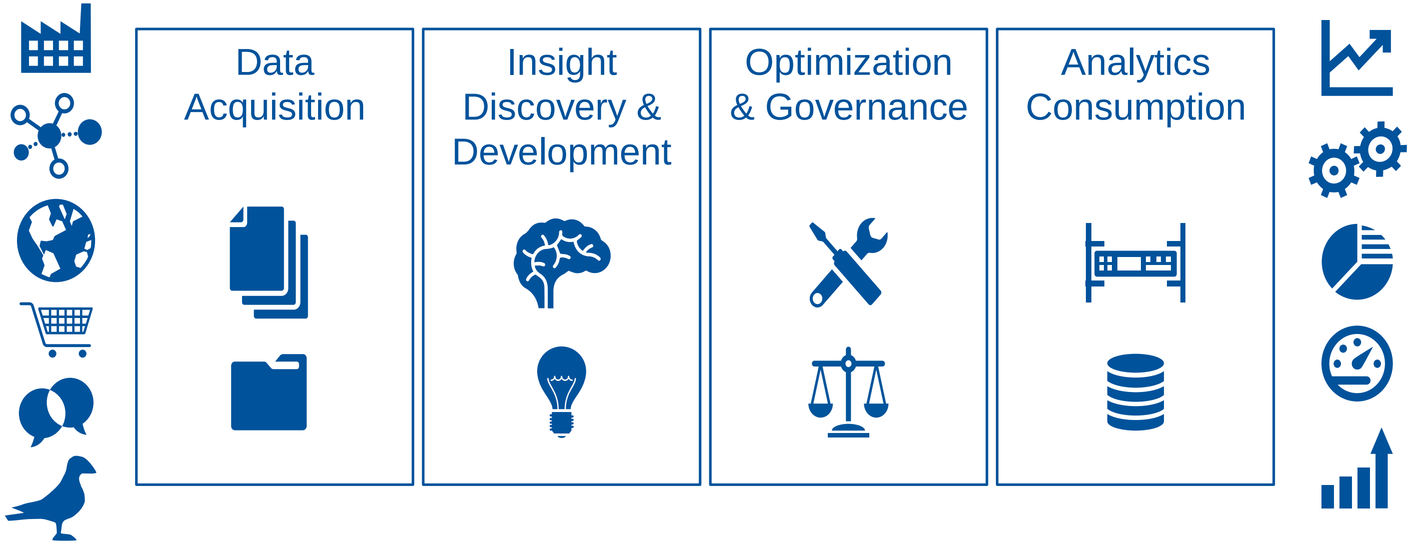
Sensor Streams

Complex event processing, content brokers, entity resolution, master data management

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

Data Lakes for Analytics Discovery

- Outgrowth of the DW staging area
- Stores raw data for exploration, analysis
- Optimization still required for broad consumption



Data Acquisition

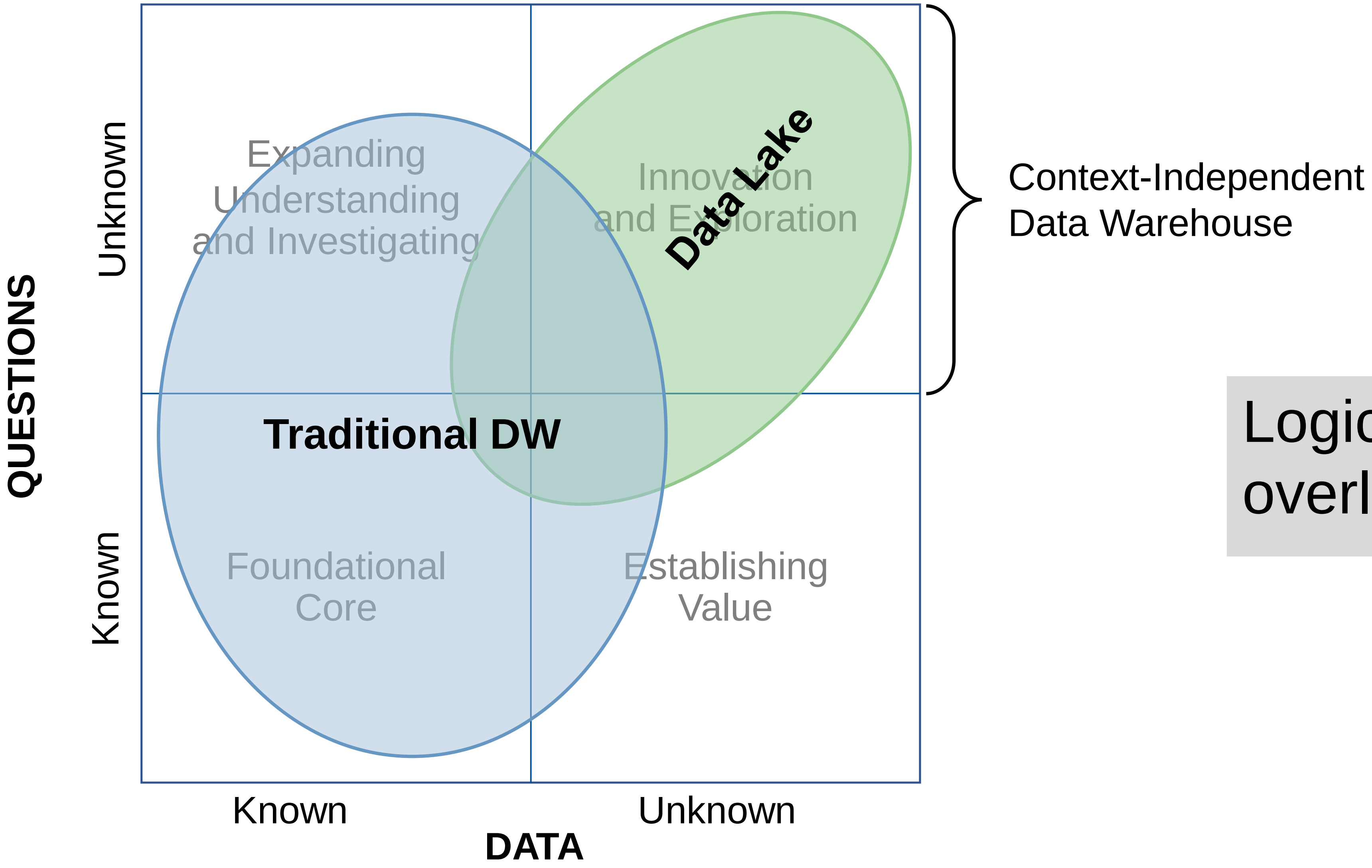
Insight Discovery & Development

Optimization & Governance

Analytics Consumption

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

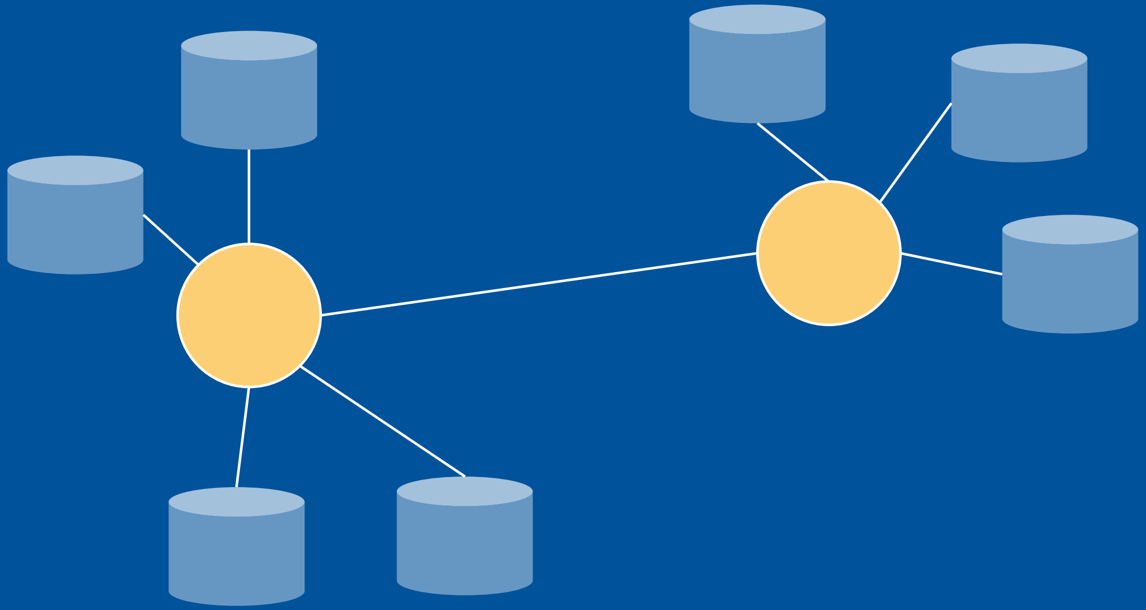
How Do Lakes and Warehouses Relate?


QUESTIONS
Unknown
Known
Expanding Understanding and Investigating
Innovation and Exploration
Data Lake
Traditional DW
Foundational Core
Establishing Value
Known
Unknown
DATA
Context-Independent Data Warehouse

Logical data warehouse overlays the **whole** chart

Data Hubs for Semantic Mediation and Integration

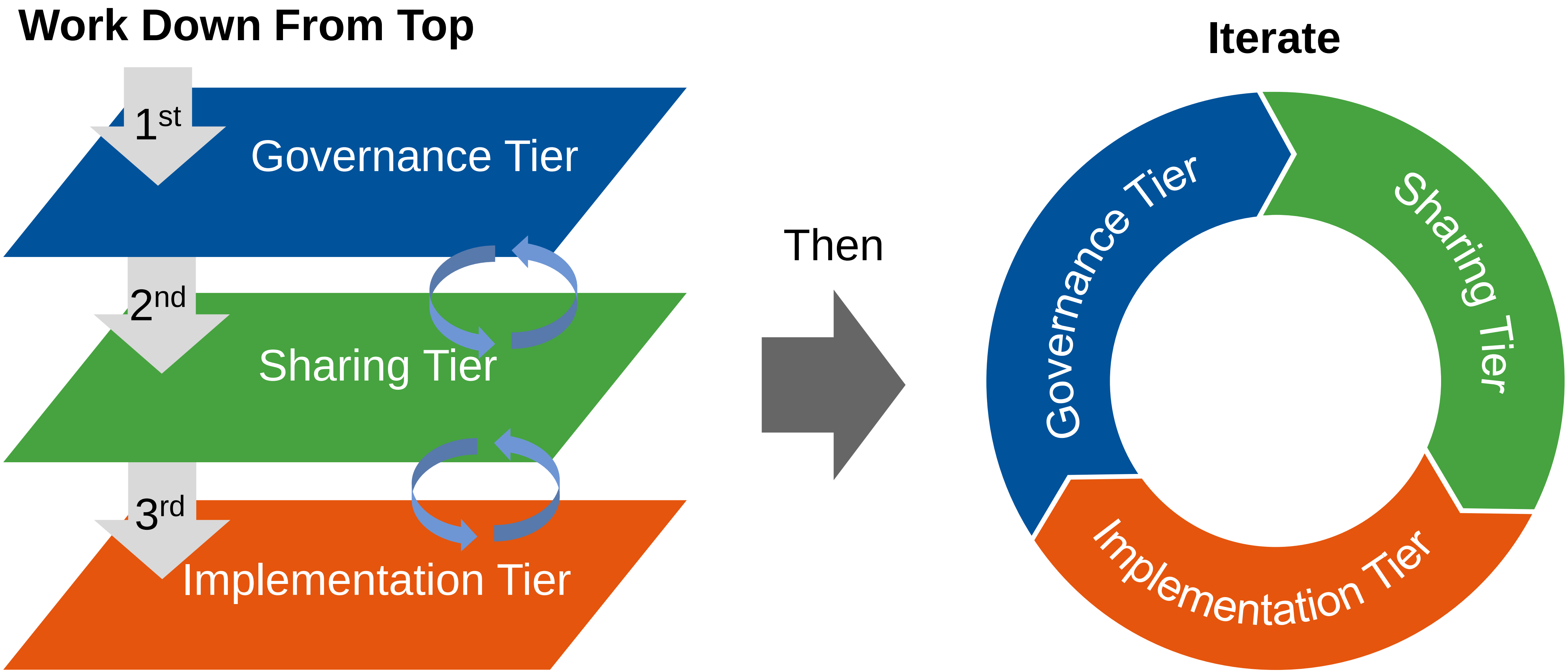
- Use cases:
  - Mediation and sharing of datasets
  - Distributed governance/policy enforcement
  - Operationally focused



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



Work Down From Top

1st

Governance Tier

2nd

Sharing Tier

3rd

Implementation Tier

Then

Iterate

Governance Tier

Sharing Tier

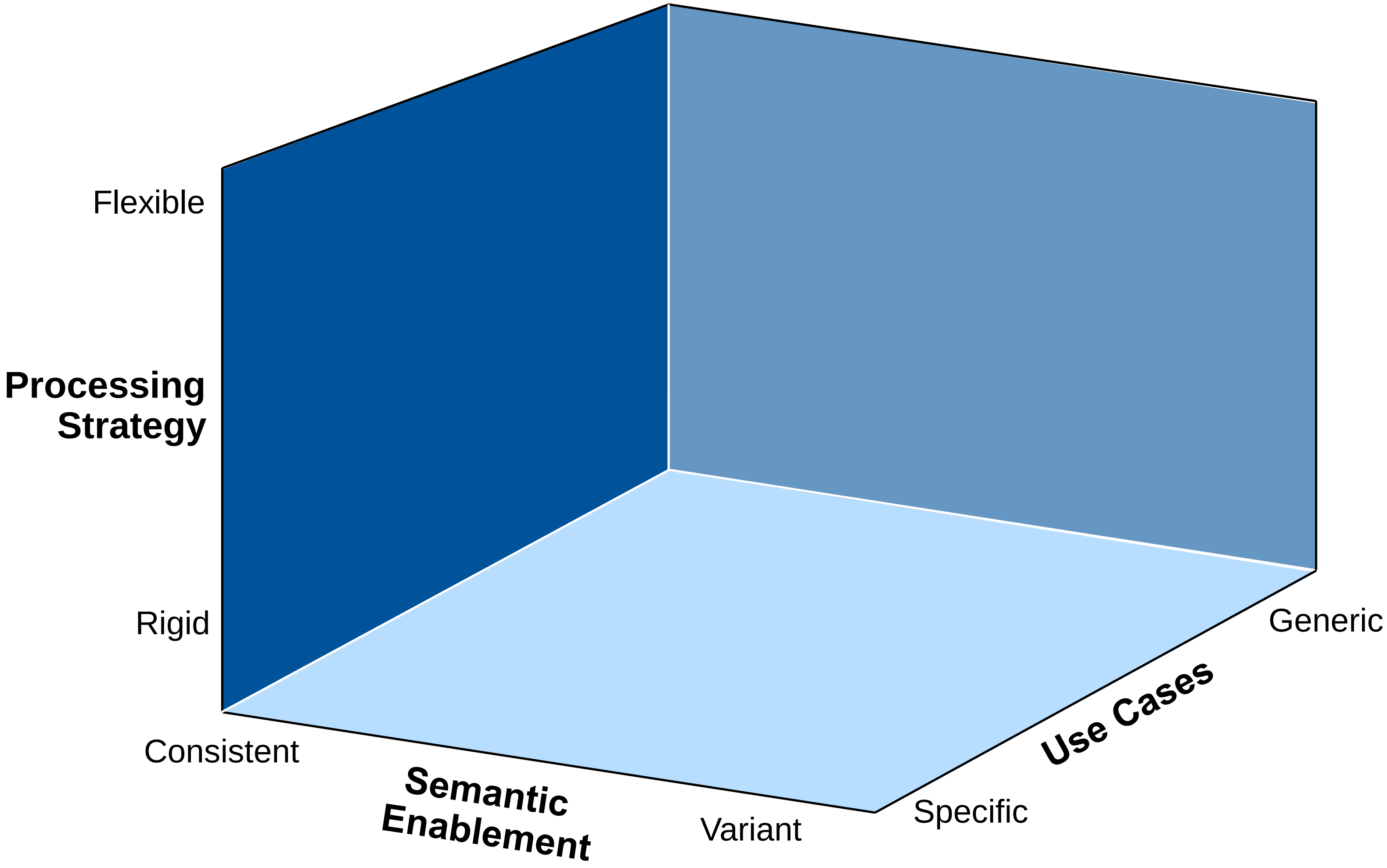
Implementation Tier

Gartner

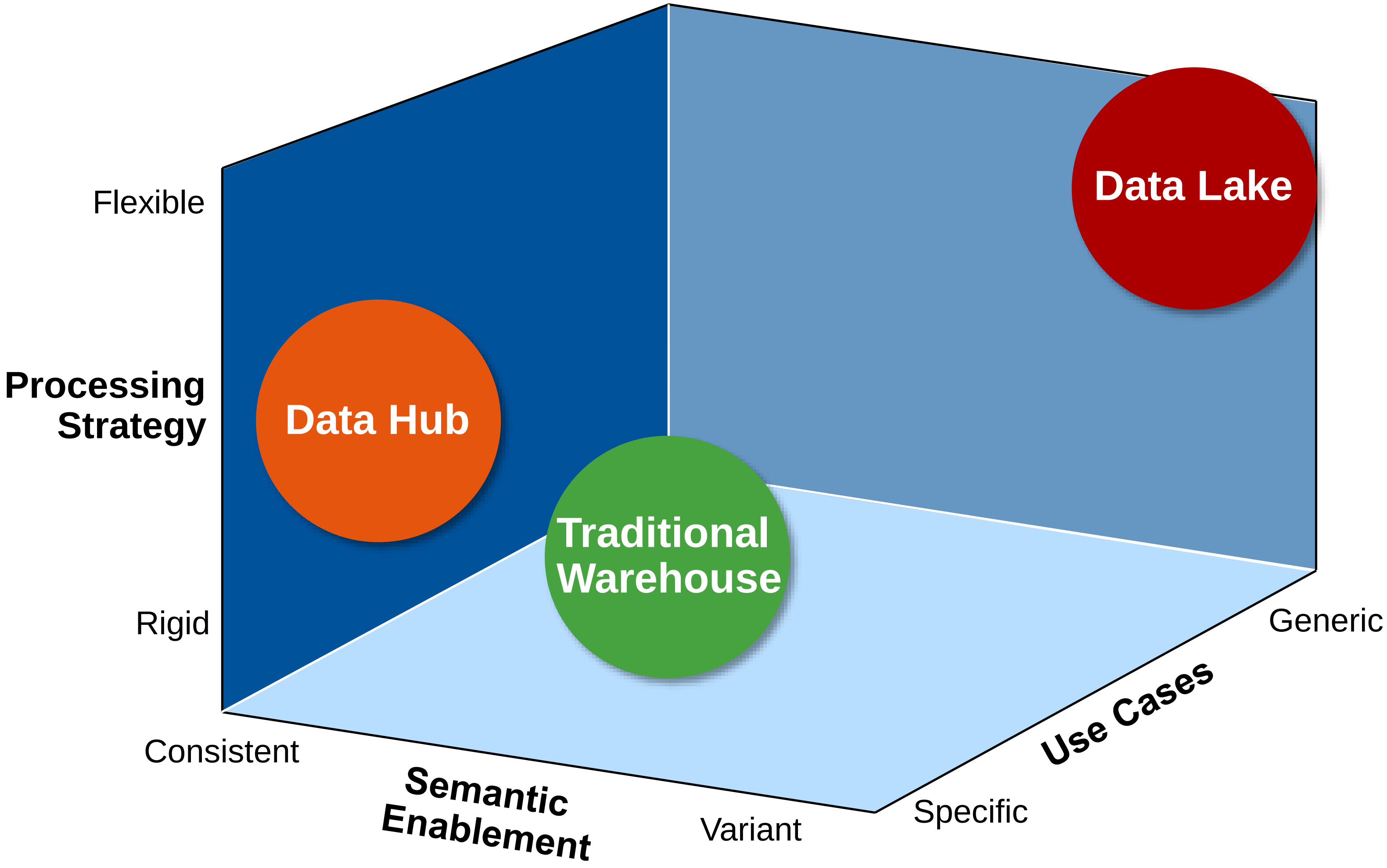
Key Issues

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?

Each Choice Optimizes Data Management Differently



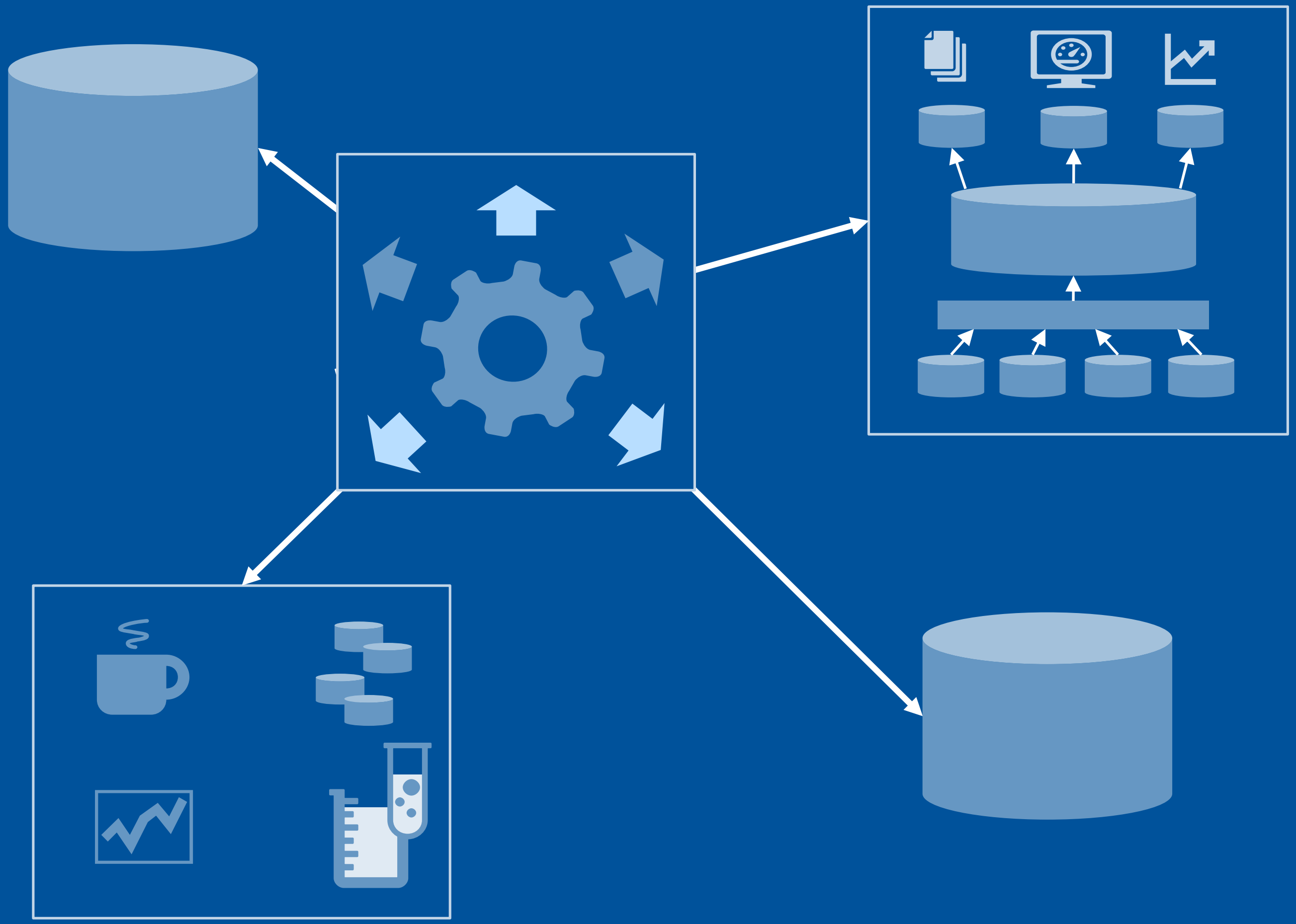
Each Choice Optimizes Data Management Differently



Hubs, Lakes and Warehouses Aren't Exclusive Choices

Hub-centric strategy:

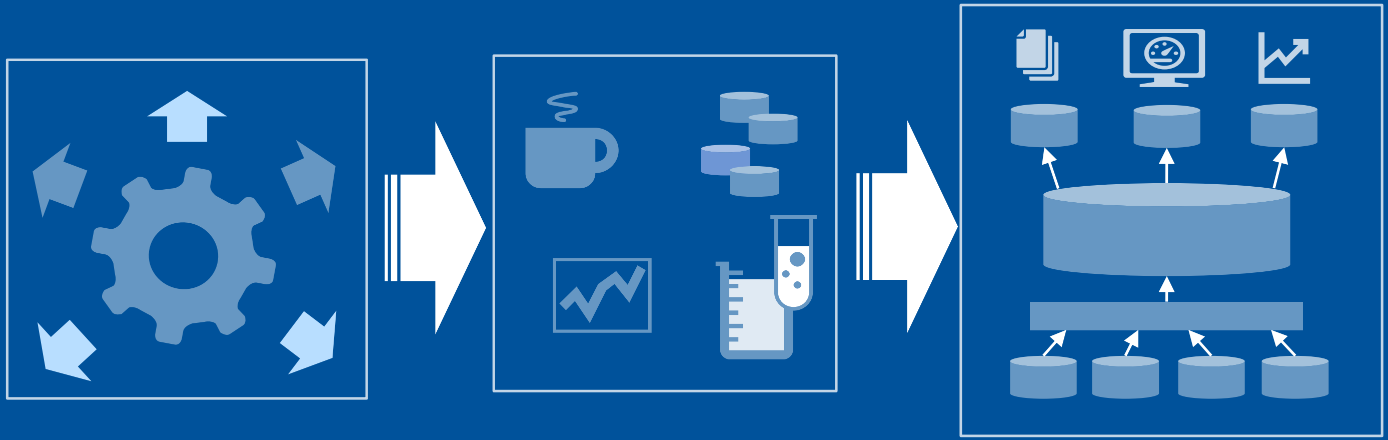
- Link, share and govern diverse datasets for operational use cases
- While uncommon, lakes and warehouses can be data sources for hubs



Hubs, Lakes and Warehouses Aren't Exclusive Choices

Collect-centric analytics strategy:

- Support discovery, self-service and optimized analytics delivery
- Enables the broadest range of analytics producers and consumers



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?

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

Data Lake Implementation Technologies

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

Strategic Planning Assumption

By 2020, 30% of data lakes will be built on standard relational technology at equal or lower cost than Hadoop.

Why It Will Happen:

- RDBMSs are the enterprise standard and the ecosystem is very mature
- Application performance is superior
- Most RDBMSs support nonrelational data in multiple formats, and can support a schema-on-read approach
- Not all "native format" data is nonrelational
- Most data going into data lakes is relational, from operational systems
- RDBMSs are not more expensive

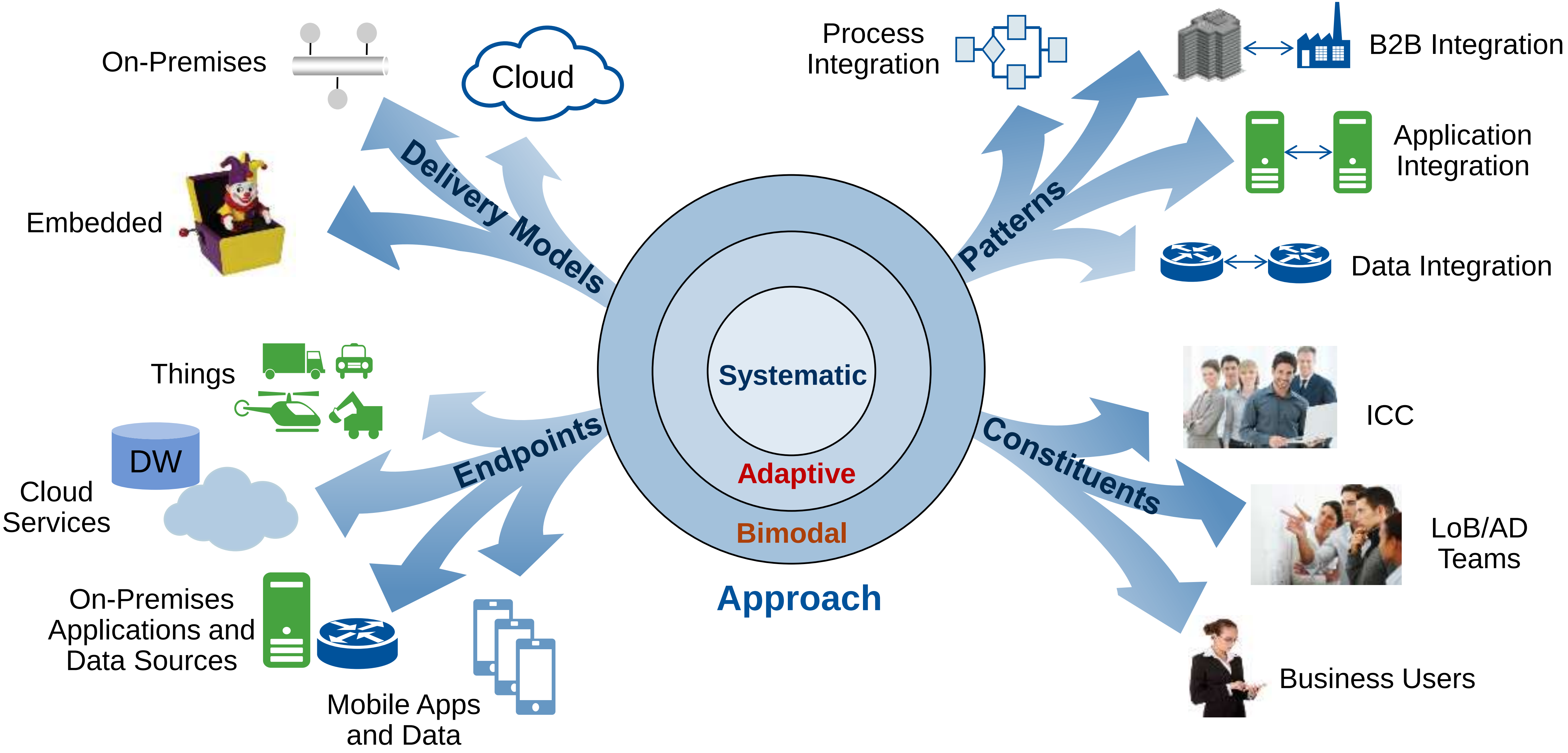
Why It Won't Happen:

- Rapid ingest of data into schema-on-read platforms is easier than conforming to a relational model
- Increasing demand for analysis of nonrelational data that does not fit easily (or efficiently) into an RDBMS
- Cloud object stores replace HDFS

Data Hub Technologies and Tools

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

Integration in Digital Business Must Be "Hybrid" in Many Dimensions

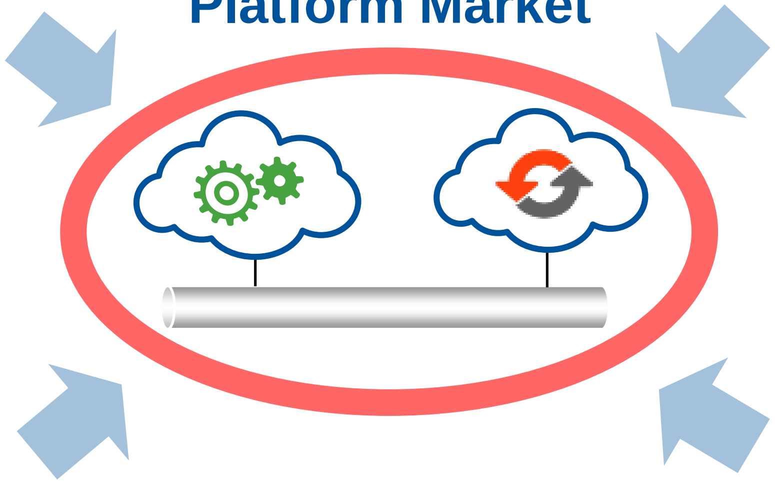


Linking the Warehouse, Lake and Hub: Diverse Vendor Landscape for Integration Technology

On-Premises Application/Data Integration Providers

- Actian
- Adeptia
- Axway
- Cisco
- Denodo
- Fiorano
- Fujitsu
- IBM
- Informatica
- Information Builders
- Infor
- InterSystems
- Magic Software
- Microsoft
- MuleSoft
- Oracle
- Red Hat
- SAP
- SAS
- Scribe Software
- Syncsort
- Talend
- TIBCO Software
- SEEBURGER
- Software AG
- WSO2

Hybrid Integration Platform Market



iPaaS Providers

- Adaptris
- Attunity
- Celigo
- Cloud Elements
- Dell Boomi
- DBSync
- Flowgear
- Fujitsu
- IBM
- Informatica
- Infor
- Jitterbit
- Microsoft
- Moskitos
- MuleSoft
- Oracle
- SAP
- Scribe Software
- SnapLogic
- Talend
- TerraSky
- Youredi

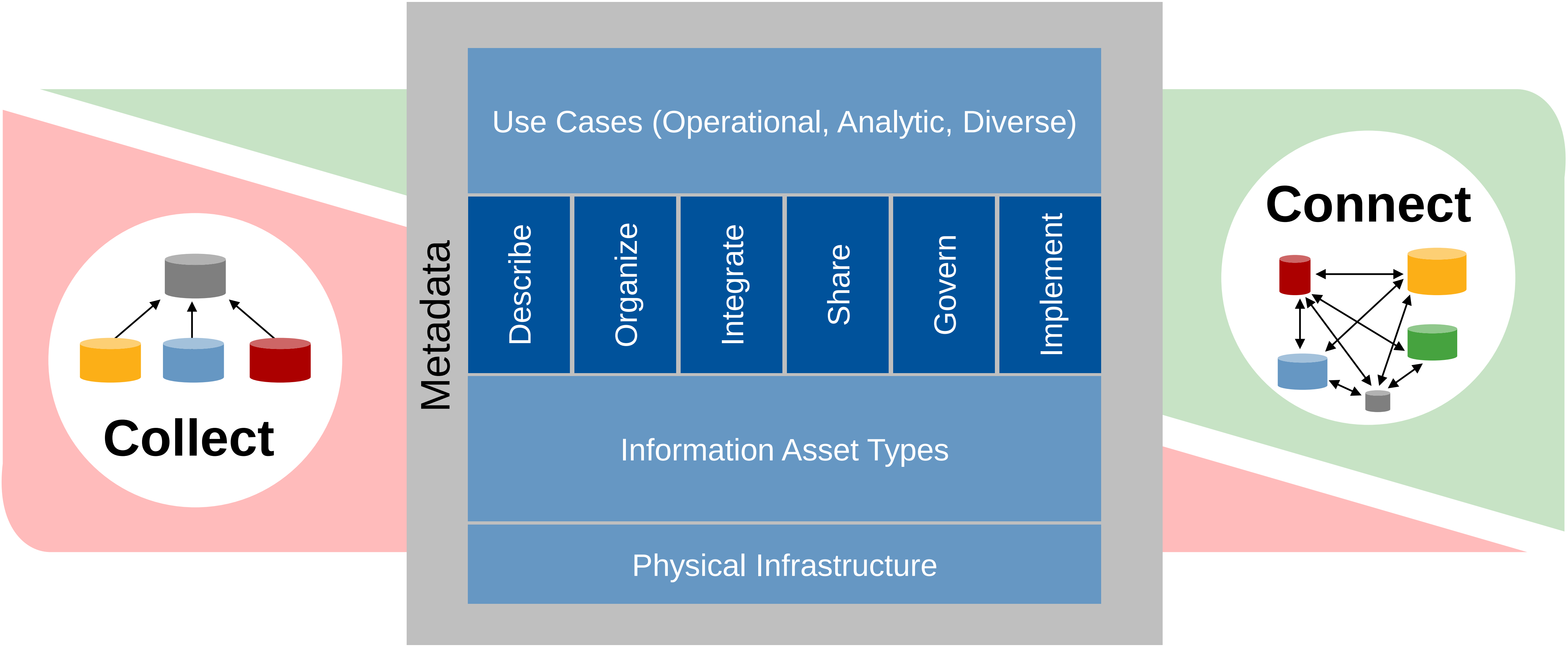
iSaaS Providers

- Actian
- Adeptia
- Azuqua
- bip.io
- cloudHQ
- Cirruspath
- elastic.io
- IFTTT
- itDuzzit
- MuleSoft
- Nubera (CloudWork)
- OneSaas
- SnapLogic
- TIBCO Software
- Wappwolf
- We Wired Web
- Zapier

Embedded Integration

- iBPM suites
- IoT platforms
- Mobile app development platforms/mBaaS
- Packaged applications/SaaS
- Self-service data preparation tools
- Others

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



Collect

Metadata

Use Cases (Operational, Analytic, Diverse)

Describe

Organize

Integrate

Share

Govern

Implement

Information Asset Types

Physical Infrastructure

Connect

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.

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 M. Ronthal (G00336273)

For information, please contact your Gartner representative.