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

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

Use Cases (Operational, Analytic, Diverse)

- Describe
- Organize
- Integrate
- Share
- Govern
- Implement

Information Asset Types

Physical Infrastructure

Collect

Connect

1 © 2019 Gartner, Inc. and/or its affiliates. All rights reserved. Gartner is a registered trademark of Gartner, Inc. and its affiliates.
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 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?
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?
How Do Lakes and Warehouses Relate?

Logical data warehouse overlays the whole chart.
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?
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.
The Elements of a Data Hub Strategy

Work Down From Top

- First: Governance Tier
- Second: Sharing Tier
- Third: Implementation Tier

Iterate

Then

Governance Tier
Sharing Tier
Implementation Tier
Types of Data Hubs

Data Hub Strategy

- **MDM** (Master Data Management): e.g., Master Data Operational Data. Business Process Integrity Complex, End-to-End Processes.
- **Integration**: e.g., Varied Data Mixed Use. Effective and Efficient Data Access, Synchronization and Provisioning.
- **RDM** (Reference Data Management): e.g., Reference Data Mixed Use. Effective and Efficient Data Look-Up and Synchronization.
- **Analytics**: e.g., Analytics Data Mixed Use. Effective and Efficient Analytical Data Synchronization and Provisioning.

Governance, Sharing, Integration.
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?
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?
How Do Hubs, Lakes and Warehouses Differ?

- **Data Warehouse**
  - Optimized Processing: 5.0
  - Consistent Semantics: 2.5
  - Specific Use Cases
  - Flexible Processing

- **Data Lake**
  - Optimized Processing: 0.0
  - Consistent Semantics: 0.0
  - Specific Use Cases
  - Flexible Processing

- **Data Hub**
  - Optimized Processing: 0.0
  - Consistent Semantics: 0.0
  - Specific Use Cases
  - Flexible Processing
Hubs, Lakes and Warehouses Aren’t Exclusive Choices

Operational Context — Governed Data Sharing

- ERP
- CRM
- SCM
- HCM

External Suppliers
Customers
Partners
Hubs, Lakes and Warehouses Aren’t Exclusive Choices

Analytical Context — Hub-Centric Consolidation

- ERP
- CRM
- SCM
- HCM

External Suppliers
Customers
Partners
Hubs, Lakes and Warehouses Aren’t Exclusive Choices

Analytical Context — Lake-Centric Consolidation

ERP
CRM
SCM
HCM

External Suppliers
Customers
Partners
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.
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

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

**Use Cases**
- Analytics and Data Warehousing
- Data Consistency
- Data Migration
- Master Data Management
- Interenterprise Data Sharing
- Self-Service Integration

**Data Delivery Styles**
- Bulk/Batch Data Movement
- Data Virtualization
- Stream Data Delivery
- Data Replication and Synchronization
- Message-Oriented Movement of Data

**Data Sources**
- Transactional Data
- Cloud Data
- Documents
- Social
- Hadoop
- IT/OT
- IoT
- Image
- Audio
- Text
- Video

**Common design, metadata, admin., optimization, governance**

**Hybrid mode of deployment:** On-premises, cloud, distributed
Apply the Right Combination of Lakes, Warehouses and Hubs to Best Enable Data Sharing and Analytics

- Use Cases (Operational, Analytic, Diverse)
  - Describe
  - Organize
  - Integrate
  - Share
  - Govern
  - Implement

- Information Asset Types
- Physical Infrastructure

Collect

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 Ronthal (G00336273)