Are You Killing the Benefits of Your Data Lake?

denodo

R20 CONSULTANCY

Speakers

Rick van der Lans

Independent Business Intelligence Analyst

R20 Consultancy

@rick_vanderlans

Lakshmi Randall

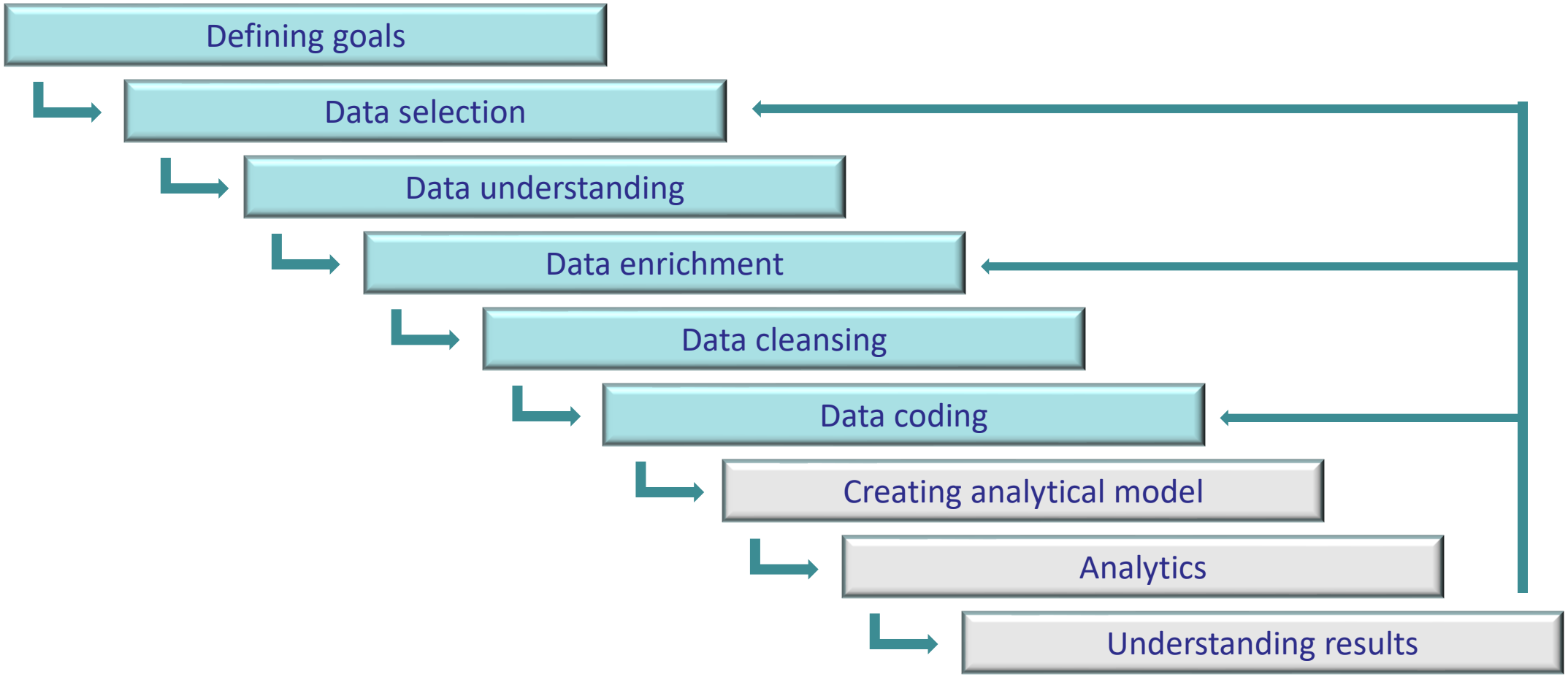
Director of Product Marketing

Denodo

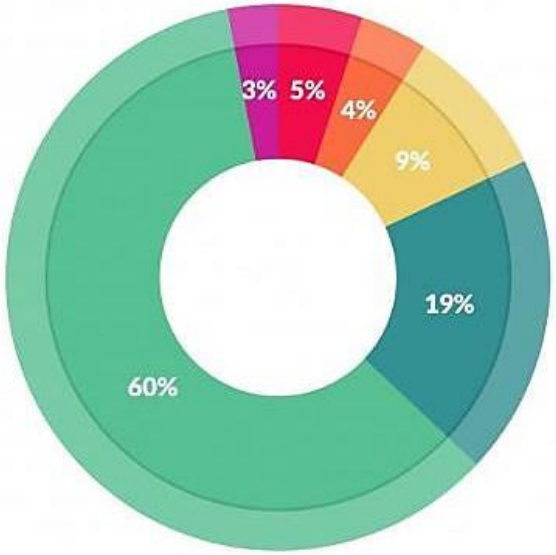
@LakshmiLJ

Wikipedia: Data science is an interdisciplinary field of scientific methods, processes, algorithms and systems to extract knowledge or insights from data in various forms, either structured or unstructured, similar to data mining.

Data Science Steps and Data Preparation

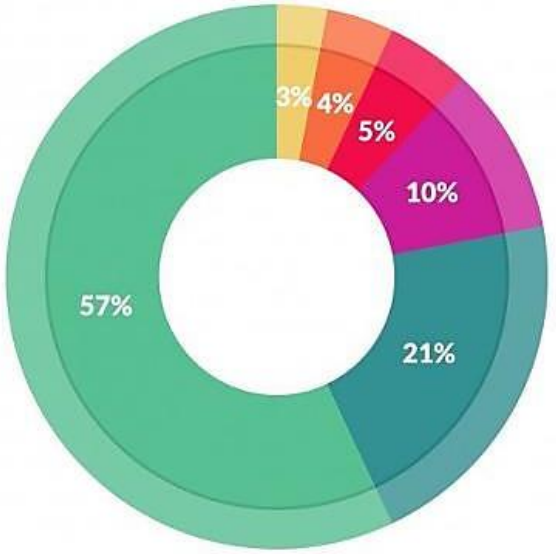


Data Preparation is Time-Consuming



What data scientists spend the most time doing

- *Building training sets: 3%*
- *Cleaning and organizing data: 60%*
- *Collecting data sets; 19%*
- *Mining data for patterns: 9%*
- *Refining algorithms: 4%*
- *Other: 5%*



What's the least enjoyable part of data science?

- *Building training sets: 10%*
- *Cleaning and organizing data: 57%*
- *Collecting data sets: 21%*
- *Mining data for patterns: 3%*
- *Refining algorithms: 4%*
- *Other: 5%*

Source: Gill Press, "Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says", March 2016

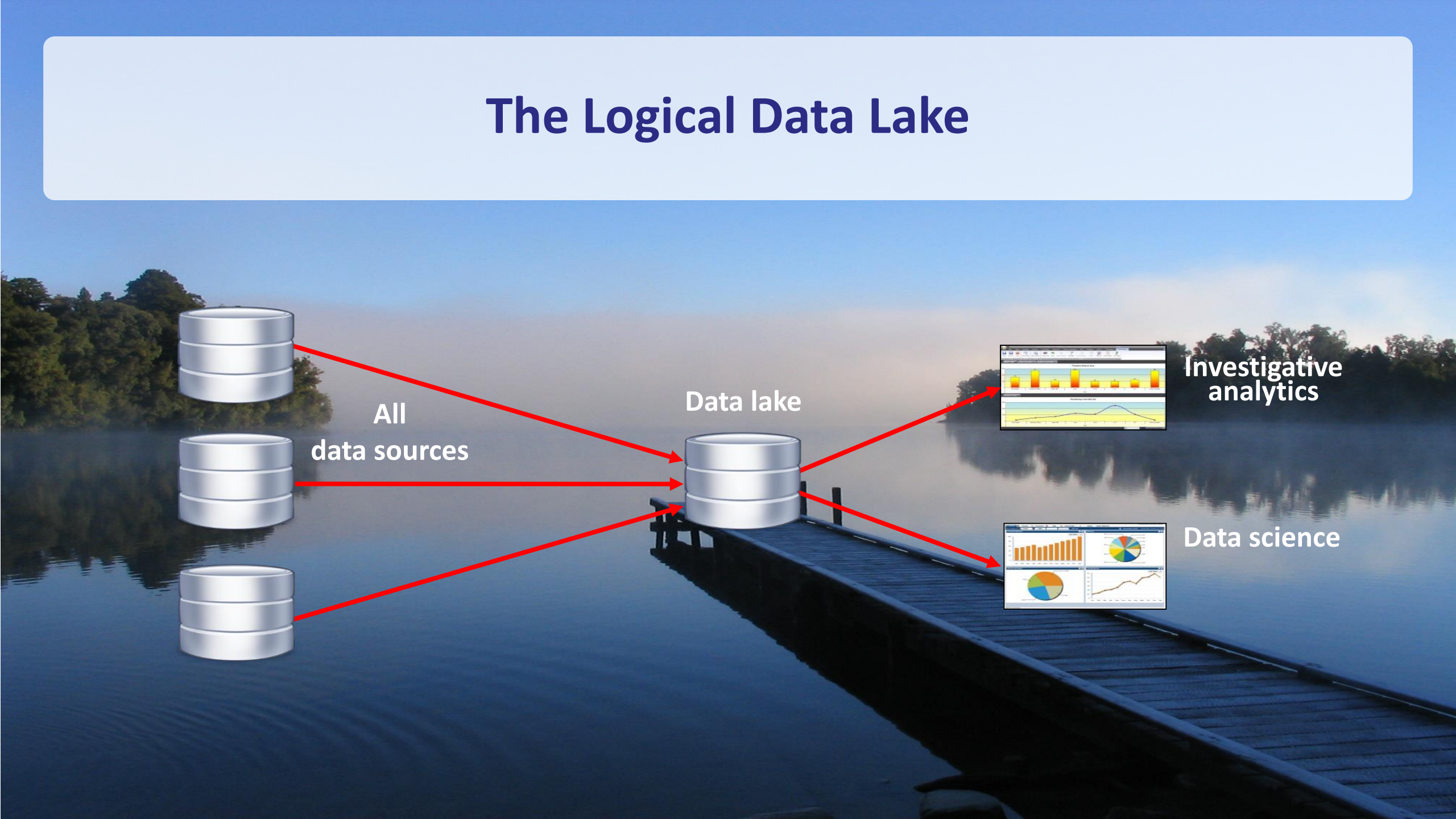
Common Definition of Data Lake

James Serra:

> A “data lake” is a storage repository, usually in Hadoop, that holds a vast amount of raw data in its native format until it is needed. It’s a great place for investigating, exploring, experimenting, and refining data, in addition to archiving data.

Source: http://www.jamesserra.com/archive/2015/04/what-is-a-data-lake/

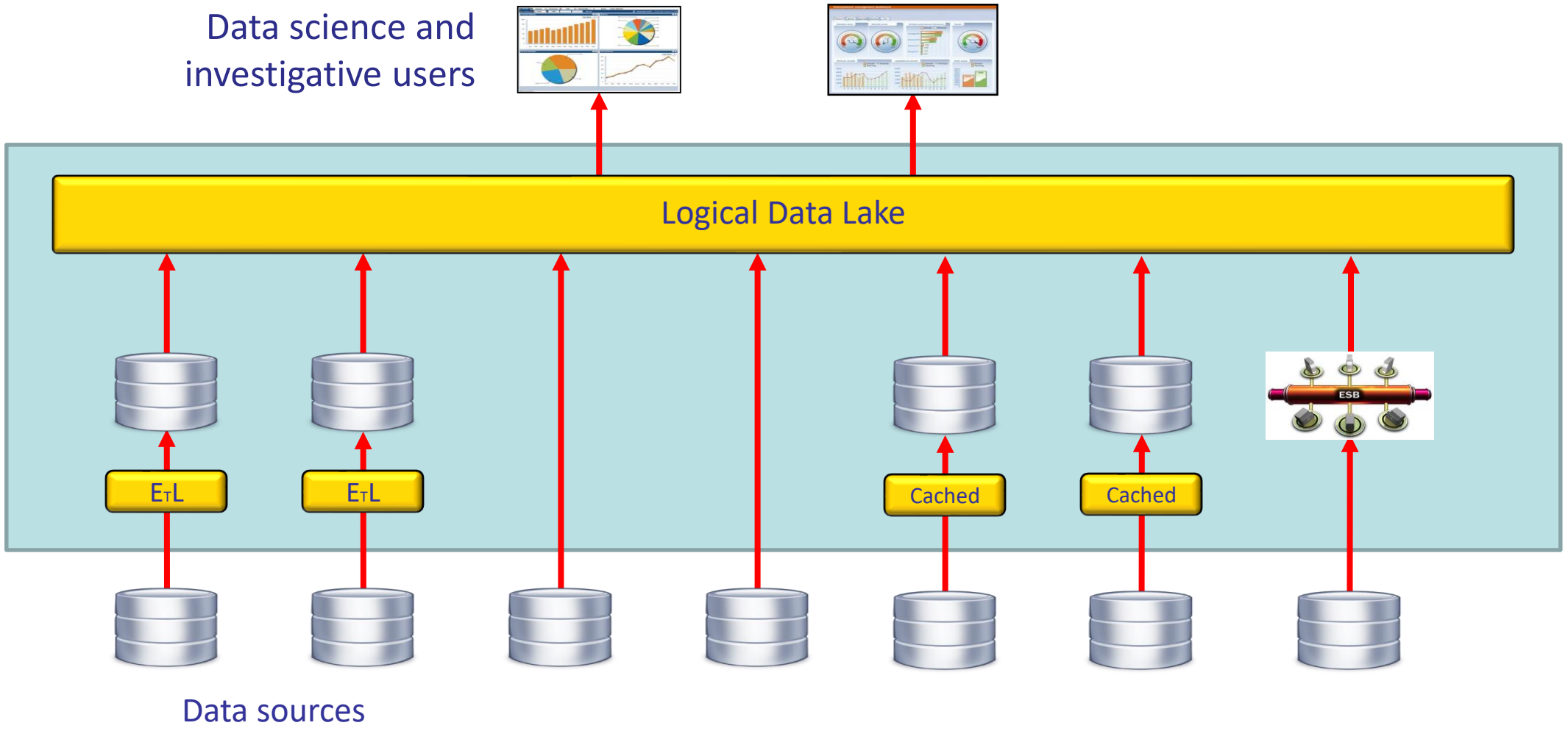
The Logical Data Lake



Challenges of a Physical Data Lake

- Complex “T” moved to data usage
- Big data too big to move
  - Too slow to copy and bandwidth issues
- Uncooperative departments - company politics
- Restricting data privacy and protection regulations
- Data in data lake is stored outside original security realm
- Missing metadata to describe data
- Some sources are hard to copy
  - For example, mainframe data
- Refreshing of data lake
- Management of data lake required
- …

The Logical (Virtual) Data Lake

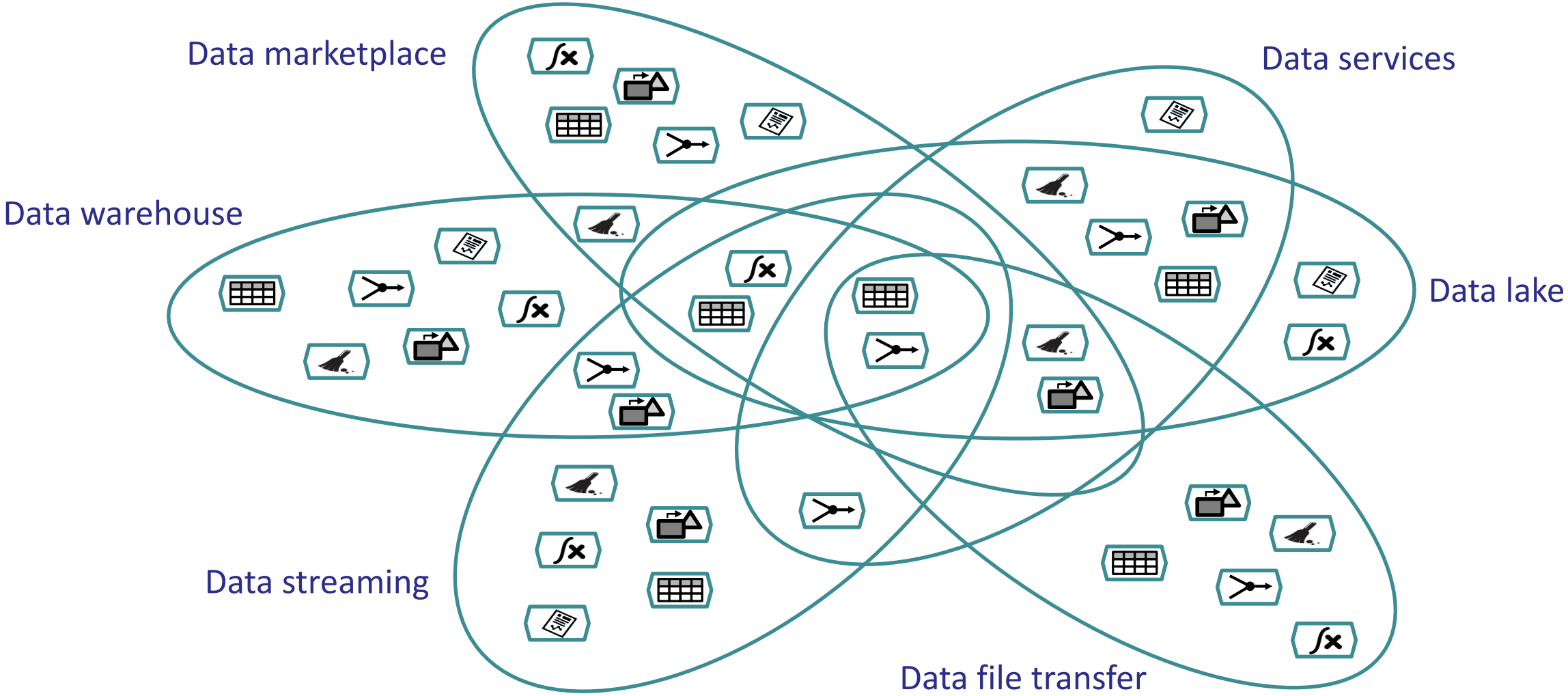


“Data is too valuable an asset to be used for reporting only.”

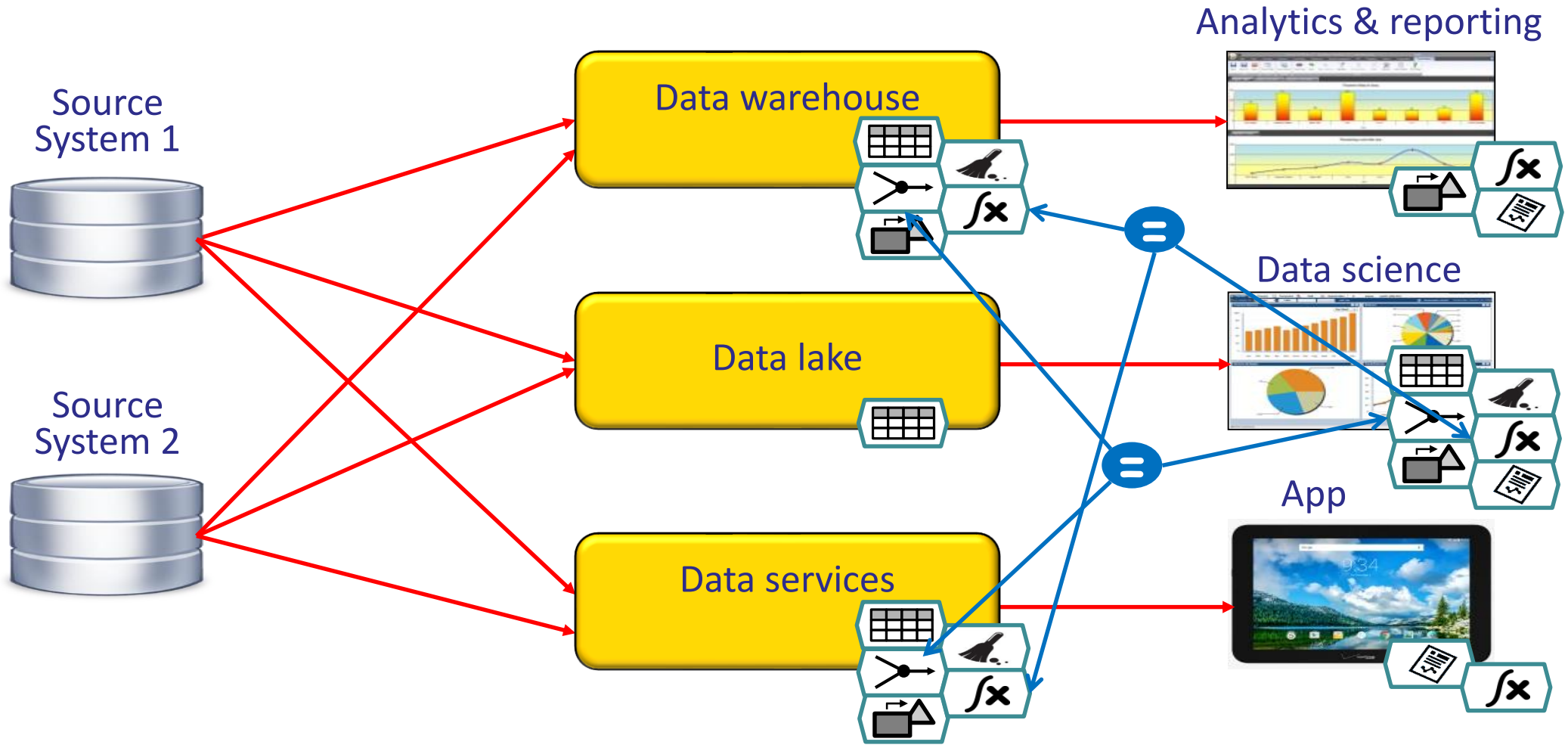
A Multitude of Data Delivery Systems

- The classic data warehouse architecture
- The data lake
- The data marketplace
- Data services
- Managed file transfer
- Data streaming
- …

Drawback: Replicated Specifications

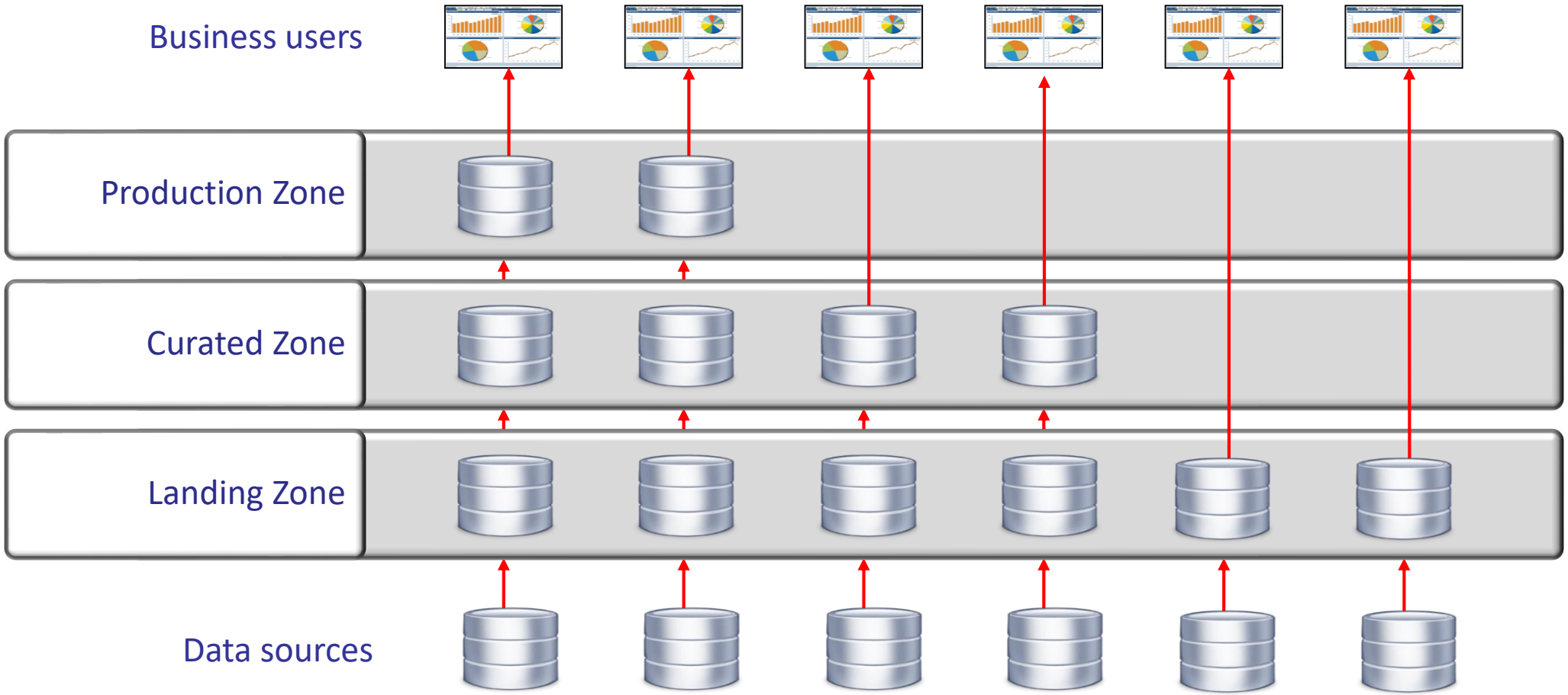


Drawback: Replicated Specifications

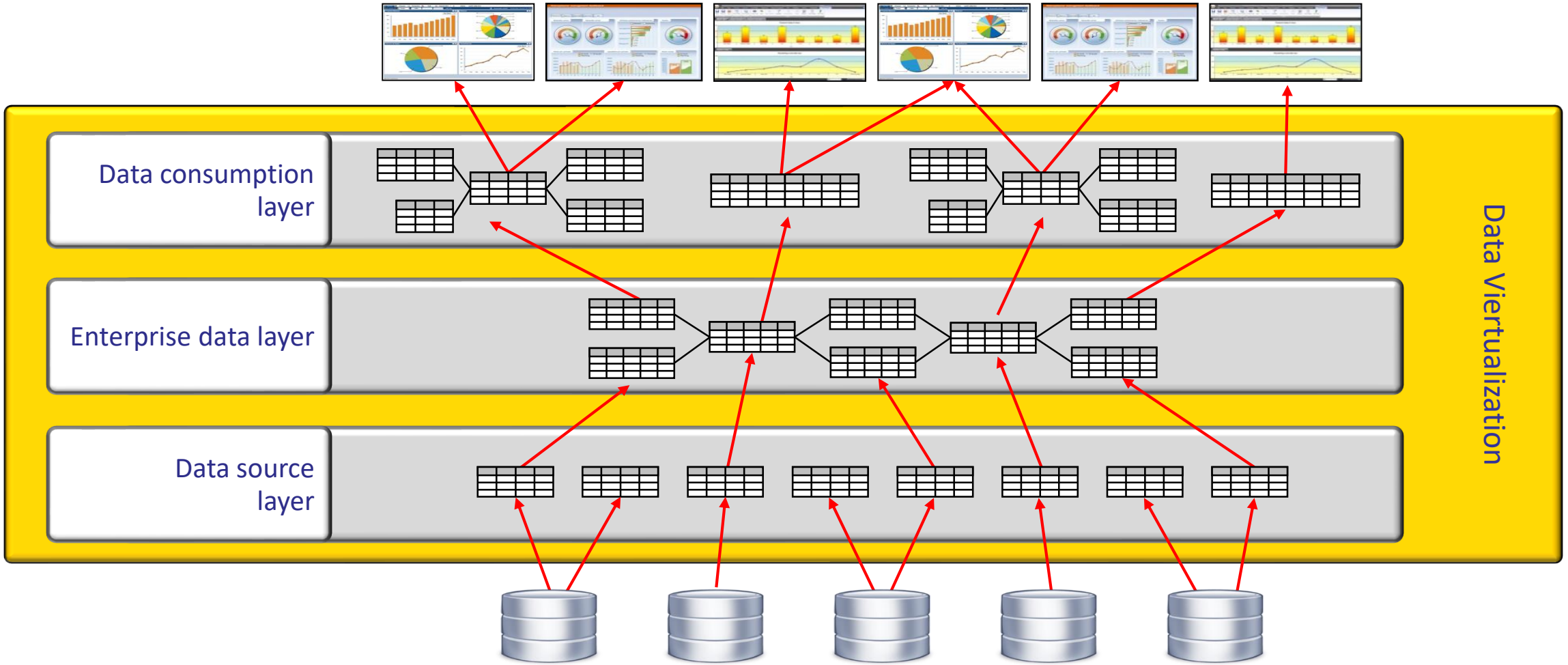


Siloed Data Delivery Systems

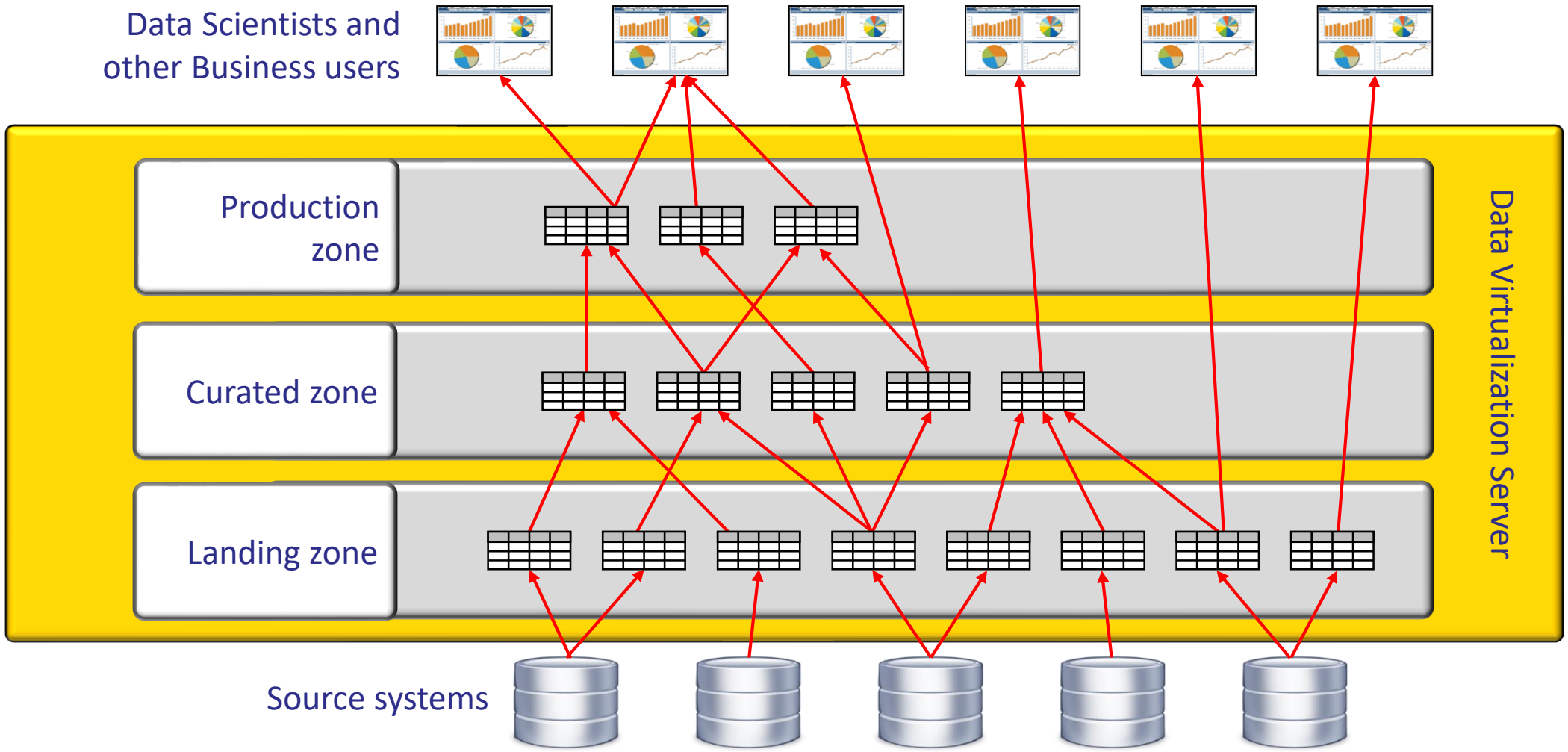
A Physical Data Lake With Multiple Zones



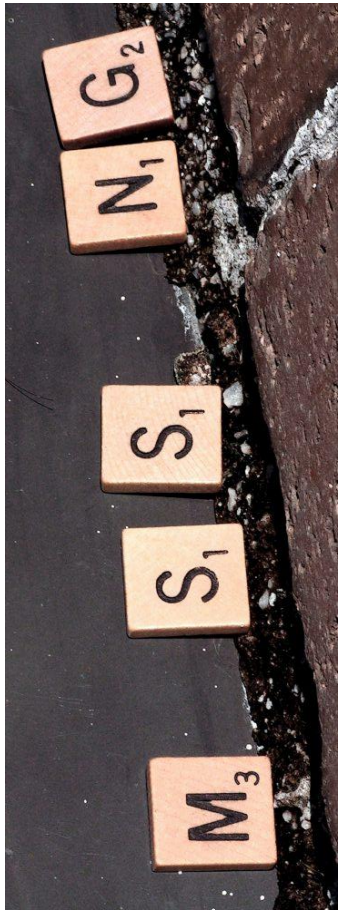
The Logical Data Warehouse Architecture



The Logical, Multi-Purpose Data Lake



Key Features Missing in SQL-on-Hadoop Engines



- Allowing applications and users to access all the data through another interface than SQL
- Allowing all types of data sources to be accessed
- Detailed lineage and impact analysis capabilities
- A searchable data catalog
- Advanced query optimization techniques for federated queries
- Advanced query pushdown and parallel processing capabilities
- Centralized data security

Single-Purpose versus Multi-Purpose Data Lake (1)

- The Single-Purpose Data Lake
  - Not always practical or feasible
  - The data in a data lake is potentially too valuable to be used by data scientists exclusively
  - Other user groups may be interested in the data lake
  - Siloed data delivery system operating independently of others
  - Multiple physical layers of lakes is complex

Single-Purpose versus Multi-Purpose Data Lake (2)

- The Multi-Purpose Data Lake
  - Some data is physically stored centrally (through copying or caching), and some is accessed remotely
  - The data offered can be accessed by any type of business user
  - The data in the data sources can be transformed to any form that is required by other user groups
  - A logical, multi-purpose data lake can be the foundation for several data delivery systems
  - Working with logical layers is easy to manage and maintain

Advantages Multi-Purpose Data Lakes

- Reduction of development costs
  - Metadata specifications are defined once and reused many times
  - Analytical solutions developed by one data scientist can easily be reused
  - Data-related solutions developed by non-data scientists can be reused
- Acceleration of development
  - Data scientists don’t need to spend time on data selection
  - Physically copying data is not mandatory, but optional
  - Business user don’t have to learn the technical languages and APIs of the original data sources
- Increase report and analytical consistency
  - Reusing analytical and data-related solutions improve the reporting and analytical consistency
  - Definitions, descriptions, tags, and categories can be centrally cataloged
  - Access to all the data can be centrally secured

Time to Tear the Silos Down!

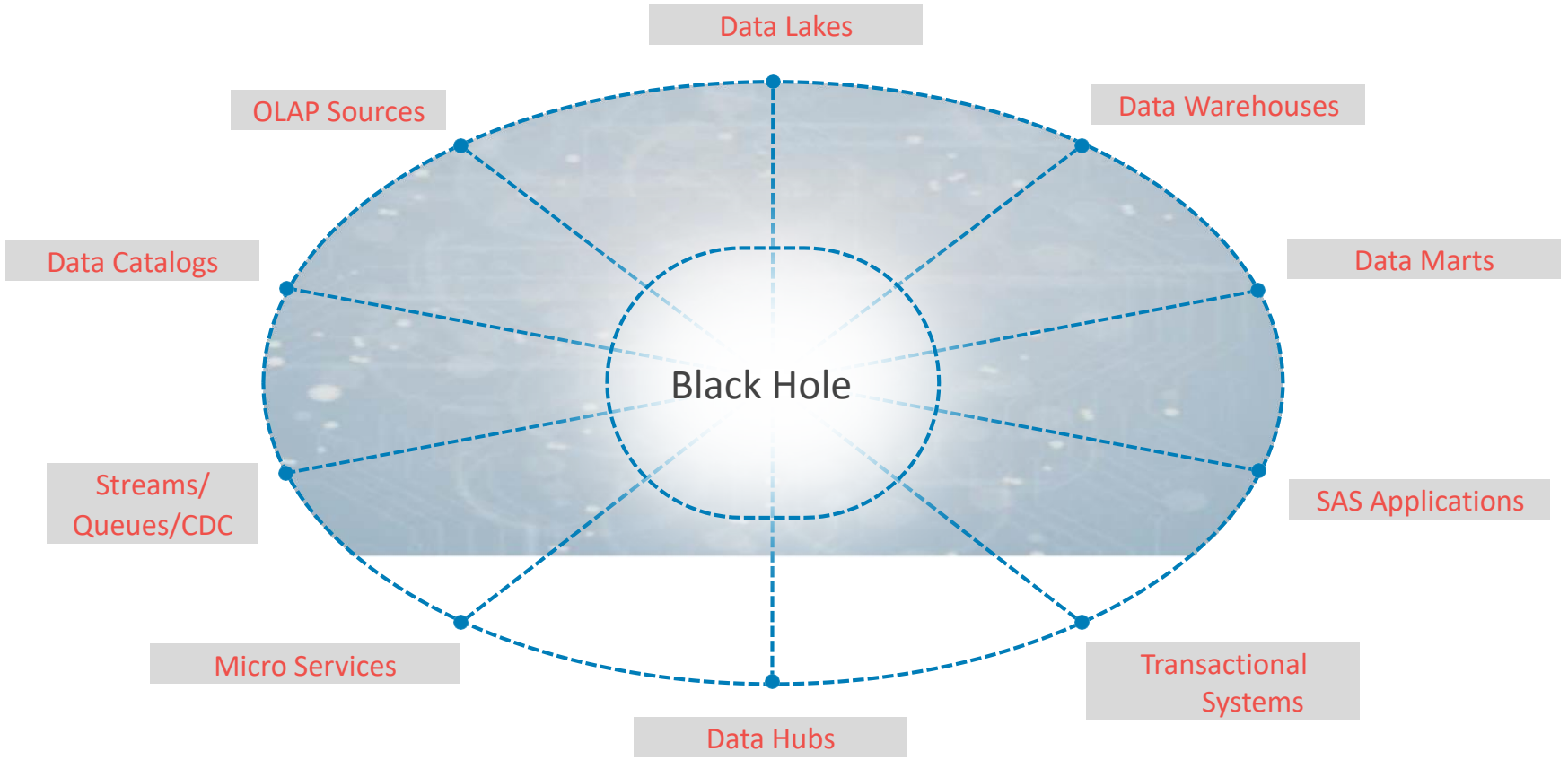
Data Virtualization

Big Data Hadoop Deployments

Shhh… the ugly little secret is that big data deployment is hard!

1) Gartner Analyst, Nick Heudecker; infoworld.com, Sept 2015

Fifty Shades of Data Management



The Analytics Environment Must Have a Brain…

- For companies to benefit from their analytic efforts, data must be:
  - Easily located (wherever it resides)
  - Easily understood (with all its context in place)
  - Easily accessed (query performance is critical)
  - Easily audited (its lifecycle is clear to both IT and business users)
  - Appropriately provisioned for analysis (its management is known)

A Few Simple Rules…

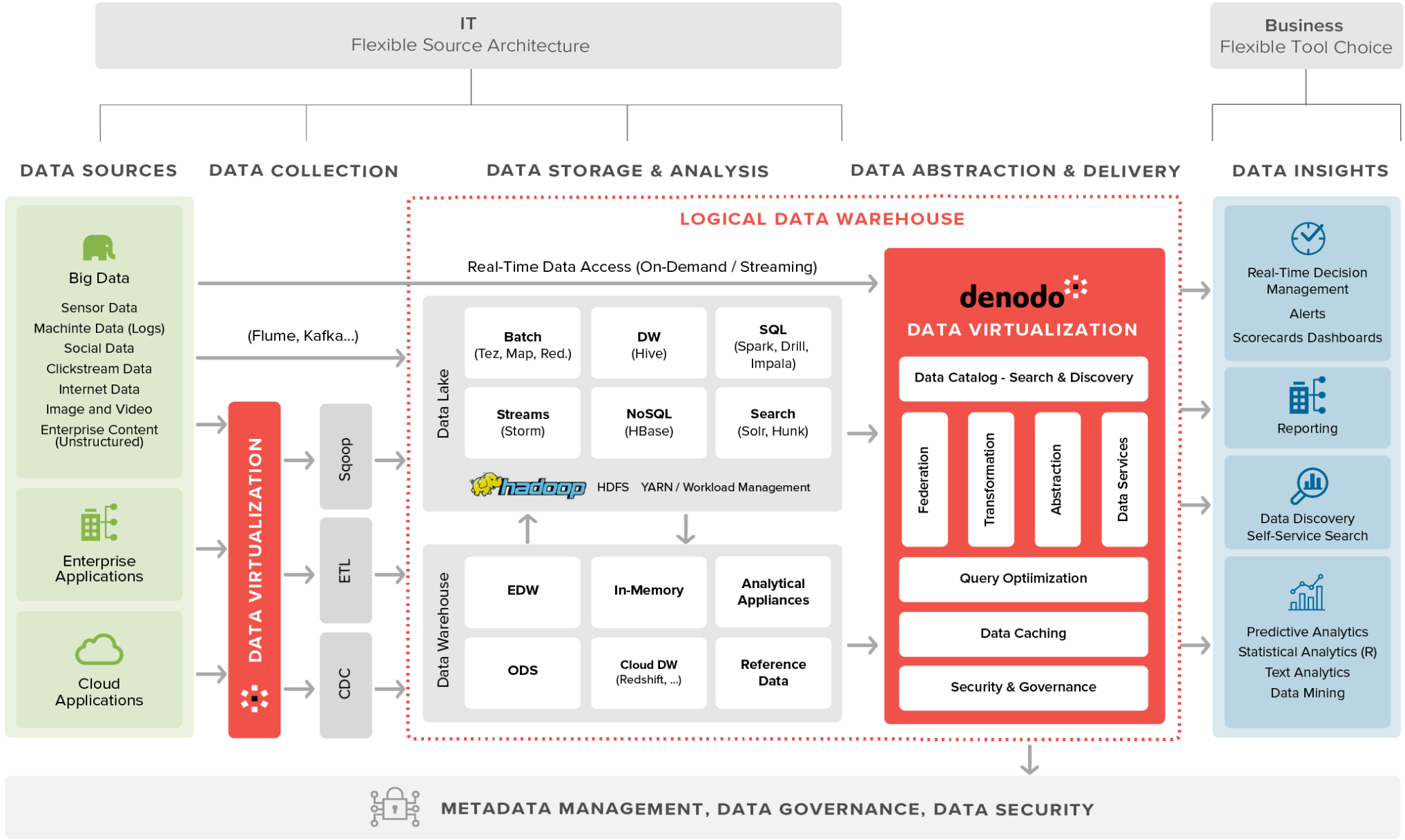
1. Build a business strategy rather than a big data strategy
2. Big data is really about ***small***
3. Users come in all shapes and sizes
    - Who are they? What data do they need? What flexibility do they want?
4. Connect to all of the data (but start with the most important)
    - What data is needed by the users? Open access or pre-aggregated or pre-calculated?
5. Use the language that the business understands
    - Don’t force people to change terminology…support multiple models, e.g., to Finance it’s an ‘account’, to Customer Care it’s a ‘customer’.

Self-Service With Guardrails

- Don’t build just for the ‘data cowboys’
- Create pre-integrated, pre-calculated data
  - Eliminating this burden from the users.
  - Ensures consistency of calculations, etc.
- But allow the cowboys to ‘roam and wrangle’
- Even the cowboys can only access ‘approved’ data sources

A Single, Logical, Multi-Purpose Data Lake

Multi-Purpose Data Lake With Data Virtualization



ENABLING TECHNOLOGIES

Virtualize Data, Don't Migrate it

- Distributed heterogeneity is a challenge for the MDA
  - Plague of data standards, models, quality metrics, interfaces…
- Consolidating diverse data is not a compelling solution
  - Migration & consolidation alleviate complexity, but have other problems
  - Time consuming, risky, disruptive, distracting
- DV is effective alternative to consolidation
  - Fraction of the time, risk, cost and disruption of migration and consolidation projects
  - Software/hardware advances give DV the speed/scale required of most SLAs & use cases

Big Data Queries Faster With Denodo Platform

Performance comparison of 5 different queries

1. Data virtualization delivers better performance without the need to replicate data into Hadoop.
2. Data virtualization leverages data source architectures for what they are good at.

	Impala Hadoop-only Runtime (s)	Denodo Runtime (s)	Denodo Runtime w/ Cache (s)	Data Volumes
Query 1	199	**120**	**68**	**Queries 1,2,3,5** •Exadata Row Count: ~5M •Impala Row Count: ~500k **Query 4** •Exadata Row Count: ~5M •Impala Row Count: ~2M
Query 2	187	**96**	**88**	
Query 3	120	212	**115**	
Query 4	timeout	**328**	**69**	
Query 5	**46**	91	56	

Anadarko

COMPANY PROFILE

Anadarko employs approximately **4,500** men and women and invested about **$4 billion** in 2017 to **find** and **develop** the **oil** and **natural gas** resources that are essential to modern life

FOCUS AREAS
U.S. ONSHORE & DEEPWATER GOM

CASH GENERATION
INTERNATIONAL OIL

FUTURE VALUE
EXPLORATION & LNG

Changing Commodity Cycle

HONED FOCUS

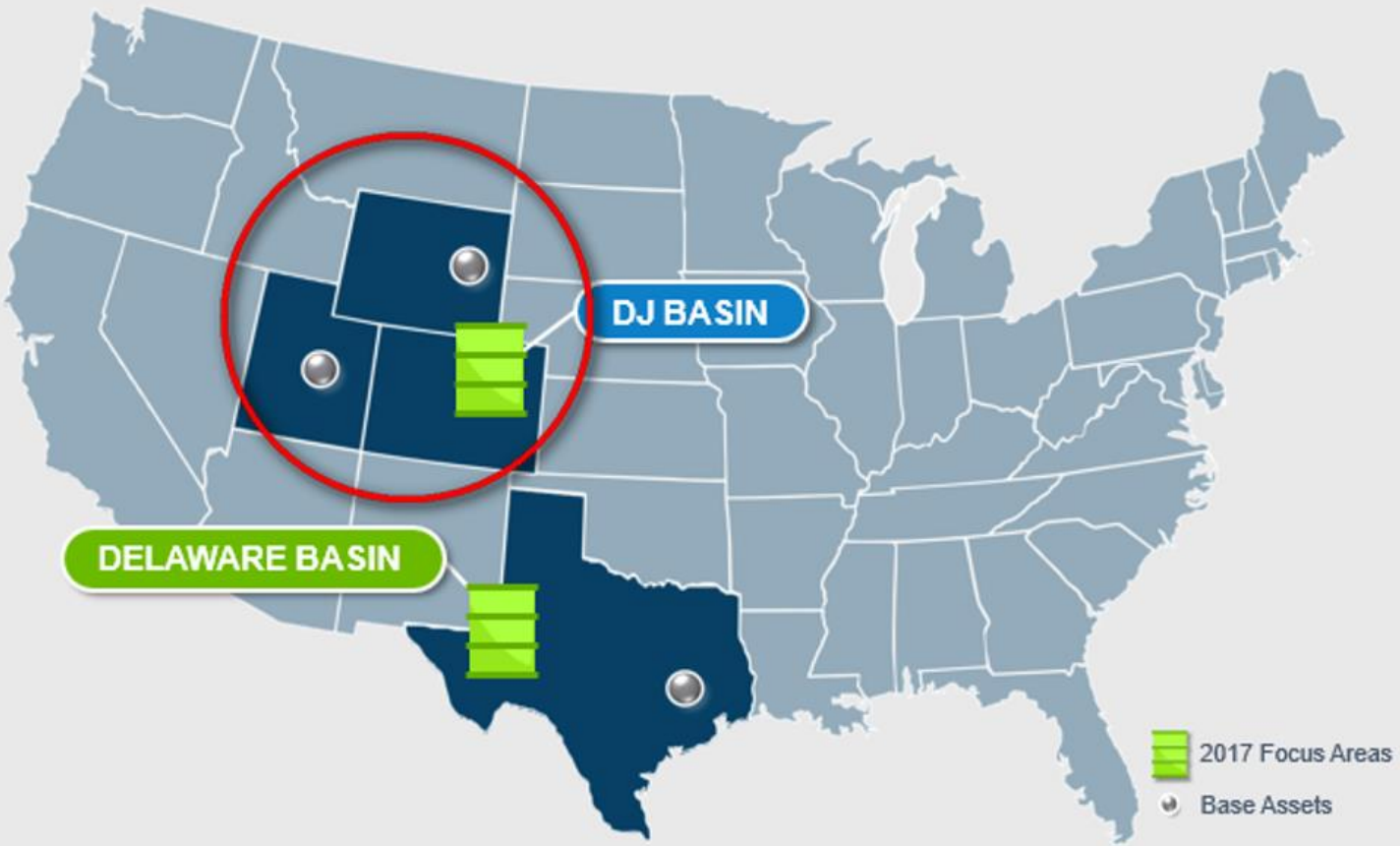
better *data*

ADJUSTED ORG

faster *data*

ENHANCED TECH

more *data*



Self-Service Data Delivery Environment

To create and use data services for analytics, reports, and apps

Results (from 2017 roll-out/implementation)…

20 corporate repositories; several non-corporate

200+ corporate views; **100+** user-defined views

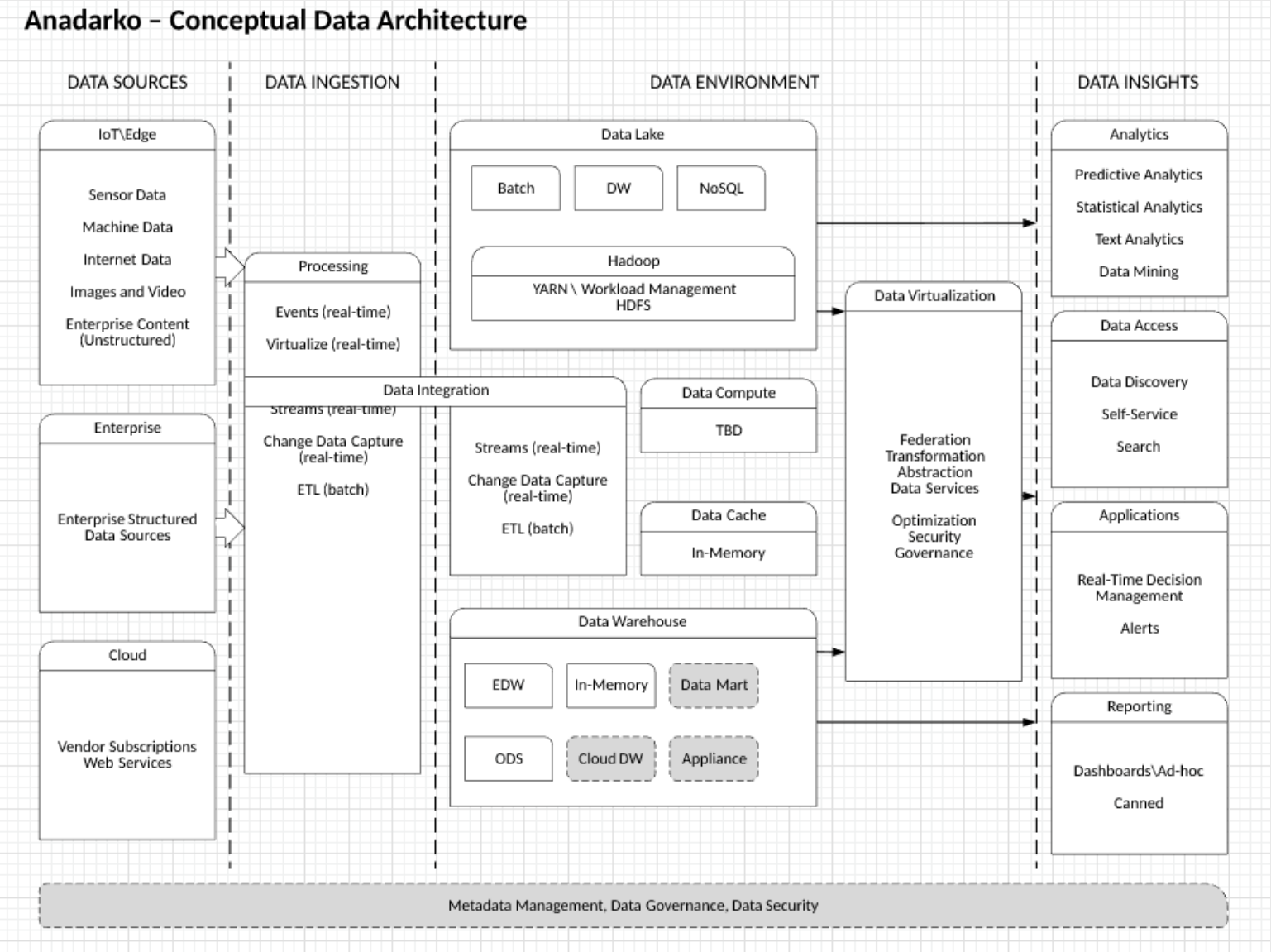
30 developers using/trained

150 direct users; **~700** indirect users

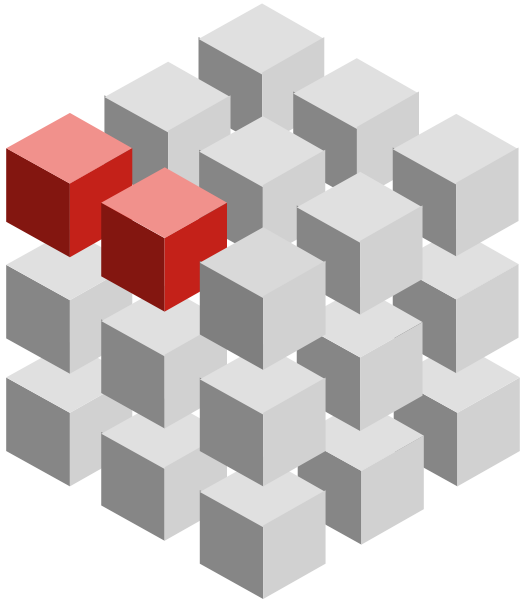
Examples

- reduced ad valorem taxes for finance
- improved (production) completion design from multi-variate analysis using virtual views
- more (combined) access to vendor subscription data exploration for competitor intelligence

Data Architecture at Anadarko



Why Multi-Purpose Data Lake?

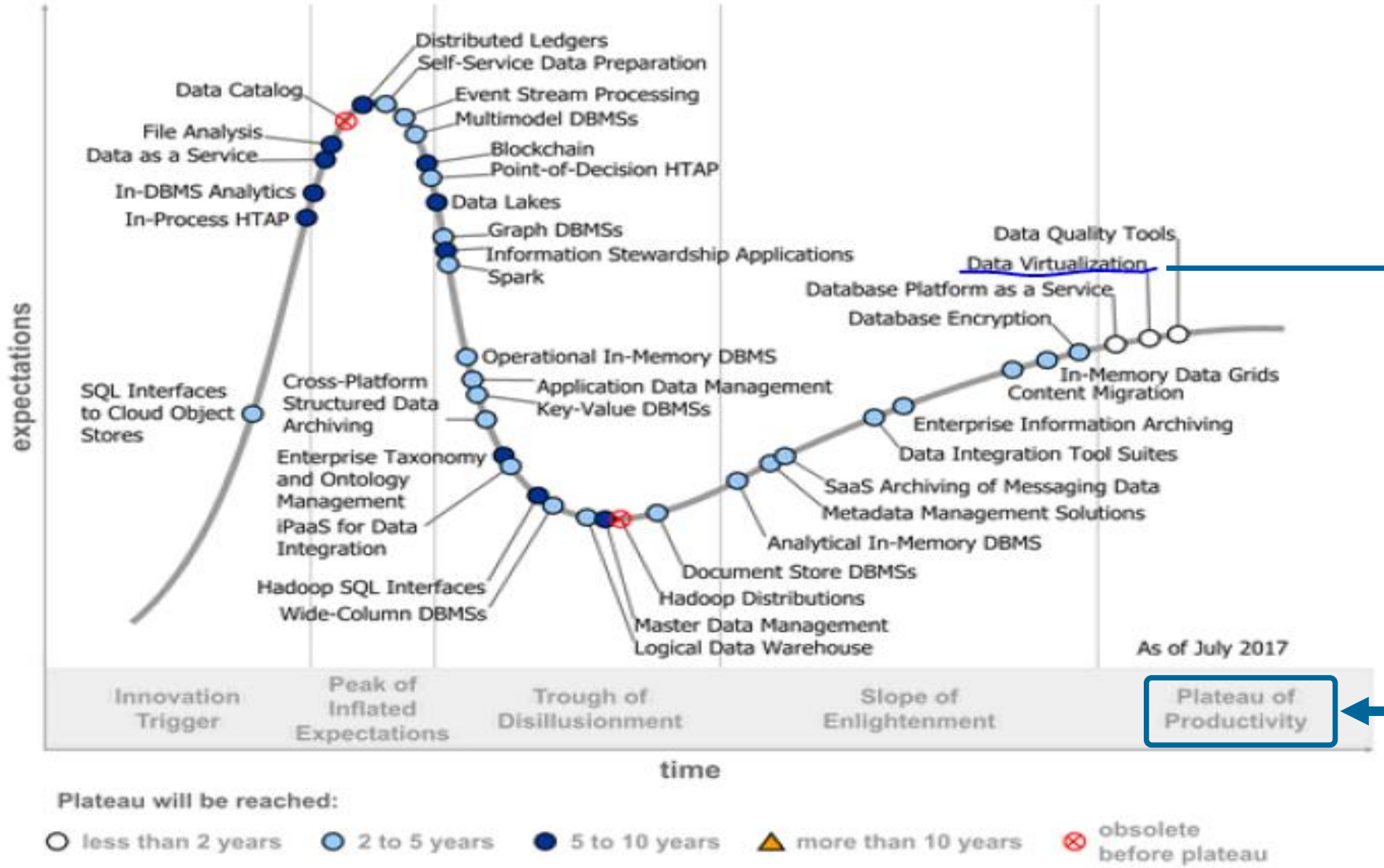


- Surface all company data without the need to replicate all data to the Hadoop lake
- Improve governance and metadata management to avoid “data swamps”
- Allow for on-demand combination of real-time (from the original sources) with historical data (in the cluster)
- Leverage the processing power of the existing data lake clusters using Denodo’s optimizer

> **Denodo's key strength is delivering a unified and centralized data services fabric with security and real-time integration across multiple traditional and big data sources, including Hadoop, NoSQL, cloud, and software-as-a-service (SaaS).**"

- Source: "Forrester Wave™: Big Data Fabric Q4 2016"

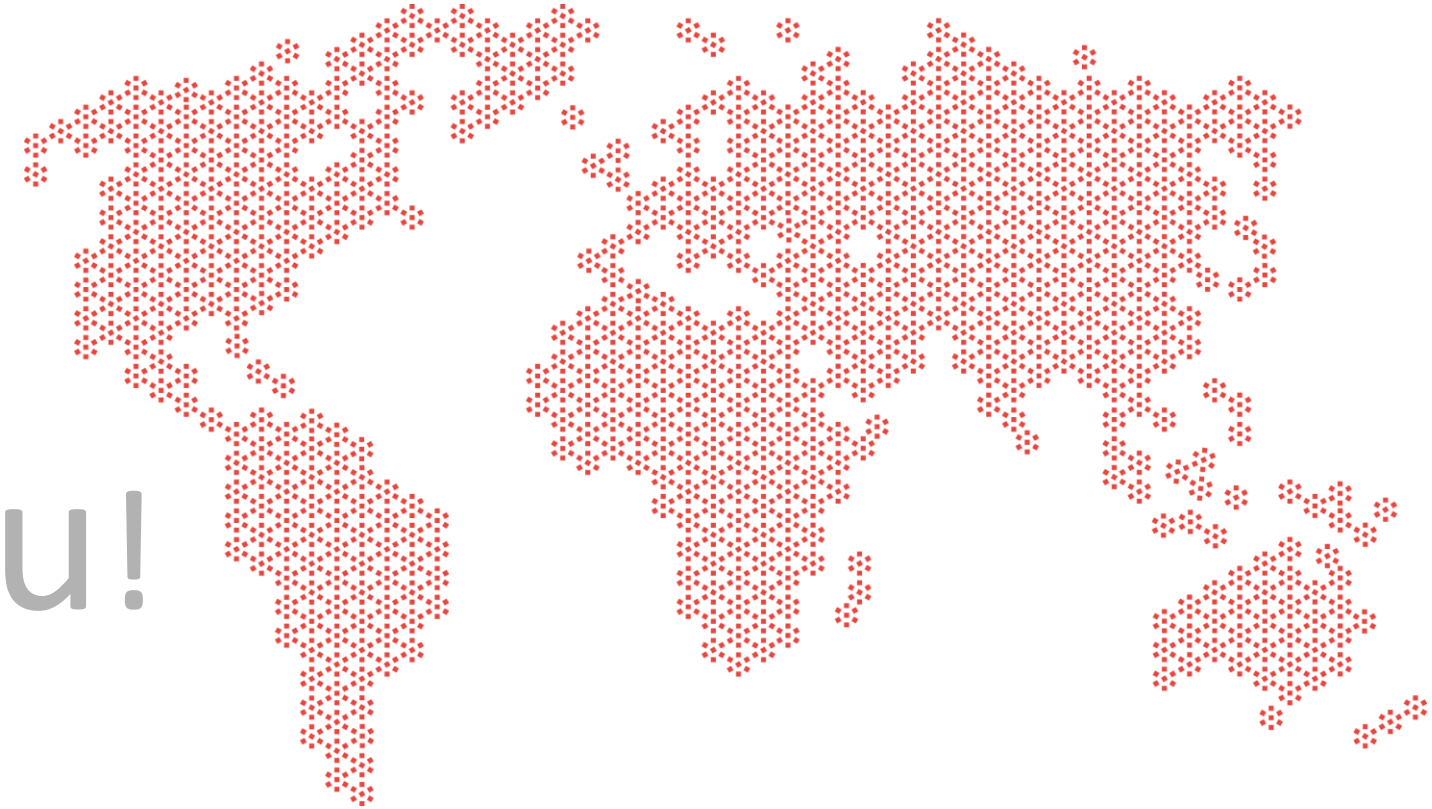
Gartner Gives DV Its Highest Maturity Rating



"Data Virtualization can be deployed with low risk and effort to achieve maximum value."

Q&A

Thank you!



denodo | www.denodo.com | info@denodo.com