Building a Single Logical Data Lake

For Advanced Analytics, Data Science, and Business Operations

Ravi Shankar, CMO
September, 2018

denodo

- Competition from a low cost vendor

- Lower the price, affecting margins?

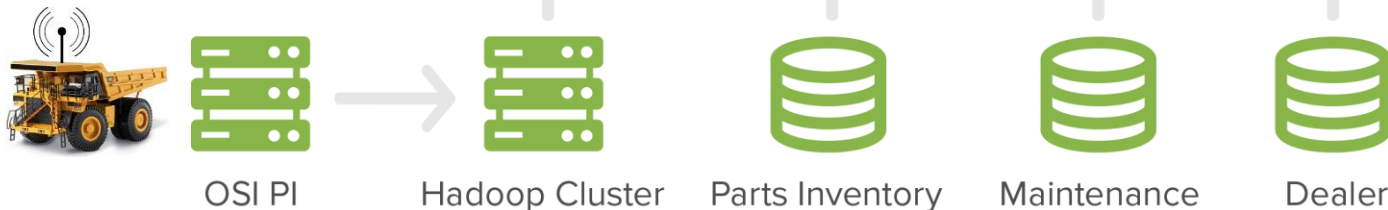- Or, maintain high price, but differentiate in other ways?

Large Heavy Equipment Manufacturer

Self-service / Predictive Analytics – IoT Integration

Tableau: Dealer / Customer Dashboard

denodo DATA VIRTUALIZATION

OSI PI   Hadoop Cluster   Parts Inventory   Maintenance   Dealer

Benefits

- Improved asset performance and proactive maintenance

- Increased revenue from sale of services and parts

- Reduced warranty costs of parts failure

denodo

"Big Data Challenges Impede Big Data Vision"

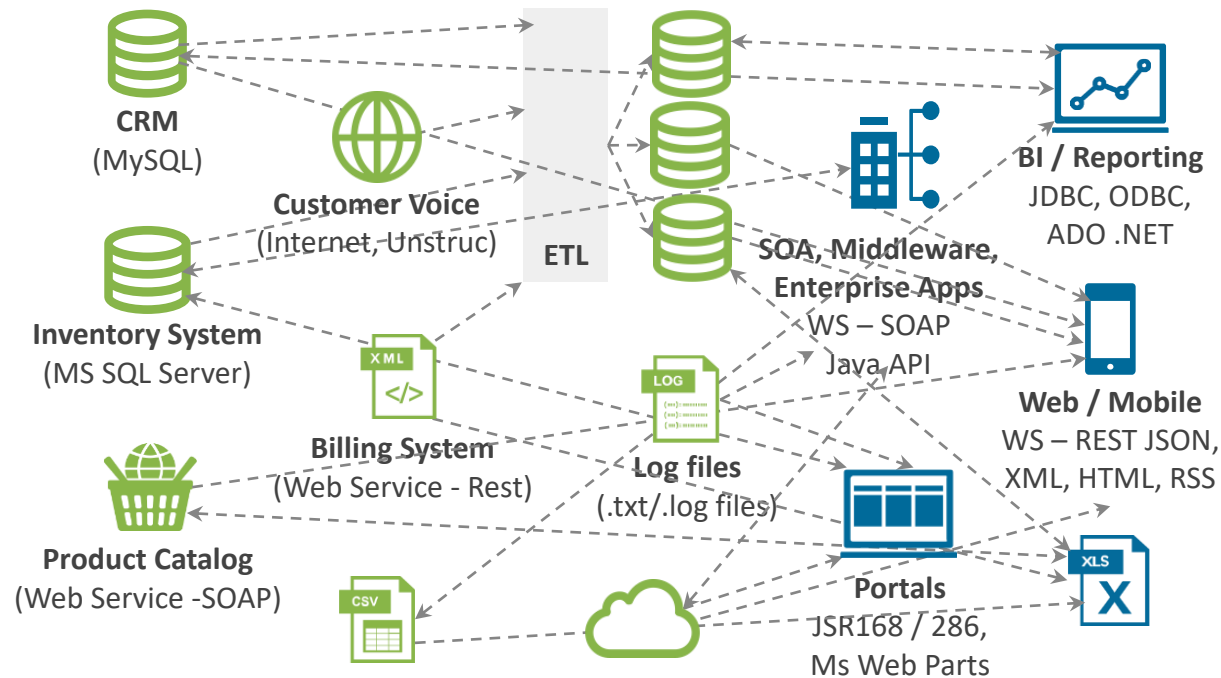| Challenge	Percentage
Lack of business competency to deal with data that is messy, diverse, or large	20%
Maturity of technology around security	20%
Lack of technology skills	19%
Legal and compliance issues	19%
Data silos created by management and organizational challenges	**18%**
Organizational business issues with data stewardship and governance	17%
Maturity of technology around governance and metadata management	16%
Maturity of technology around preparing messy, diverse, and large data	16%

denodo

IT – Business Dilemma

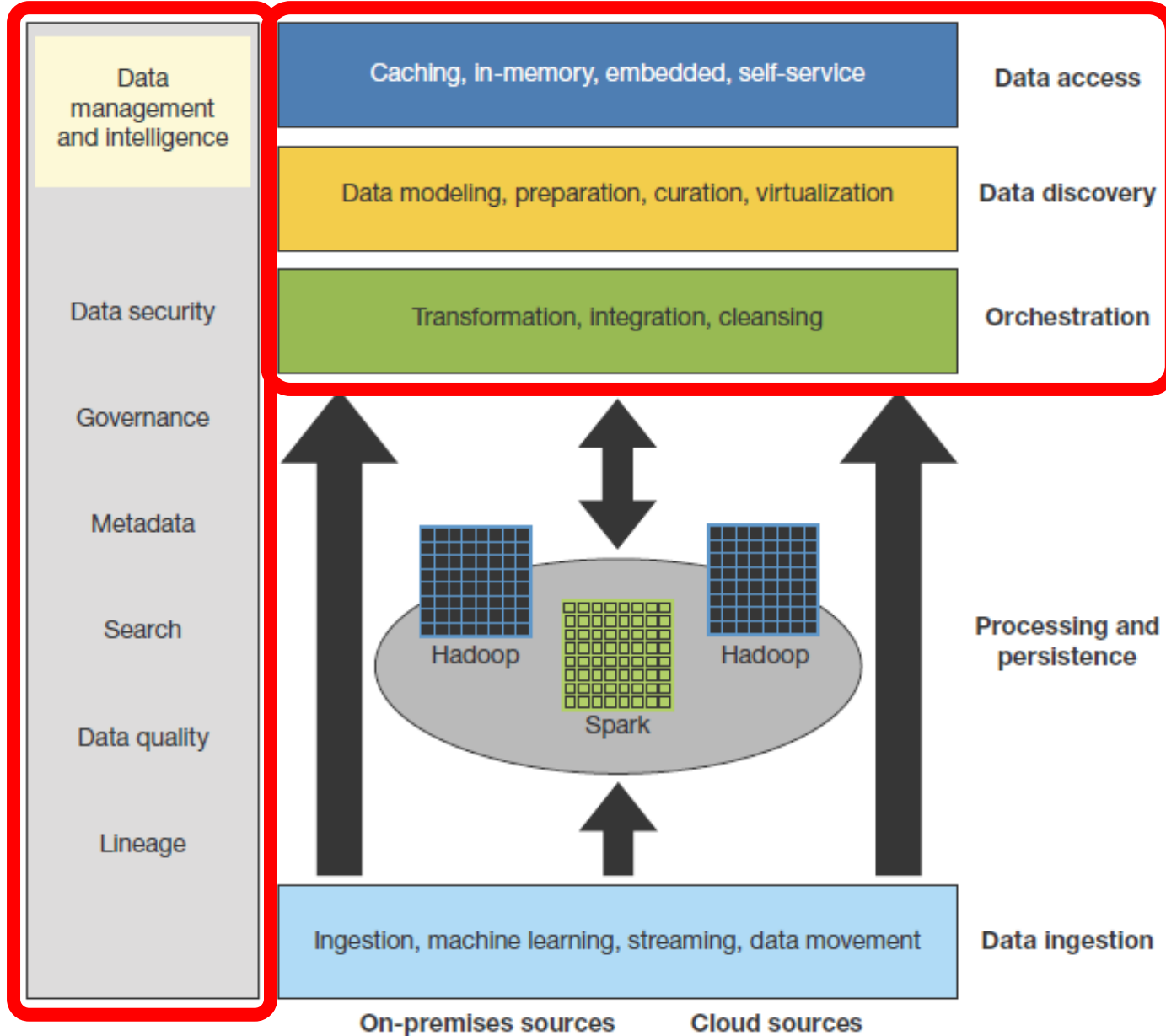IT Architecture is Unmanageable & Brittle because:

IT Focuses on Data Collection & Storage

No One Focused on Data Delivery

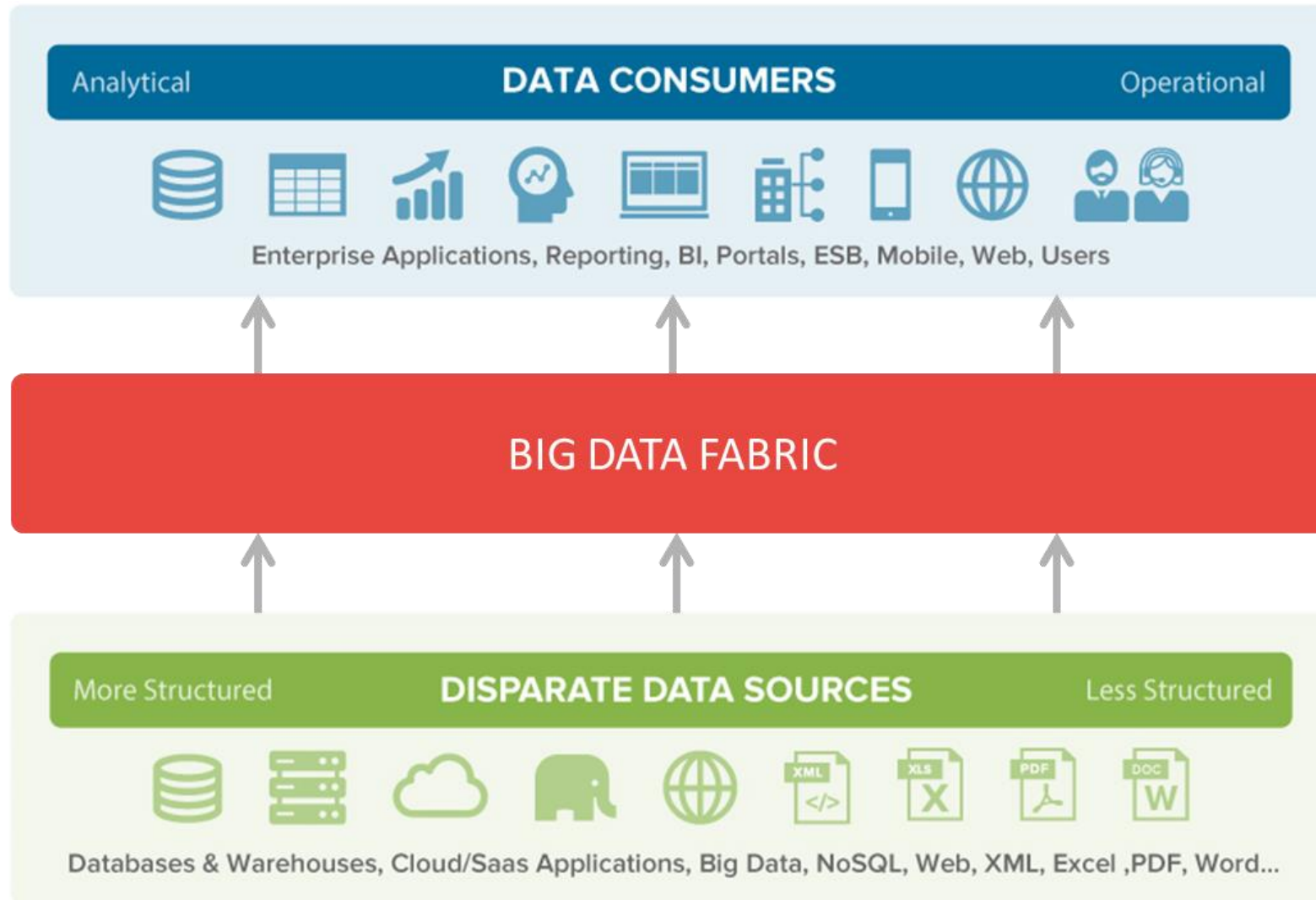– So create 100's to 1K's of brittle direct connections and replicate large volumes of data

CRM (MySQL)

Customer Voice (Internet, Unstruc)

Inventory System (MS SQL Server)

Billing System (Web Service - Rest)

Product Catalog (Web Service -SOAP)

ETL

Log files (.txt/.log files)

SOA, Middleware, Enterprise Apps WS – SOAP Java API

Portals JSR168 / 286, Ms Web Parts

BI / Reporting JDBC, ODBC, ADO .NET

Web / Mobile WS – REST JSON, XML, HTML, RSS

Business Focuses on Data Visualization & Analysis

denodo

Big Data Fabric Architecture – Forrester Research, 2016

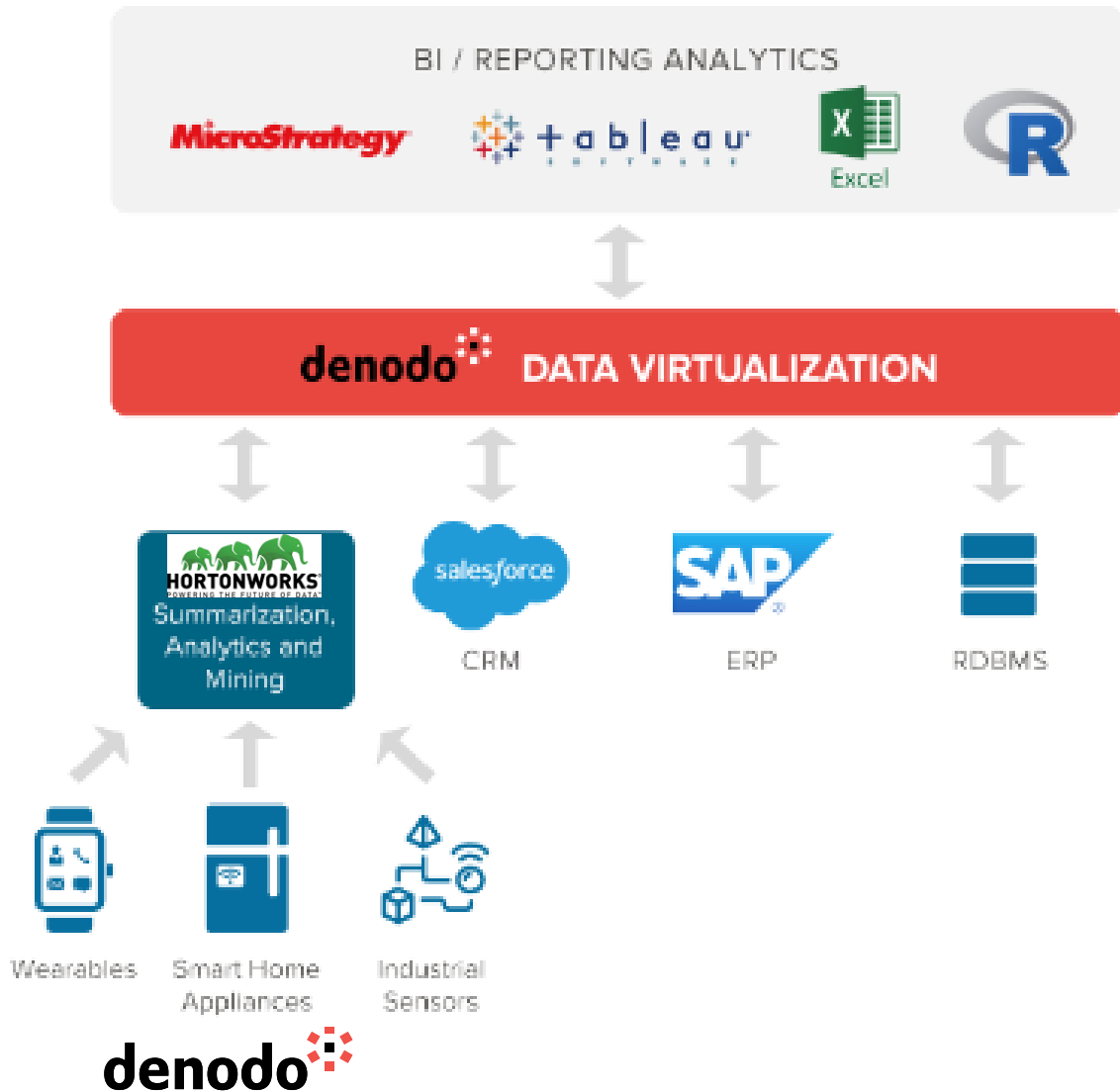| Data management and intelligence		
	Caching, in-memory, embedded, self-service	Data access
Data security	Data modeling, preparation, curation, virtualization	Data discovery
Governance	Transformation, integration, cleansing	Orchestration
Metadata		
Search	Hadoop Spark Hadoop	Processing and persistence
Data quality		
Lineage		
	Ingestion, machine learning, streaming, data movement	Data ingestion
	On-premises sources Cloud sources	

Big Data Fabric – Data Abstraction Layer

Abstracts access to **disparate data sources**

Acts as a **single repository** (virtual)

Makes data available in **real-time** to consumers

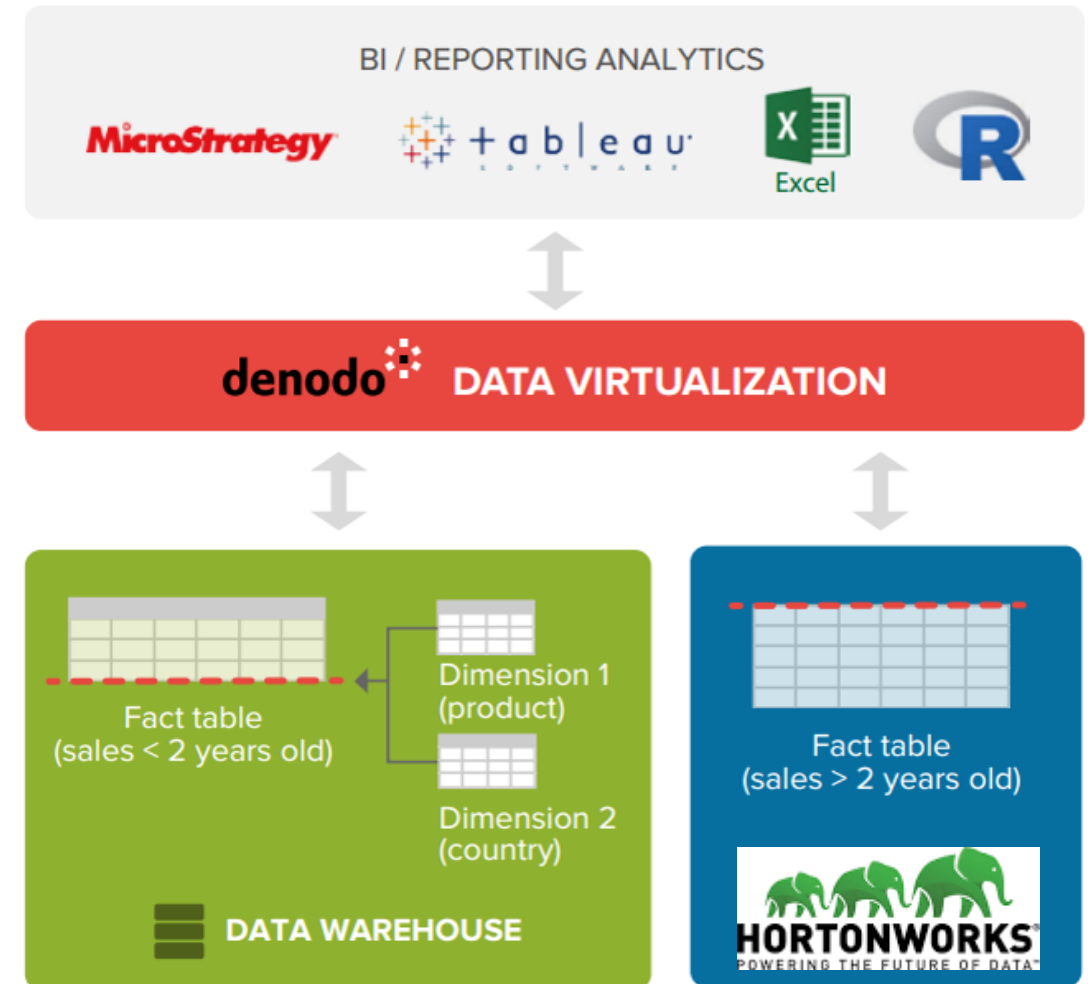# 3 DATA CONSUMERS

Consume in business applications

| Analytical | **DATA CONSUMERS** | Operational |

Enterprise Applications, Reporting, BI, Portals, ESB, Mobile, Web, Users, IoT/Streaming Data

| Multiple protocols, formats | Linked data services query, search, browse | Request/Reply, event driven | Secure delivery |

denodo DATA VIRTUALIZATION

2

Combine related data into views

Agile Development	**CONNECT** Normalized views of disparate data	**COMBINE** Discover, Transform, Prepare, Improve Quality, Integrate	**CONSUME** Share, Deliver, Publish, Govern, Collaborate	Governance & Metadata
Performance				Data Catalog
Resource Management				Security & Data Privacy
Lifecycle Management				Data Services

| Library of wrappers | Web automation | Any data or content | Read & Write |

1

Connect to disparate data sources

| More Structured | **DISPARATE DATA SOURCES** | Less Structured |

Databases & Warehouses, Cloud/Saas Applications, Big Data, NoSQL, Web, XML, Excel, PDF, Word...

denodo

8

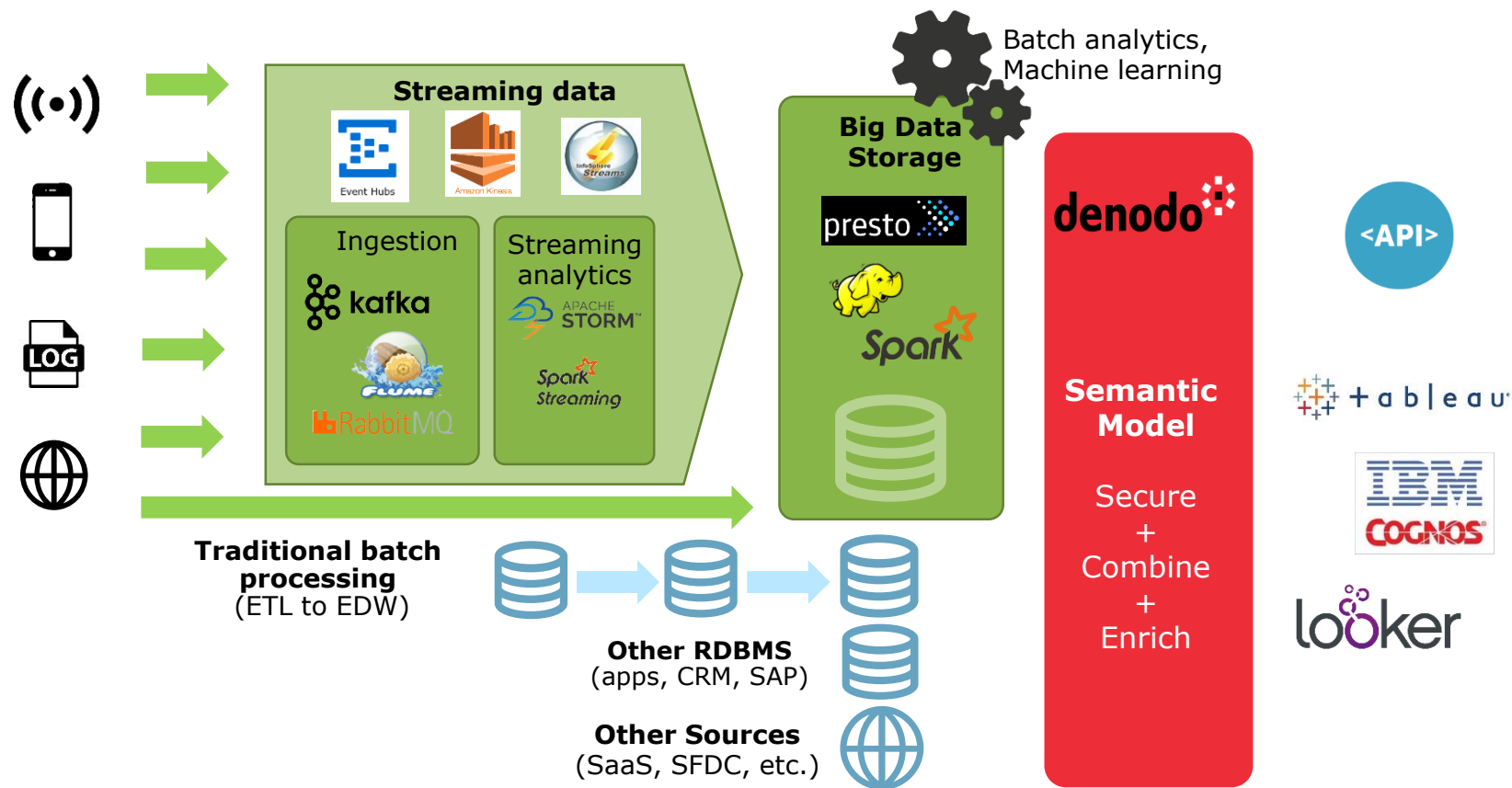# Logical Data Lake – Use Cases
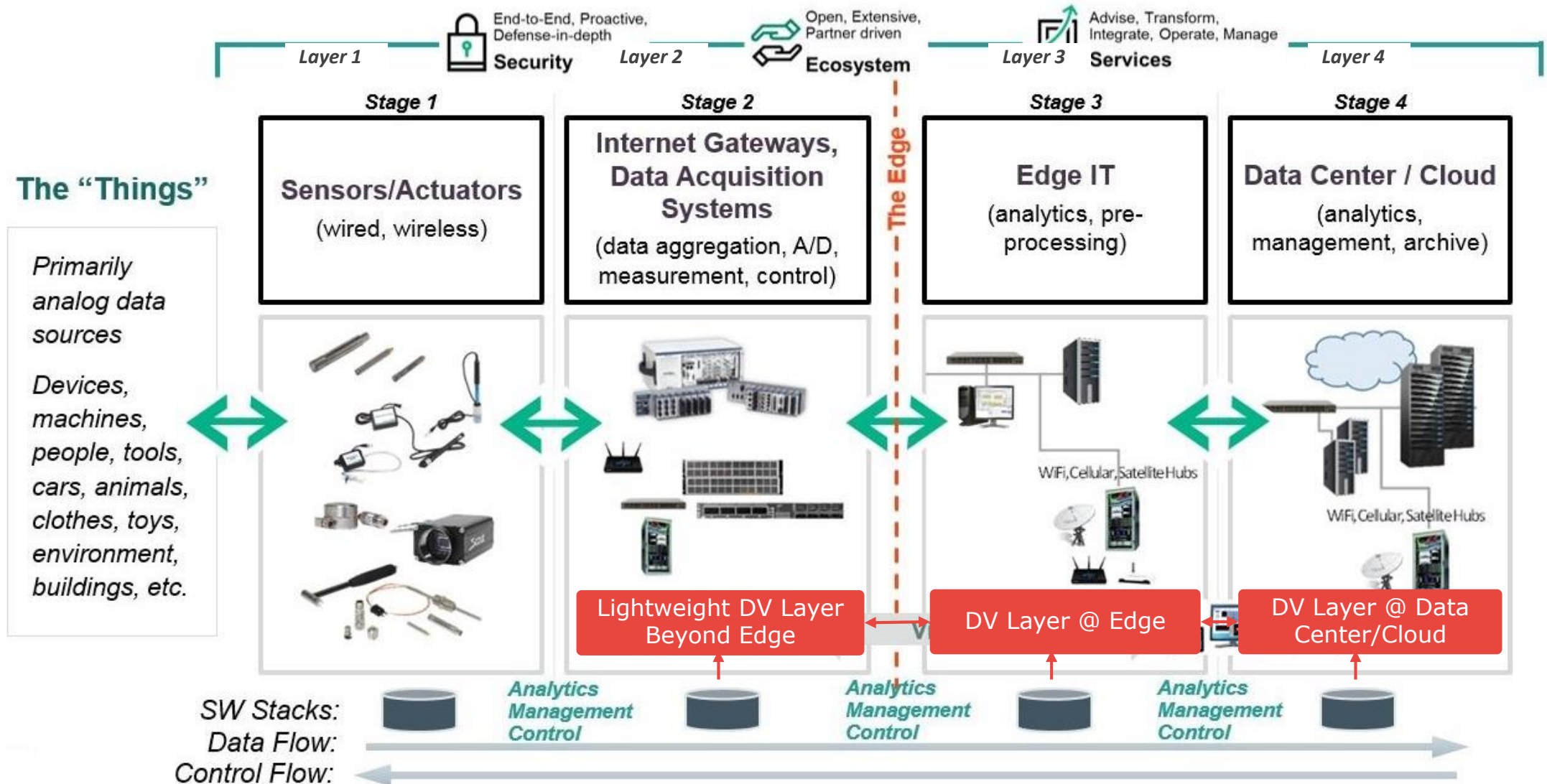

IoT Integration

Data Warehouse Offloading

Data Virtualization in the IoT Ecosystem

Edge Computing Using Data Virtualization

Six Essential Capabilities of Data Virtualization

1. Data abstraction

2. Zero replication, zero relocation

3. Real-time information

4. Self-service data services

4. Centralized metadata, security & governance

4. Location-agnostic architecture for multi-cloud, hybrid acceleration

denodo

Big Data Queries Faster with Denodo Platform

Performance comparison of 5 different queries

1. Data Virtualization delivers better performance without needing 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**	
Query 2	187	**96**	**88**	**Queries 1,2,3,5** •Exadata Row Count: ~5M •Impala Row Count: ~500k **Query 4** •Exadata Row Count: ~5M •Impala Row Count: ~2M
Query 3	120	212	**115**	
Query 4	timeout	**328**	**69**	
Query 5	**46**	91	56	

denodo

Denodo Dynamic Query Optimizer

```
SELECT c.id, SUM(s.amount) as total
FROM customer c JOIN sales s
ON c.id = s.customer_id
GROUP BY c.id
```
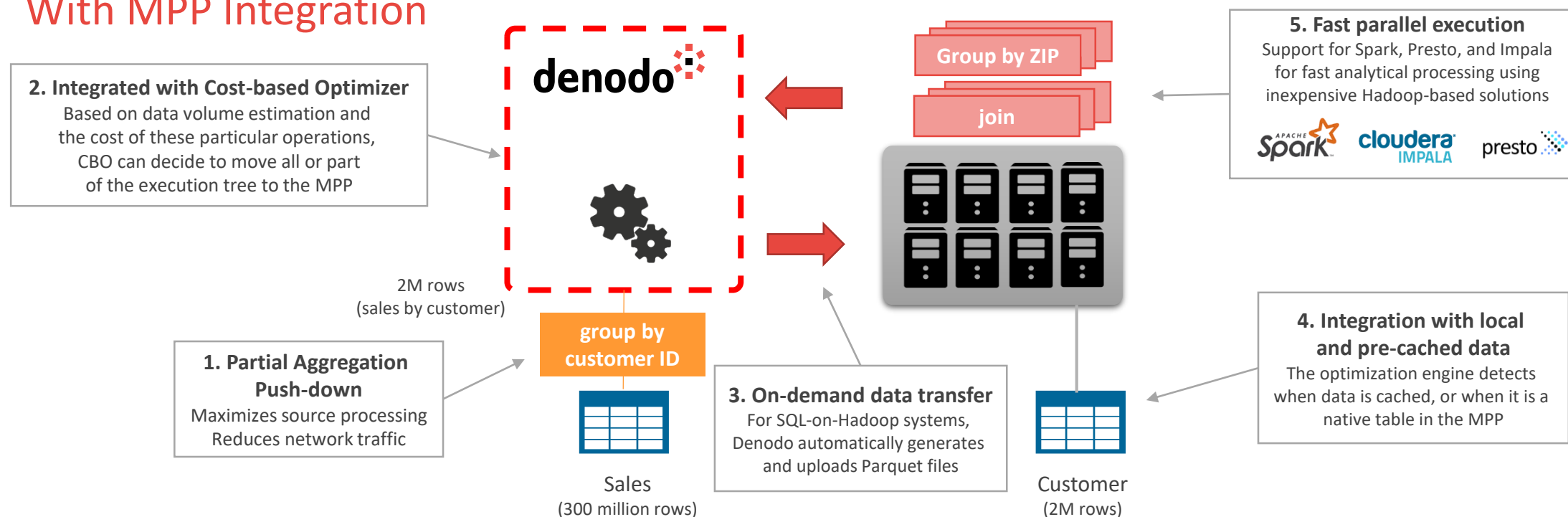
System	Execution Time	Data Transferred	Optimization Technique
Denodo	9 sec.	4 M	Aggregation push-down
Without Denodo	125 sec.	292 M	None: full scan
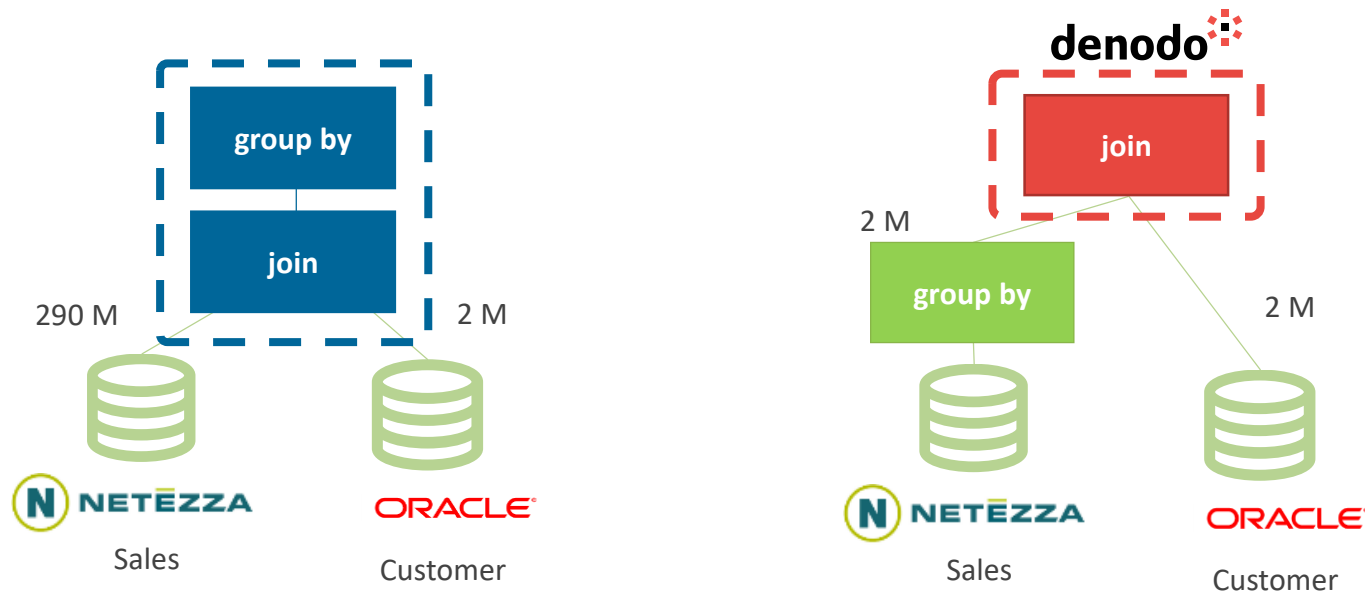
Data Virtualization with Massively Parallel Processing

With MPP Integration

2. Integrated with Cost-based Optimizer
Based on data volume estimation and the cost of these particular operations, CBO can decide to move all or part of the execution tree to the MPP

5. Fast parallel execution
Support for Spark, Presto, and Impala for fast analytical processing using inexpensive Hadoop-based solutions

1. Partial Aggregation Push-down
Maximizes source processing
Reduces network traffic

3. On-demand data transfer
For SQL-on-Hadoop systems, Denodo automatically generates and uploads Parquet files

4. Integration with local and pre-cached data
The optimization engine detects when data is cached, or when it is a native table in the MPP

2M rows
(sales by customer)

group by customer ID

Sales
(300 million rows)

Customer
(2M rows)

Group by ZIP

join

System	Execution Time	Optimization Techniques
Others	~ 10 min	Traditional Federation Tools
No MPP	43 sec	Aggregation push-down
With MPP	11 sec	Aggregation push-down + MPP integration (Impala 8 nodes)

ROI and TCO of Data Virtualization

Customer-reported projected savings by percentage

Data Integration Cost reduction

- **60-80%** savings

Traditional Call Centres, Portals

- **30-70%** savings

BI and Reporting

- **40-60%** savings

ETL and Data Warehousing

- Project timelines of 6-12 months reduced to **3-6 months**
- **Up to 85%** reduction in time

- New sources can be configured in minutes, and fully integrated within days.

- **100's** of application entities can be integrated within weeks.

- New business functionality can be added within days.

- Existing functionality can be enhanced with new data within days.

- Data proliferation can be significantly reduced.

- Common, consistent and timely access to all data via preferred visualization tools.

denodo

Denodo 'Horizontal Solution' Categories

Customer Centricity / MDM
- ✓ Complete View of Customer

Data Governance
- ✓ GRC
- ✓ GDPR
- ✓ Data Privacy / Masking

Data Services
- ✓ Data as a Service
- ✓ Data Marketplace
- ✓ Data Services
- ✓ Application and Data Migration

BI and Analytics
- ✓ Self-Service Analytics
- ✓ Logical Data Warehouse
- ✓ Enterprise Data Fabric

Big Data
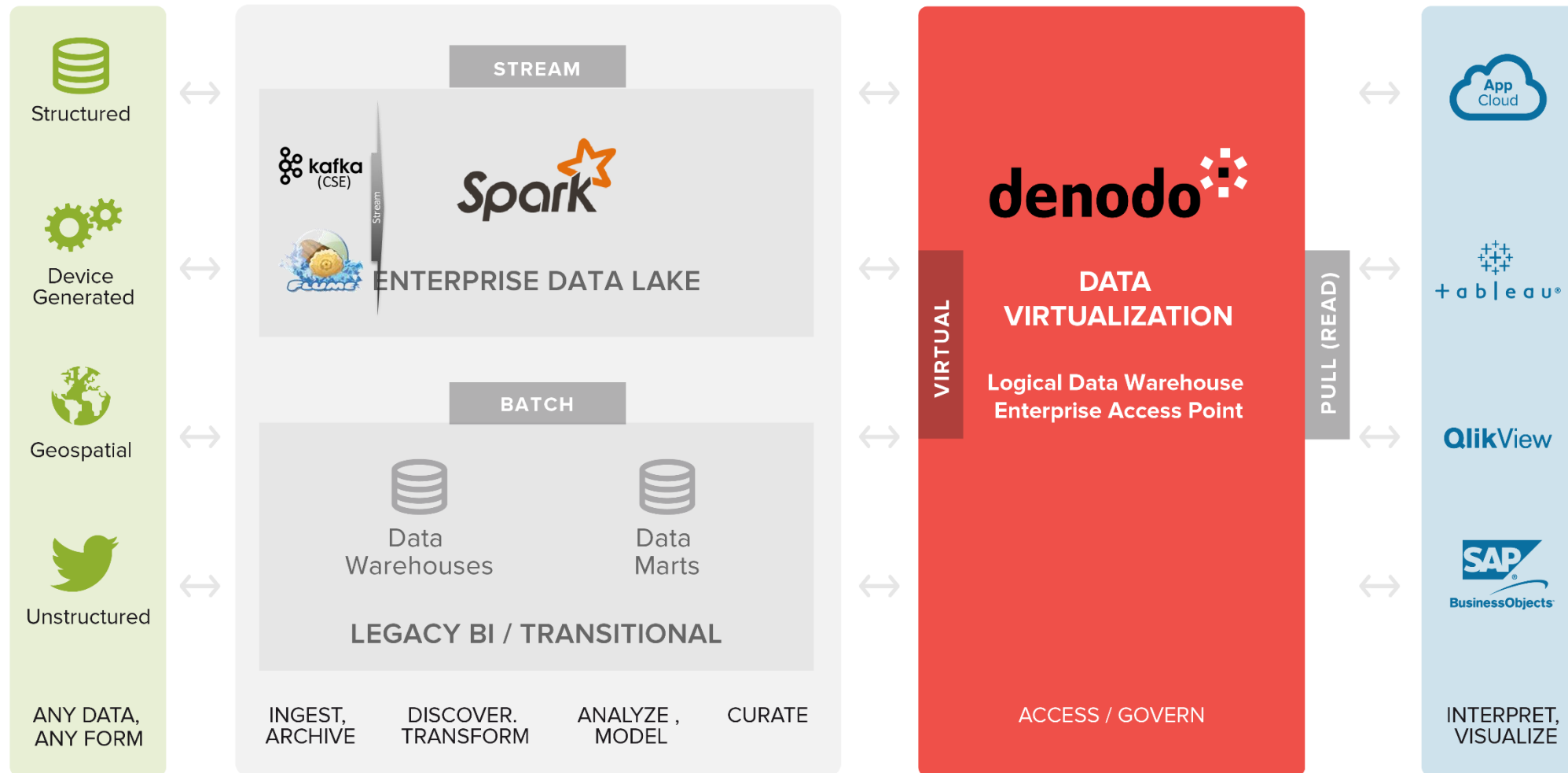- ✓ Logical Data Lake
- ✓ Data Warehouse Offloading
- ✓ IoT Analytics

Cloud Solutions
- ✓ Cloud Modernization
- ✓ Cloud Analytics
- ✓ Hybrid Data Fabric

denodo

Autodesk – Integration of Streaming Data

Asurion: Data Virtualization On-Premises and Cloud
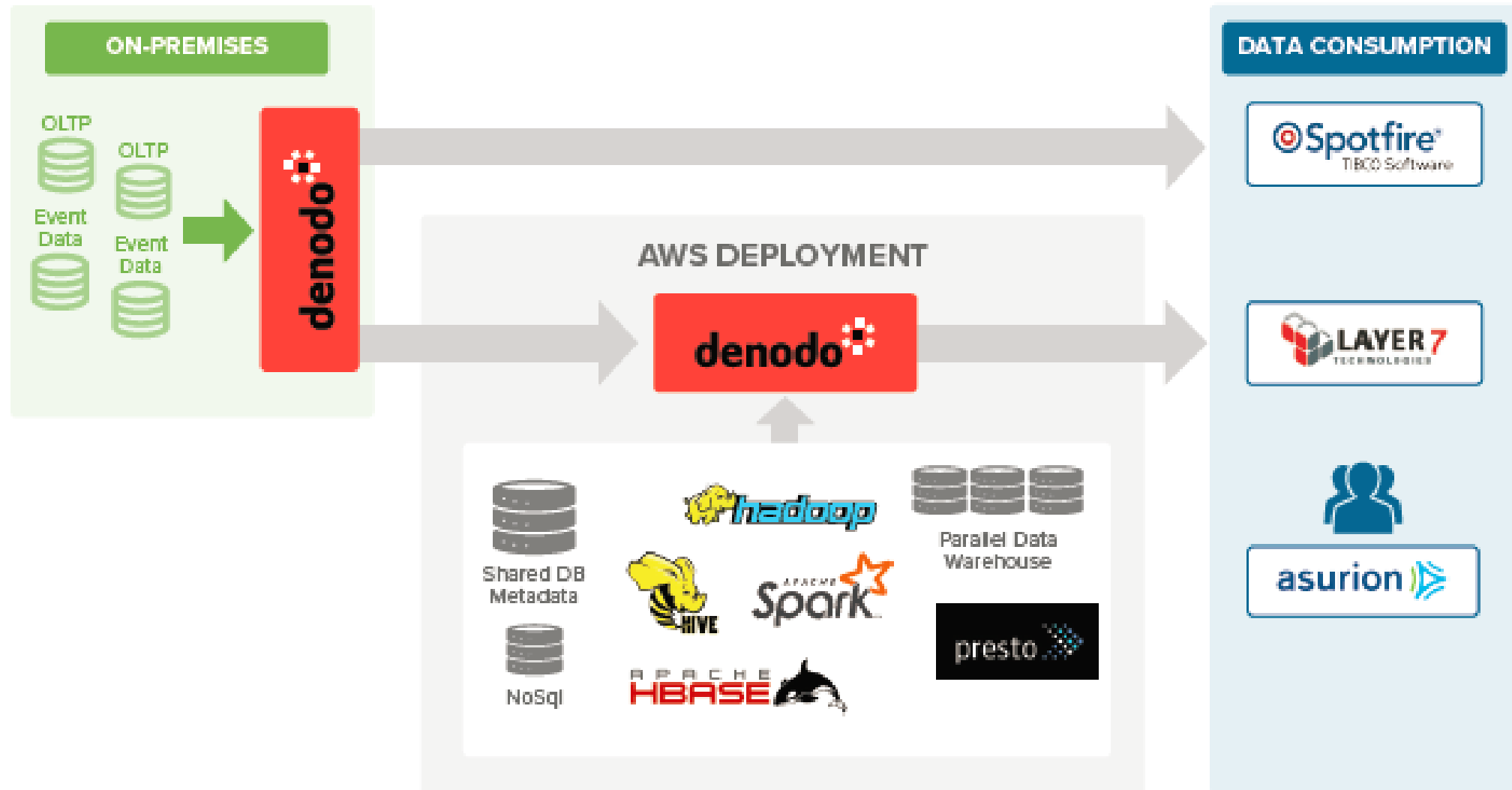
Denodo

The Leader in Data Virtualization

LEADERSHIP

- Longest continuous focus on data virtualization – since 1999
- Leader in 2018 Forrester Wave – Big Data Fabric
- Winner of numerous awards
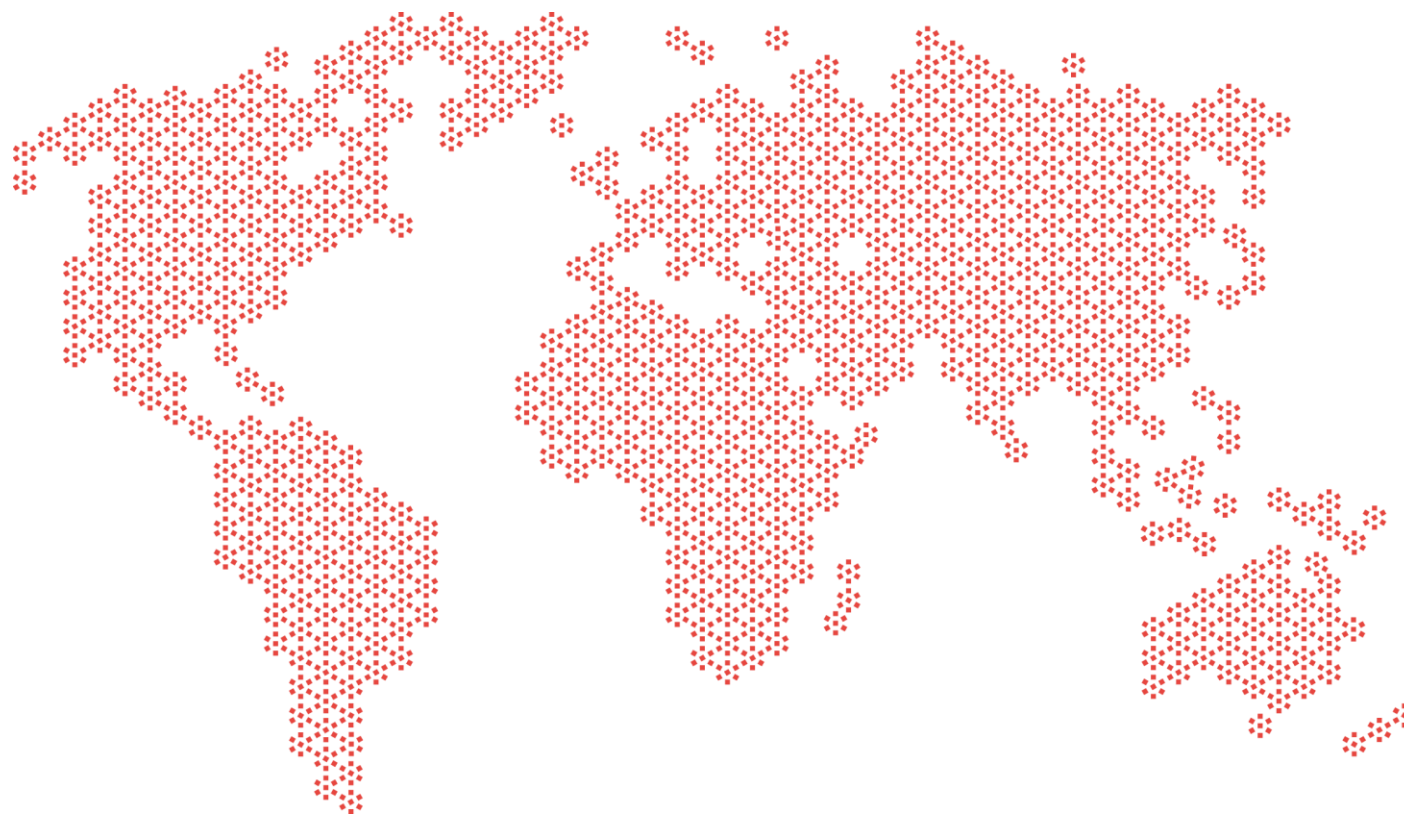

DENODO OFFICES, CUSTOMERS, PARTNERS
Palo Alto, CA.
Global presence throughout North America, EMEA, APAC, and Latin America.

CUSTOMERS
~500 customers, including many F500 and G2000 companies across every major industry have gained significant business agility and ROI.

FINANCIALS
Backed by $4B+ private equity firm.
50+% annual growth; Profitable.

denodo

www.denodo.com | info@denodo.com