denodo webinars

Designing Fast Data Architecture for Big Data using Logical Data Warehouse and Data Lakes

A Case Study presented by Kurt Jackson
Platform Lead, Autodesk

Speakers

Kurt Jackson

Platform Lead

AUTODESK

Ravi Shankar

Chief Marketing Officer

denodo

Agenda

1. Towards a Logical Data Lake – An Autodesk Case Study
2. Performance Considerations in Logical Data Warehouse/ Lakes
3. Q&A and Next Steps

Towards the Logical Data Lake

Kurt Jackson

Platform Lead

AUTODESK

What is a Logical Data Warehouse?

- A **logical data warehouse** is a data system that follows the ideas of traditional EDW (star or snowflake schemas) and includes, in addition to one (or more) core DWs, data from external sources.

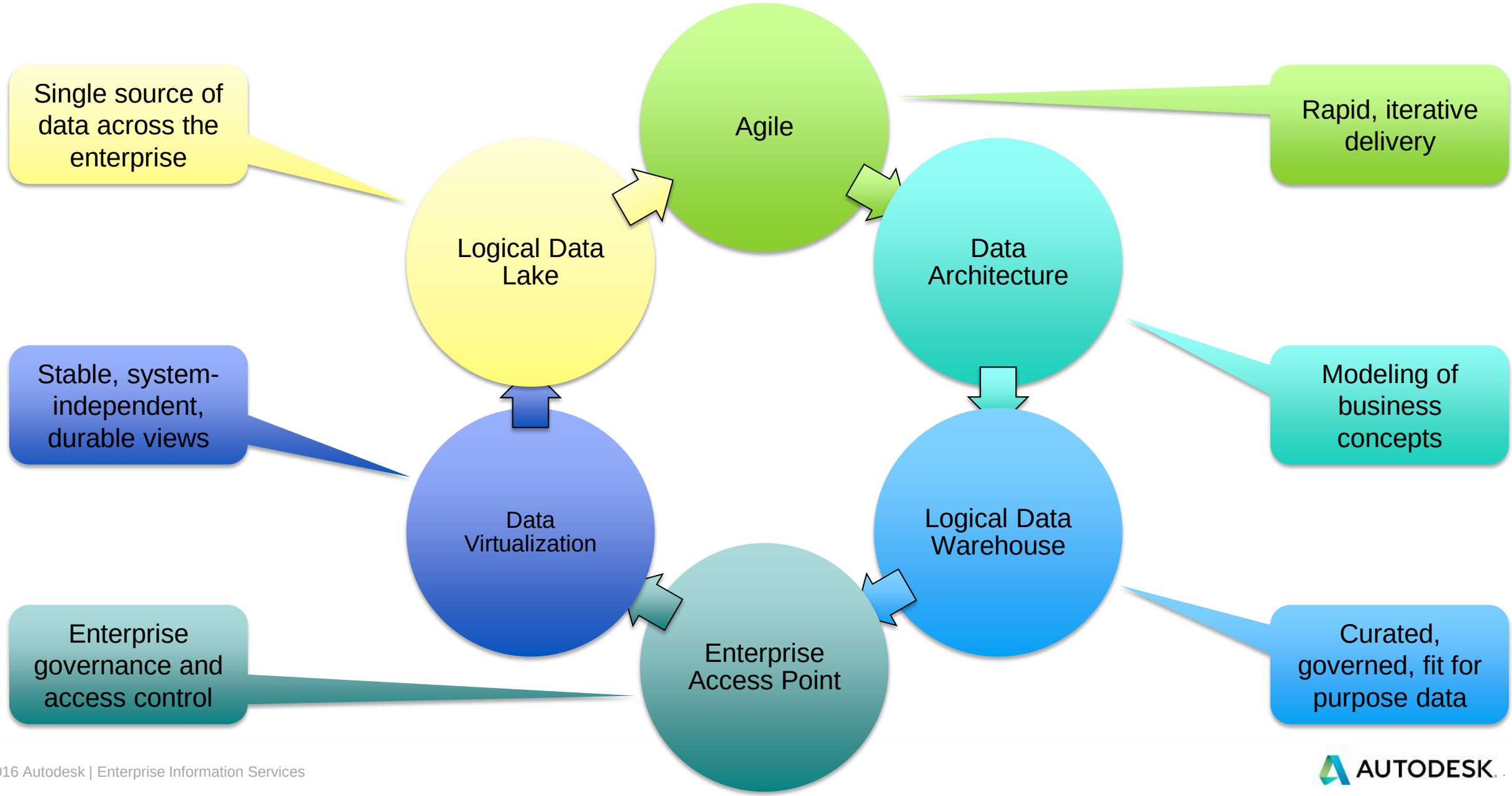
- The main motivations are improved decision making and/or cost reduction

What about the Logical Data Lake?

- A Data Lake will not have a star or snowflake schema, but rather a more heterogeneous collection of views with raw data from heterogeneous sources
- The virtual layer will act as a common umbrella under which these different sources are presented to the end user as a single system
- However, from the virtualization perspective, a Virtual Data Lake shares many technical aspects with a LDW and most of these contents also apply to a Logical Data Lake

Introduction

Agile Data Architecture Lifecycle



AUTODESK.

Business Problem

The Data-Driven Enterprise

Information Management Foundation

(Data)

- Data Governance
- EIM
- Information-driven enterprise

Organization

(People)

- **IT** – Responsible for information mgmt & provisioning
- **Business** – Responsible for analytics & acting on outcomes

Fact-based Decision Making

(Process)

- Info as an asset
- Instinct vs. analytics

Appropriate Technology Platform

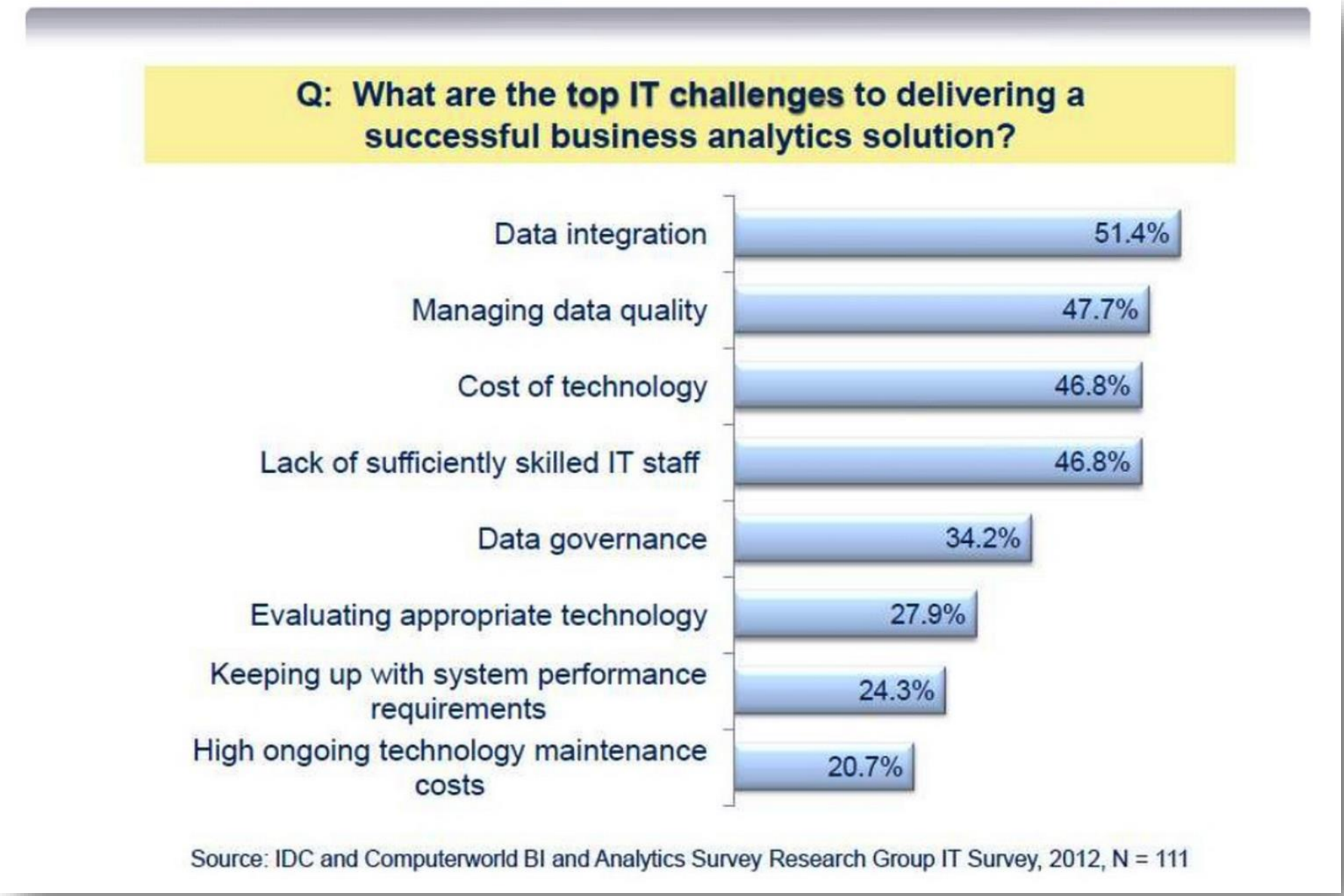
(Technology)

- Integrated toolset & delivery platform
- Not silos of capability

Gartner

4

Most of us are in the same boat

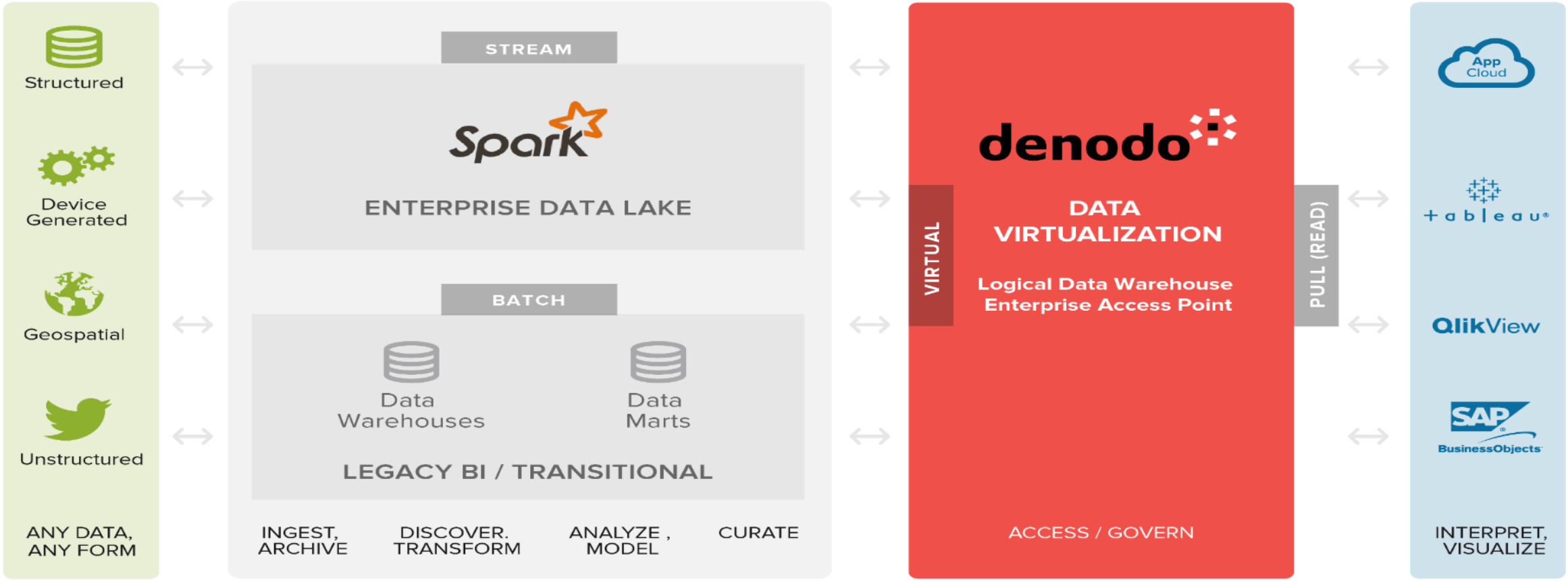


The Autodesk Agile Data Architecture

Philosophy

- Access and refine data near the source
- Published logical data interfaces
- Implementing interfaces is only an IT concern
- Agile and opportunistic retirement of legacy systems

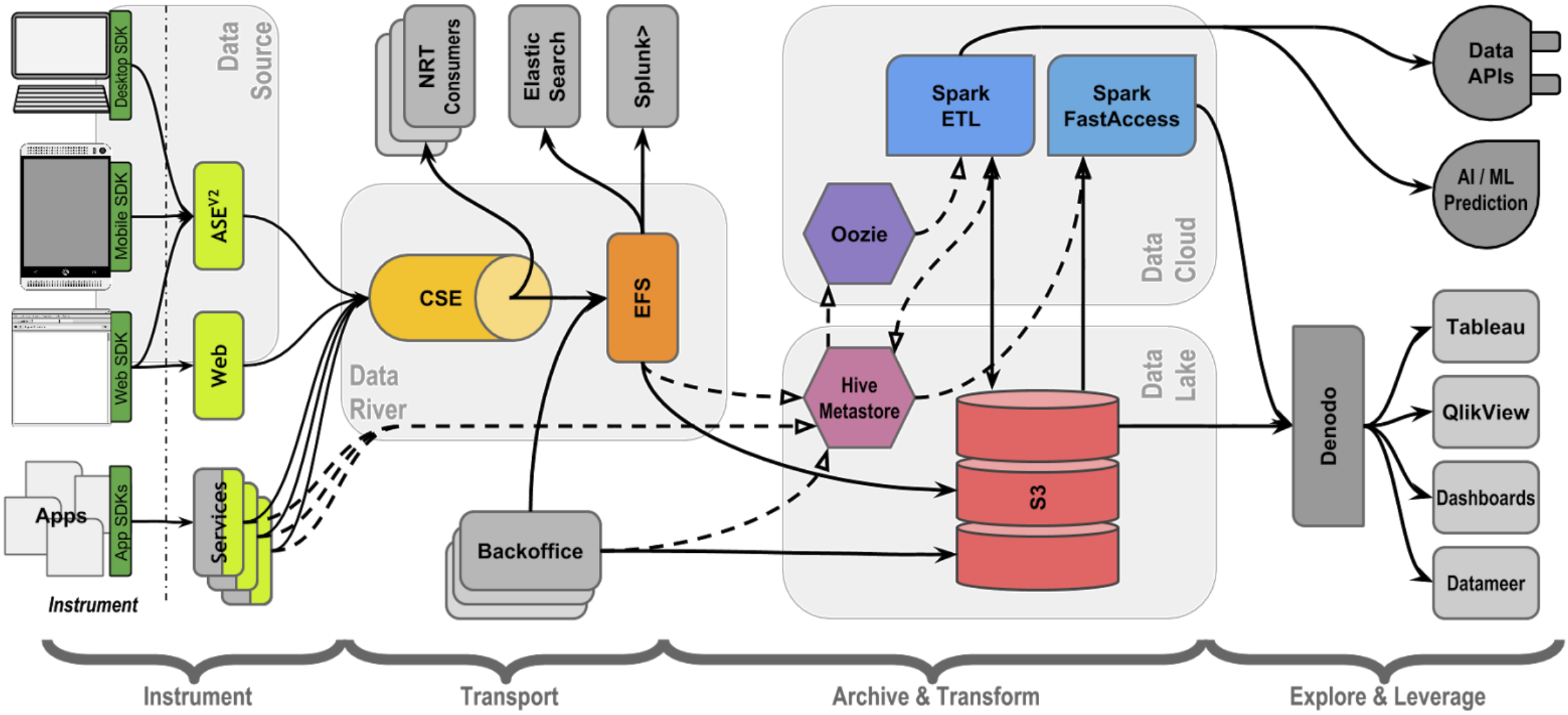
Autodesk Data Architecture



The journey to the Logical Data Lake

Data virtualization can be used throughout your data pipeline!

Big Data Ecosystem



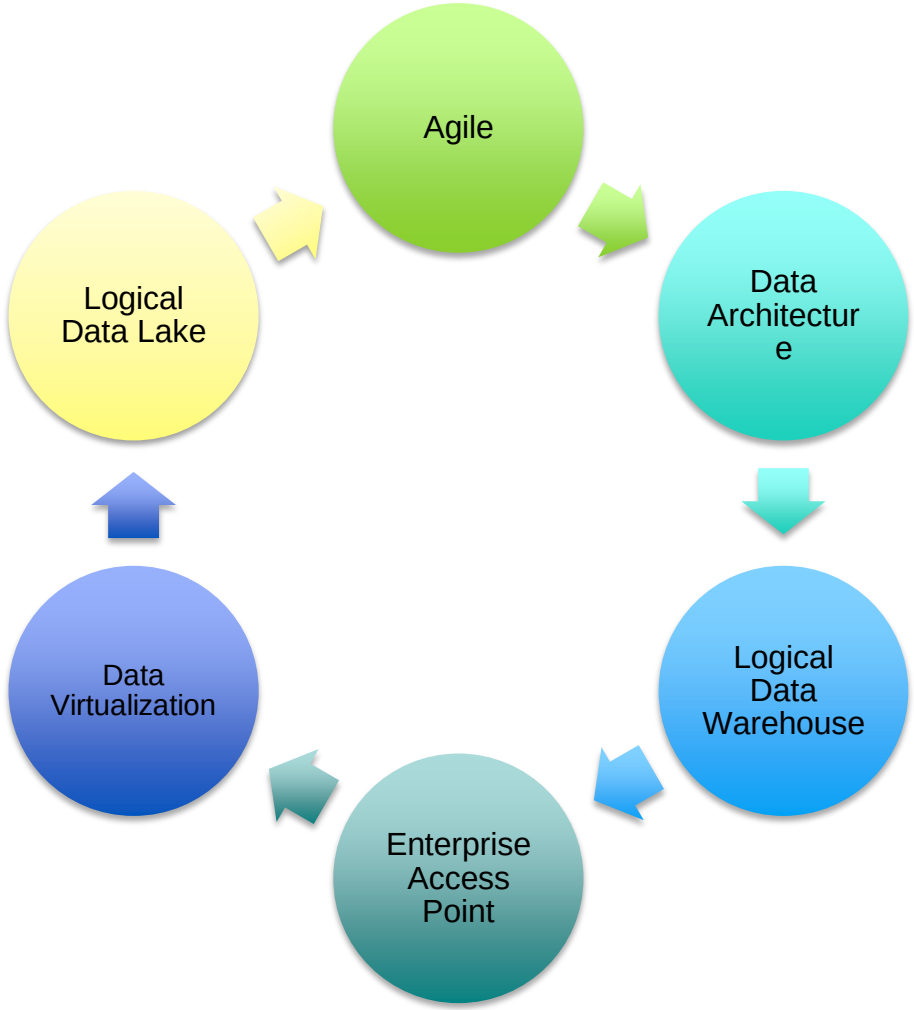
AUTODESK.

Towards the Logical Data Lake

AUTODESK

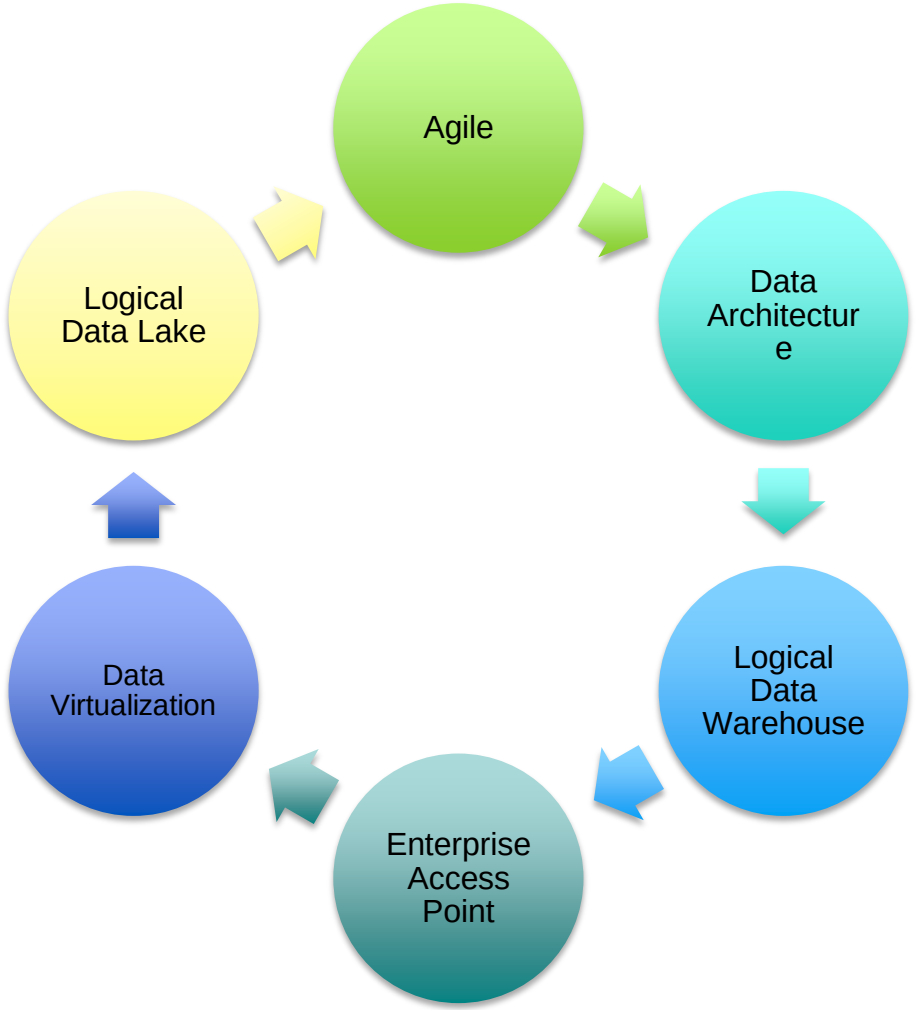
Logical Data Warehouse (LDW)

- Usability
  - Single repository for schema definitions
- Integrity
  - Only published views are publically available
  - Business ownership guarantees the quality



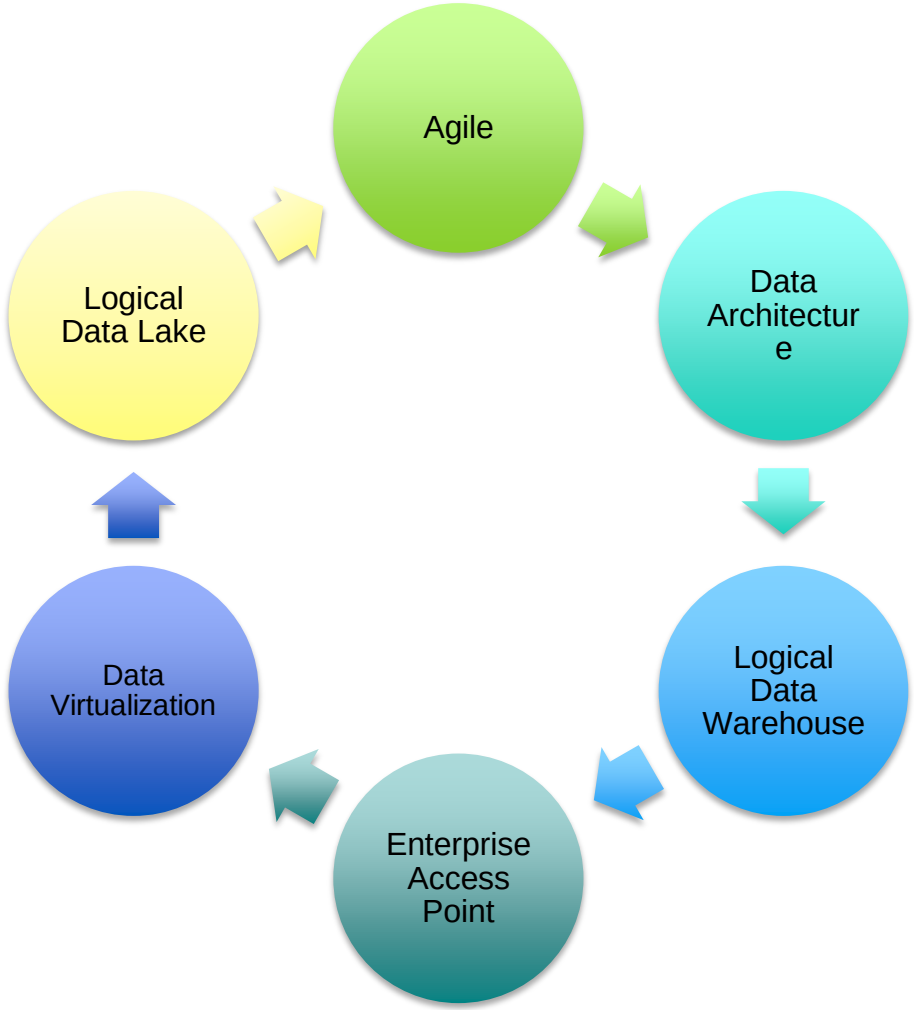
The Enterprise Access Point

- Availability
  - A single access point to administer
- Security
  - A single point for authentication, authorization and audit trail



Data Virtualization

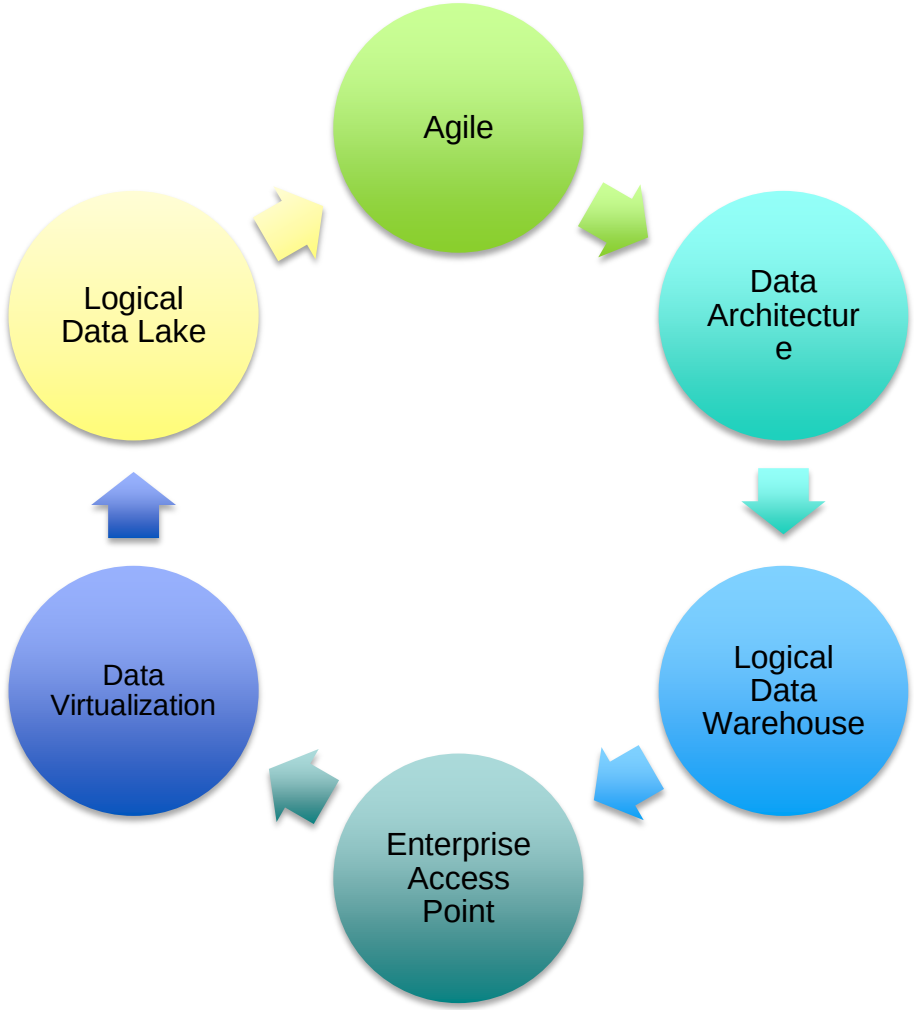
- System-agnostic
  - Data contract independent of implementation
- Enables the agile enterprise
  - Aggressively and opportunistically refactor enterprise systems
- Manifest support for federation and data blending



AUTODESK.

Logical Data Lake

- Beyond the traditional Enterprise Data Lake
  - Combine traditional data lakes with other data
- Blending of disparate, heterogeneous data sources
  - Internal, external, 3rd party…
- Approaches a true, single origin for all enterprise data



Towards the Logical Data Lake implements the philosophy

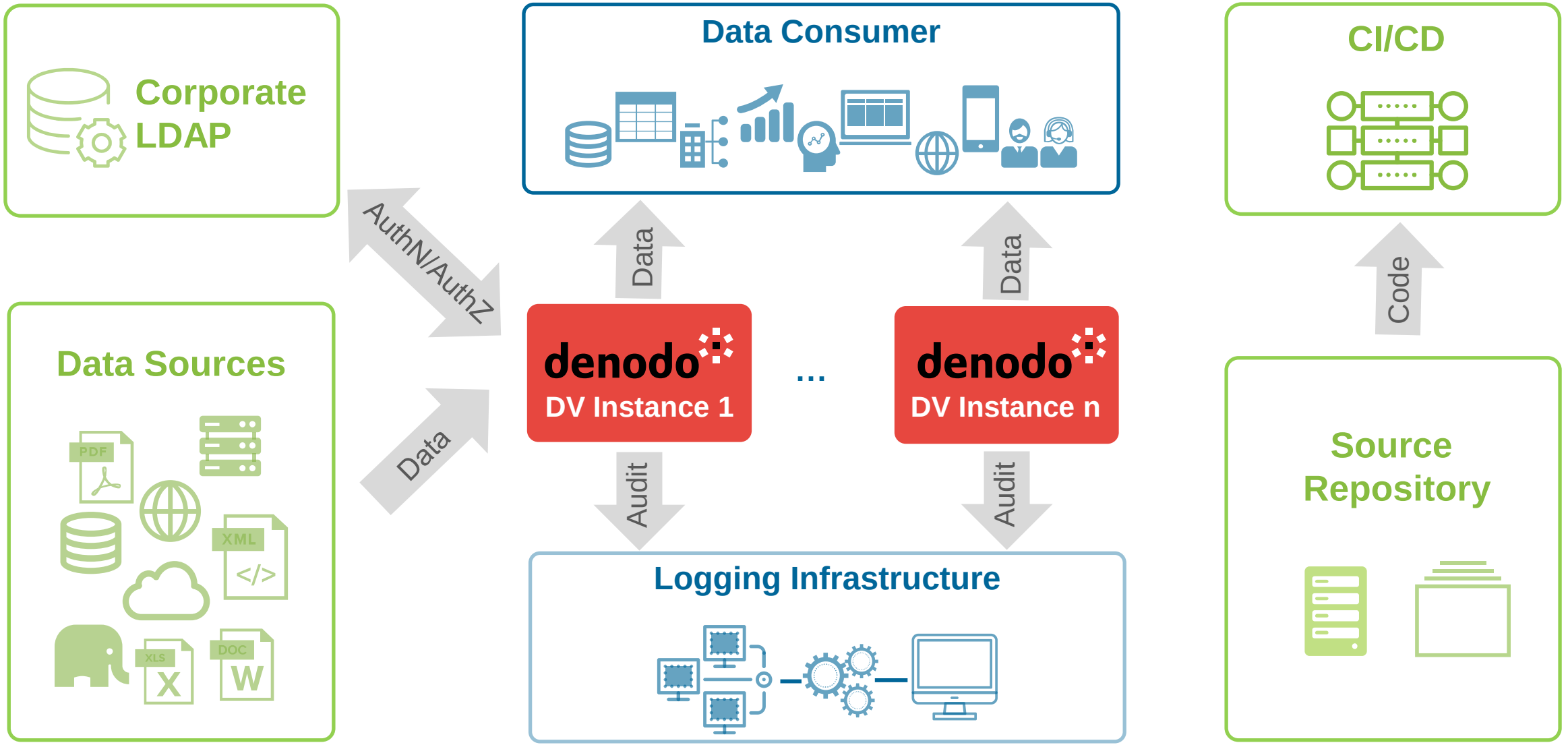
- Access and refine data near the source
  - No painful ETL pipelines for data derivation
- Published logical data interfaces
  - Single access point for all of external data sets
- Implementing interfaces is only an IT concern
  - Accelerate to best of breed solutions
- Agile and opportunistic retirement of legacy systems
  - Rip and replace becomes almost transparent

Building the Agile Data Architecture at Autodesk

Implementation Approach

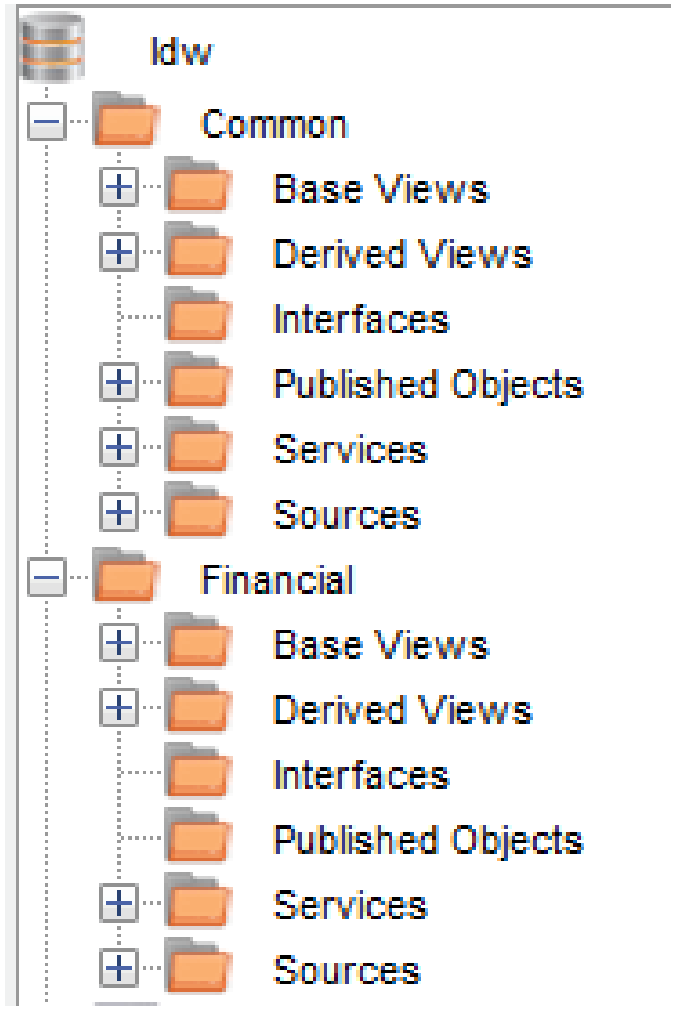
- Identify enterprise data sources
  - Harder than you think
- All new streaming, highly-available ingestion mechanism
  - Self-service or nearly so
  - Stream-based facilitates batch and streaming data processing
- Leverage highly-redundant cloud storage for the data lake
  - S3
- Leverage best-of breed for individual components
  - Open source, selected commercial vendors
- Develop canonical representations for your data sets
  - Very difficult – consider a CDO
- Virtualize the data warehouses and marts with a next generation Logical DW
  - New implementations leverage the LDW
  - Legacy migrates opportunistically

Architecting the Data Virtualization Layer



Build an Information Architecture

- Base views to abstract data sources
- Layered derived views to reflect successively refined derivations
- Create the notion of publication for curated, externally visible views
- Expose services on top of views to make views more accessible
- Separate namespaces (schemas) by project or subject area
- Build the notion of commonality for views shared across schemas
- Naming conventions for all objects
- Data portal for one-stop shopping for data consumers



Towards the Logical Data Lake can be a liberating journey!

Performance Considerations in Logical Data Warehouse/ Lakes

Ravi Shankar, CMO
Denodo

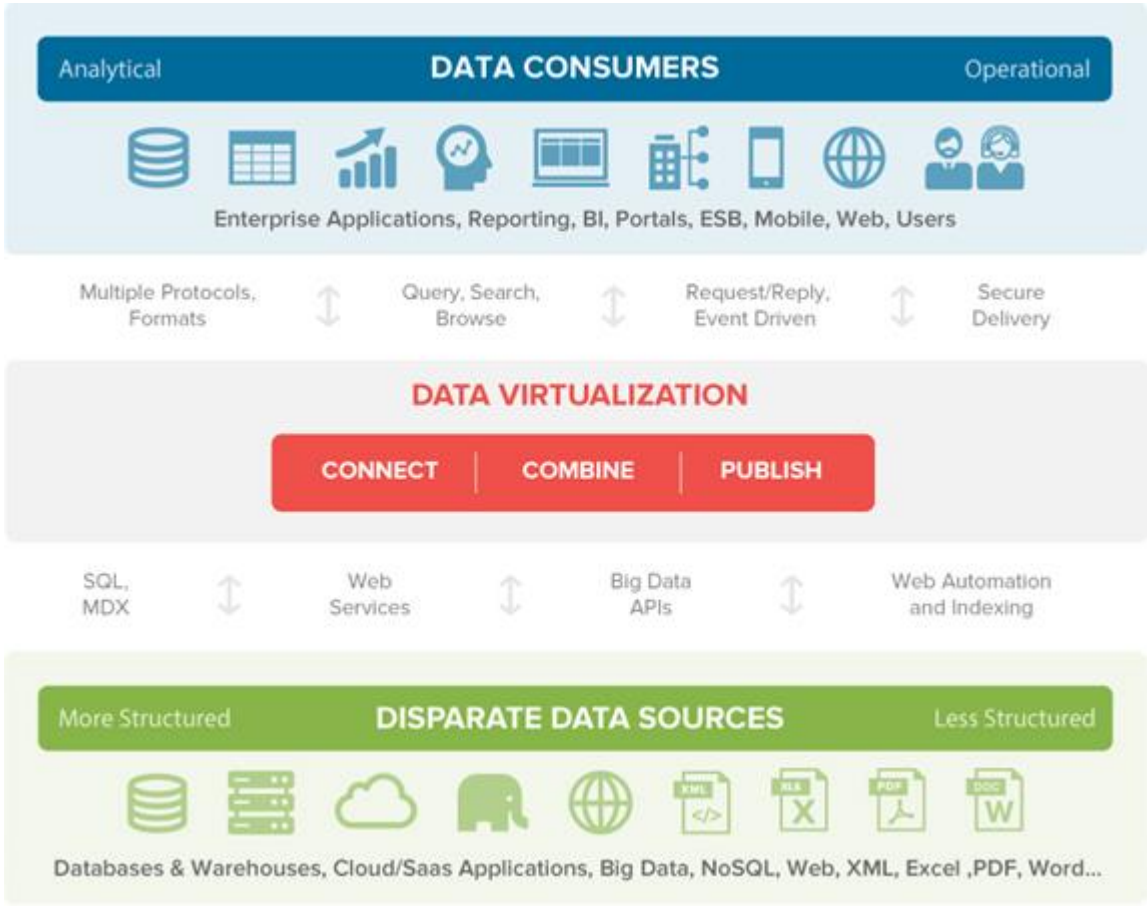
Query Optimization in Data Virtualization

Real-time Data Integration

Publishes the data to applications | 3

Combines related data into views | 2

Connects to disparate data sources | 1



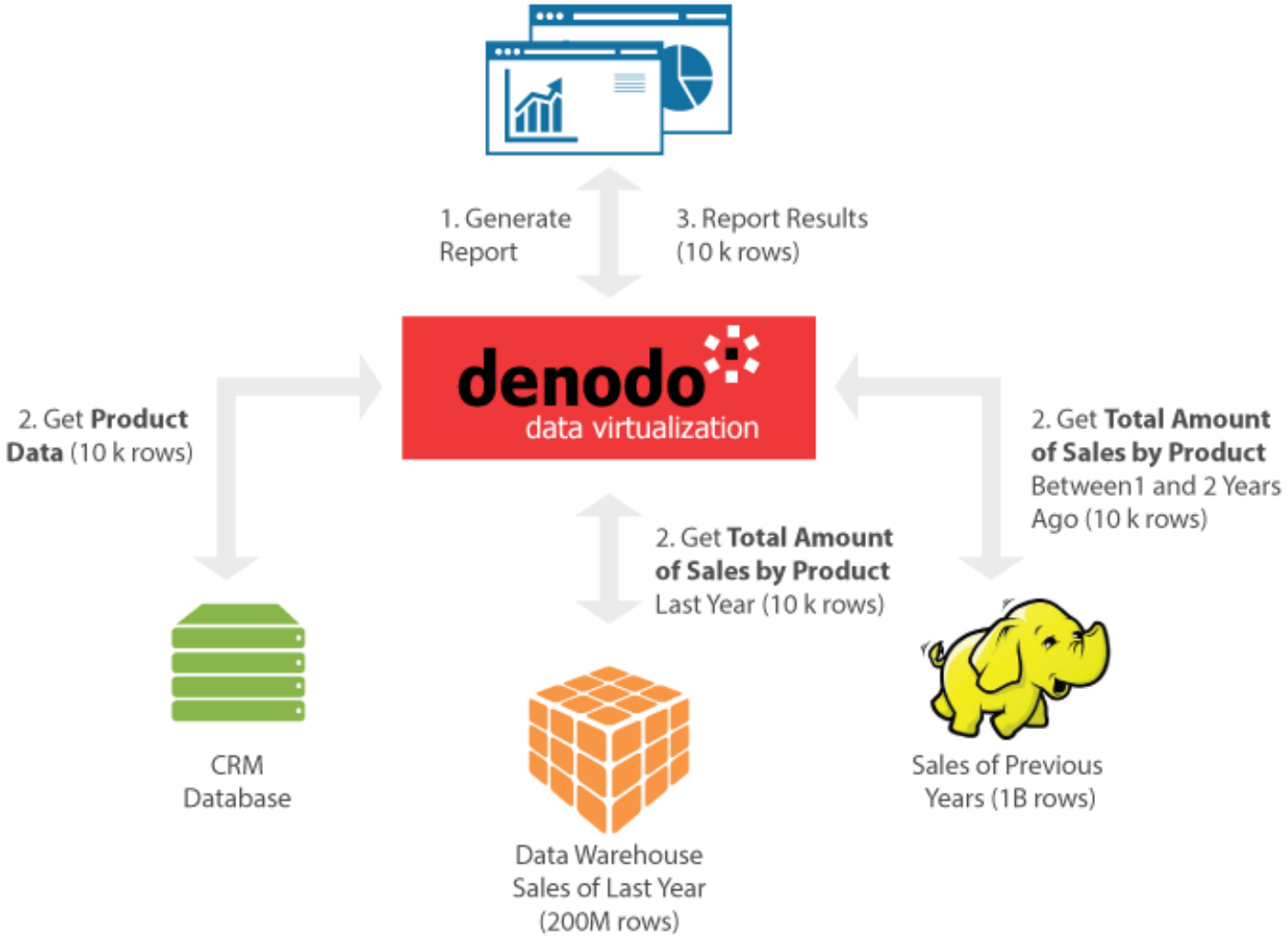
"Data virtualization integrates disparate data sources in real time or near-real time to meet demands for analytics and transactional data."

– Create a Road Map For A Real-time, Agile, Self-Service Data Platform, Forrester Research, Dec 16, 2015

denodo

Dynamic Query Rewriting for Big Data

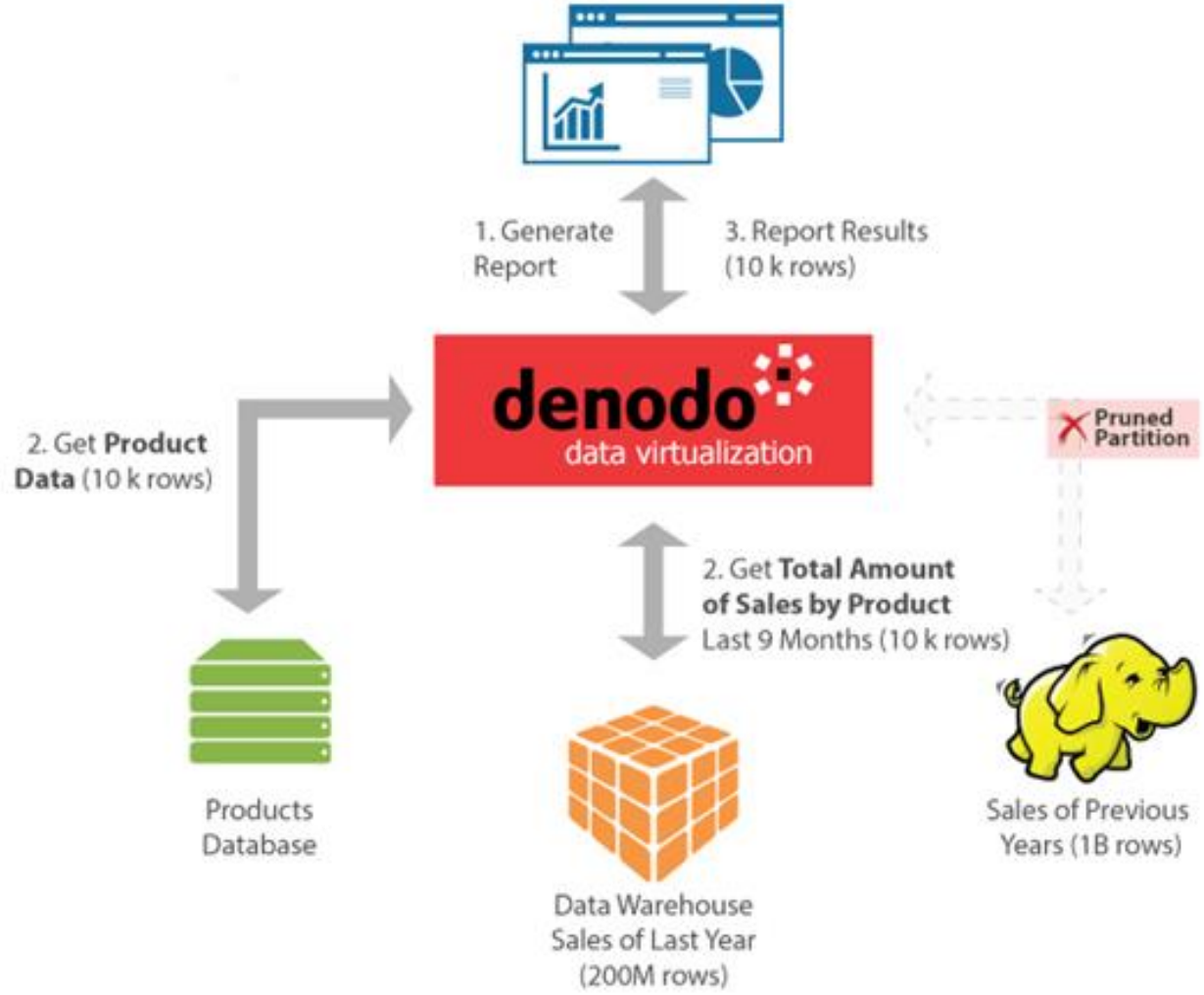
1. Full Aggregation Pushdown



Total sales for each product in the last 2 years

Dynamic Query Rewriting for Big Data

2. Partition Pruning



Total sales for each product in the last 9 months

Dynamic Query Rewriting for Big Data

3. Partial Aggregation Pushdown

1. Products and their Categories – Products DB

Product ID	Product Category
1	Phone
2	Phone
3	Computer

2. Total Sales by Product – Data Warehouse

Product ID	Total Sales
1	1000
2	1500
3	1200

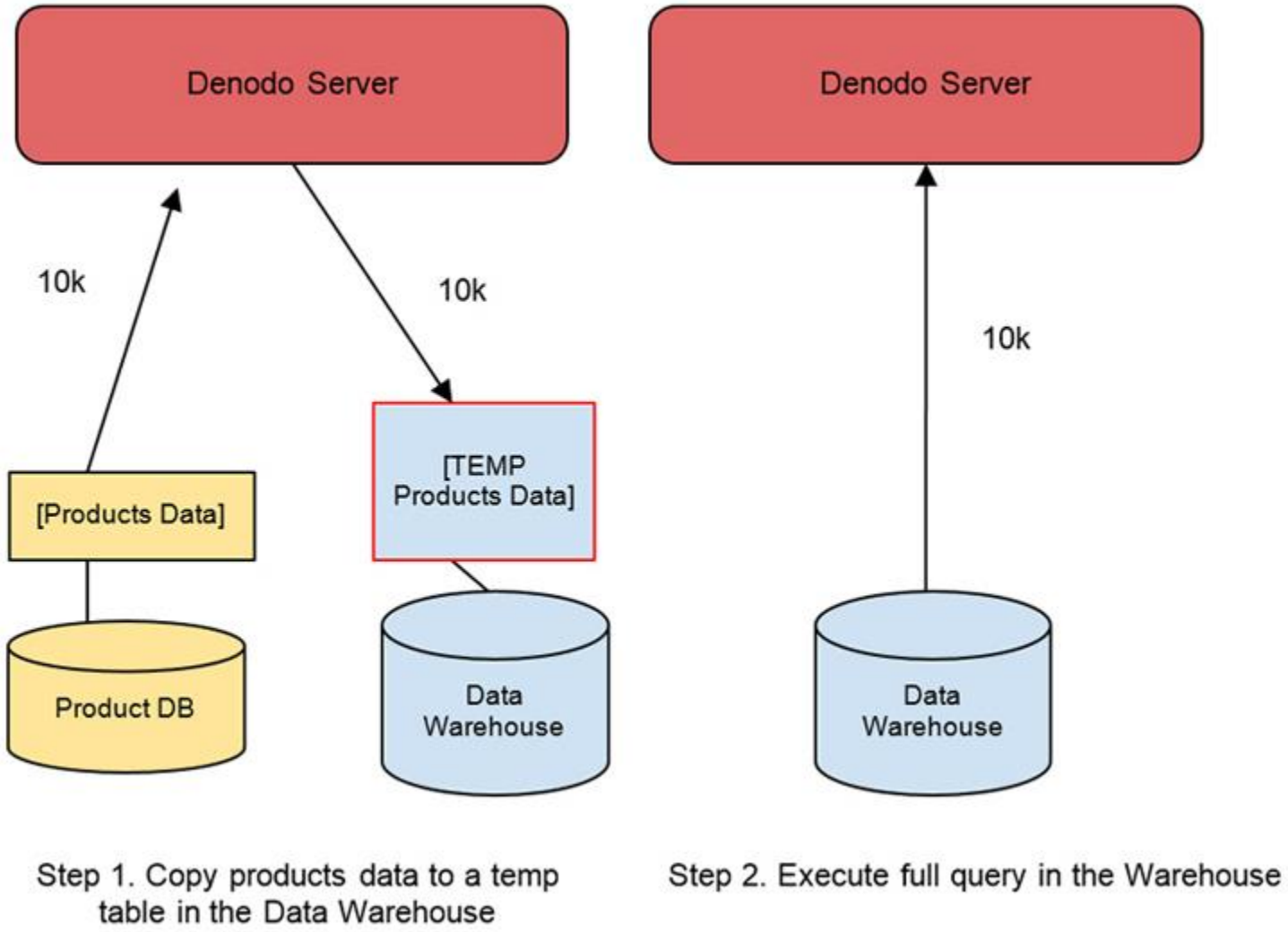
3. Join both tables and aggregate – Data Virtualization

Product Category	Total Sales
Phone	2500
Computer	1200

Total sales for each product category in the last 9 months

Dynamic Query Rewriting for Big Data

4. On the Fly Data Movement

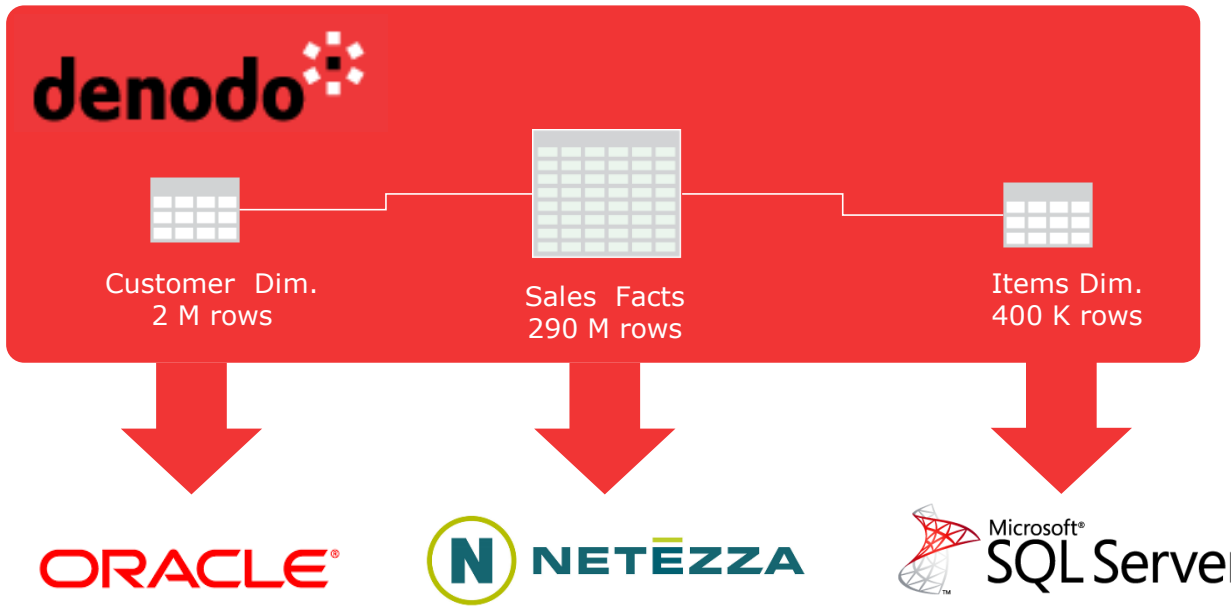


Maximum discount for each product in the last 9 months

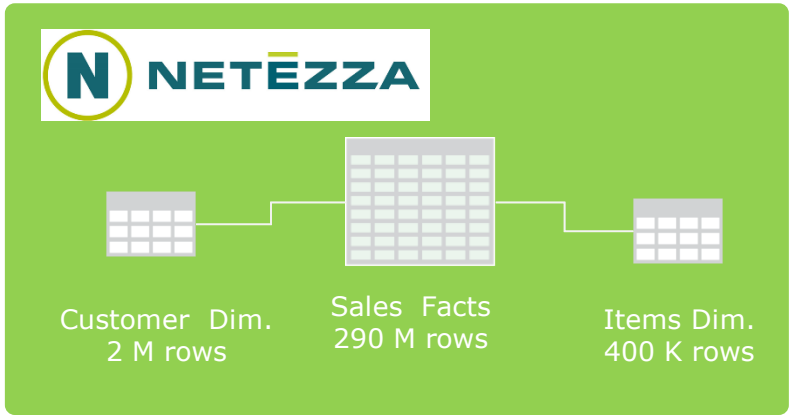
Performance Comparison

Logical Data Warehouse vs. Physical Data Warehouse

Compares the performance of a federated approach in Denodo with an MPP system where all the data has been replicated via ETL



VS.



* TPC-DS is the de-facto industry standard benchmark for measuring the performance of decision support solutions including, but not limited to, Big Data systems.

Performance Comparison

Logical Data Warehouse vs. Physical Data Warehouse

Query Description	Returned Rows	Time Netezza	Time Denodo (Federated Oracle, Netezza & SQL Server)	Optimization Technique (automatically selected)
Total sales by customer	1.99 M	20.9 sec.	**21.4 sec.**	Full aggregation push-down
Total sales by customer and year between 2000 and 2004	5.51 M	52.3 sec.	**59.0 sec**	Full aggregation push-down
Total sales by item brand	31.35 K	4.7 sec.	**5.0 sec.**	Partial aggregation push-down
Total sales by item where sale price less than current list price	17.05 K	3.5 sec.	**5.2 sec**	On the fly data movement

Dynamic Query Optimizer

Delivers Breakthrough Performance for Big Data, Logical Data Warehouse, and Operational Scenarios

Best dynamic query optimization engine in the industry.

- ✓ Dynamically determines lowest-cost query execution plan based on statistics
- ✓ Factors in all the special characteristics of big data sources such as number of processing units and partitions
- ✓ Can easily handle any number of incremental queries
- ✓ Enables connectivity to the broadest array of big data sources such as Redshift, Impala, Spark.

Q&A

Kurt Jackson

Platform Lead

AUTODESK

Ravi Shankar

Chief Marketing Officer

denodo

Next Steps

View Educational Seminar (on-demand): Logical Data Warehouse, Data Lakes, and Data Services Marketplace
Visit: www.denodo.com

Read about Data Virtualization for Logical Data Warehouse and Data Lakes
Visit: denodo.com/en/solutions/horizontal-solutions/logical-data-warehouse

Get Started! Download Denodo Express
Visit: www.denodo.com/en/denodo-platform/denodo-express
Access Denodo Platform on AWS
Visit: www.denodo.com/en/denodo-platform/denodo-platform-for-aws

Thank You

Kurt Jackson

Platform Lead

AUTODESK.

Ravi Shankar

Chief Marketing Officer

denodo