denodo
webinar series

DATA VIRTUALIZATION PACKED LUNCH WEBINAR SERIES

Sessions Covering Key Data Integration Challenges
Solved with Data Virtualization

denodo
webinar series

In Memory Parallel Processing for Big Data Scenarios

Paul Moxon

VP Data Architectures & Chief Evangelist, Denodo

Pablo Alvarez-Yanez

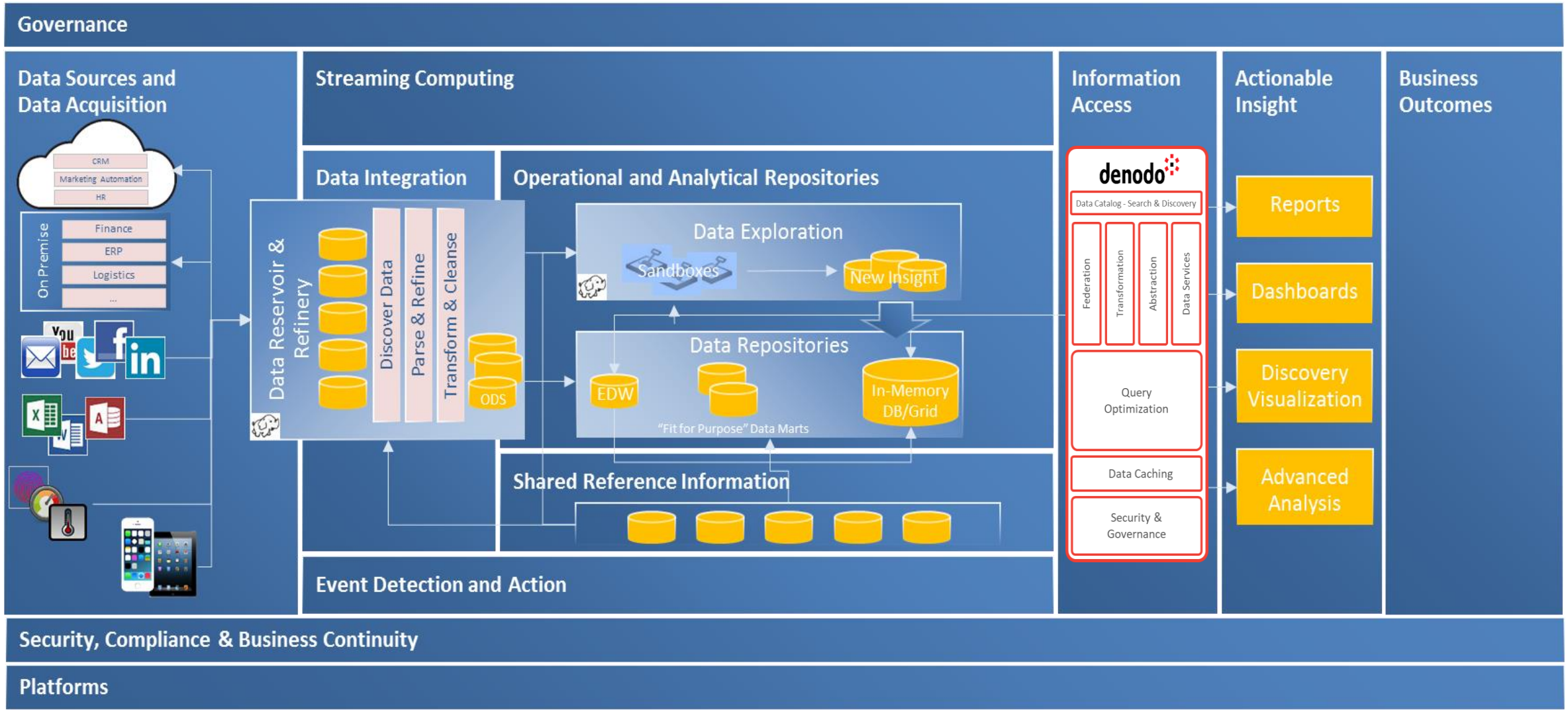
Principal Solution Architect, Denodo

Agenda

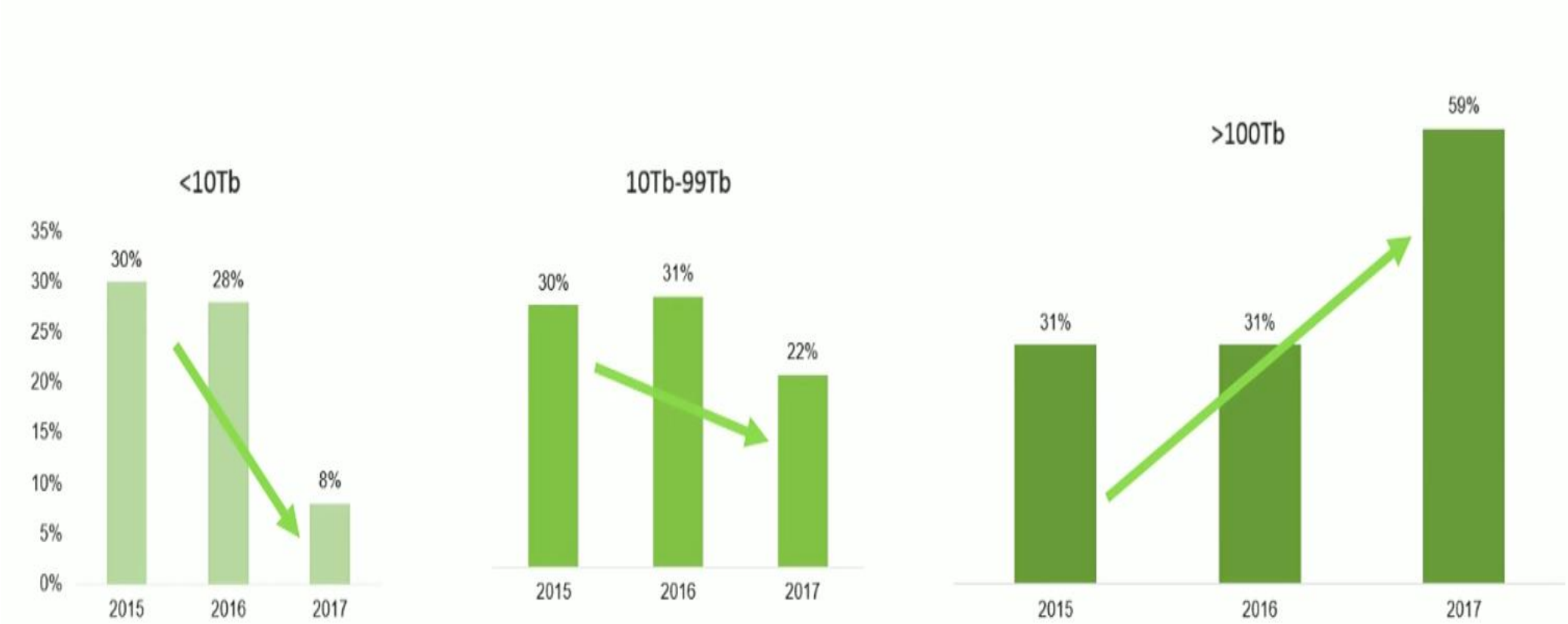
1. Modern Data Architecture
2. Denodo Platform – Big Data Integrations
3. Big Data Performance
4. Putting This All Together
5. Q&A
6. Next Steps

The Modern Data Architecture

Data Integration – A Modern Data Ecosystem

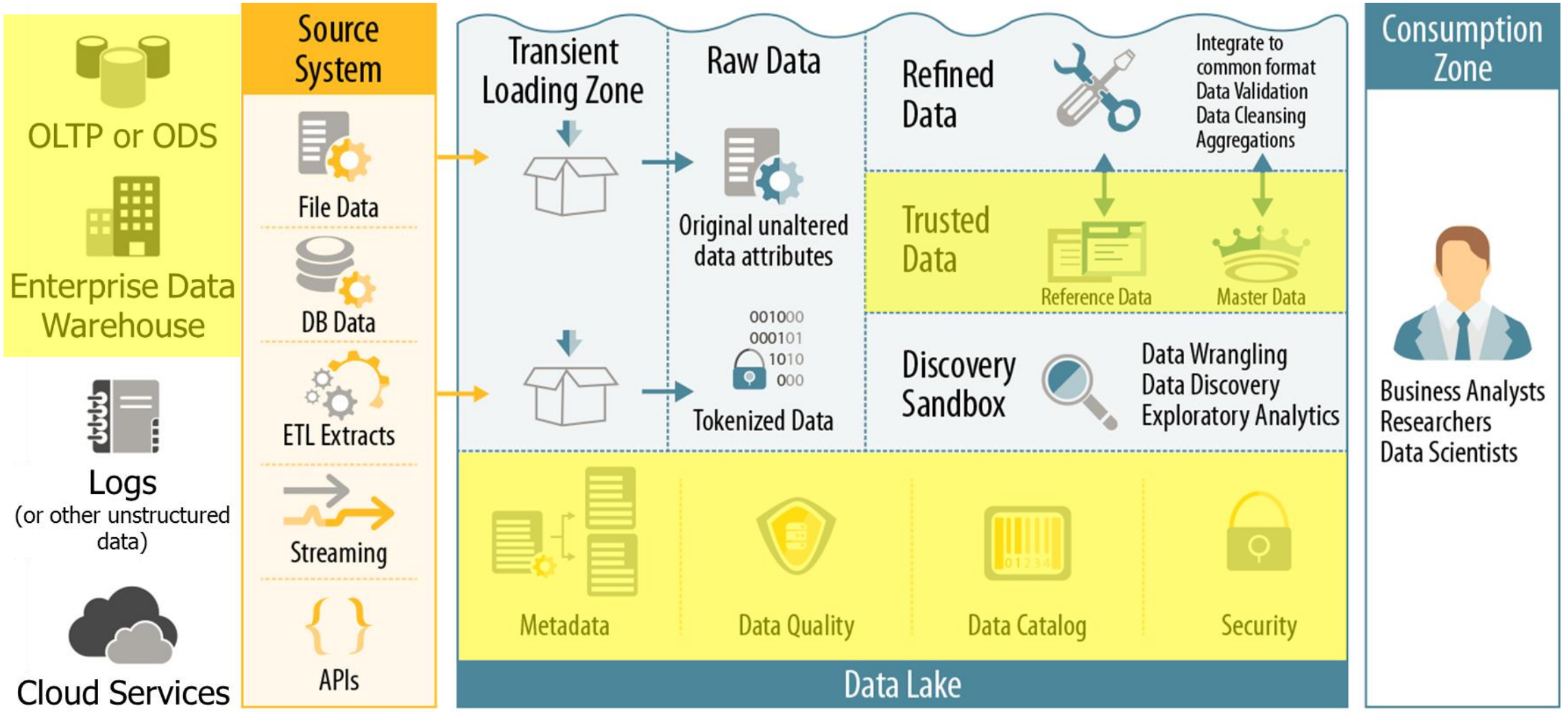


Organizations are Storing More and More Data…

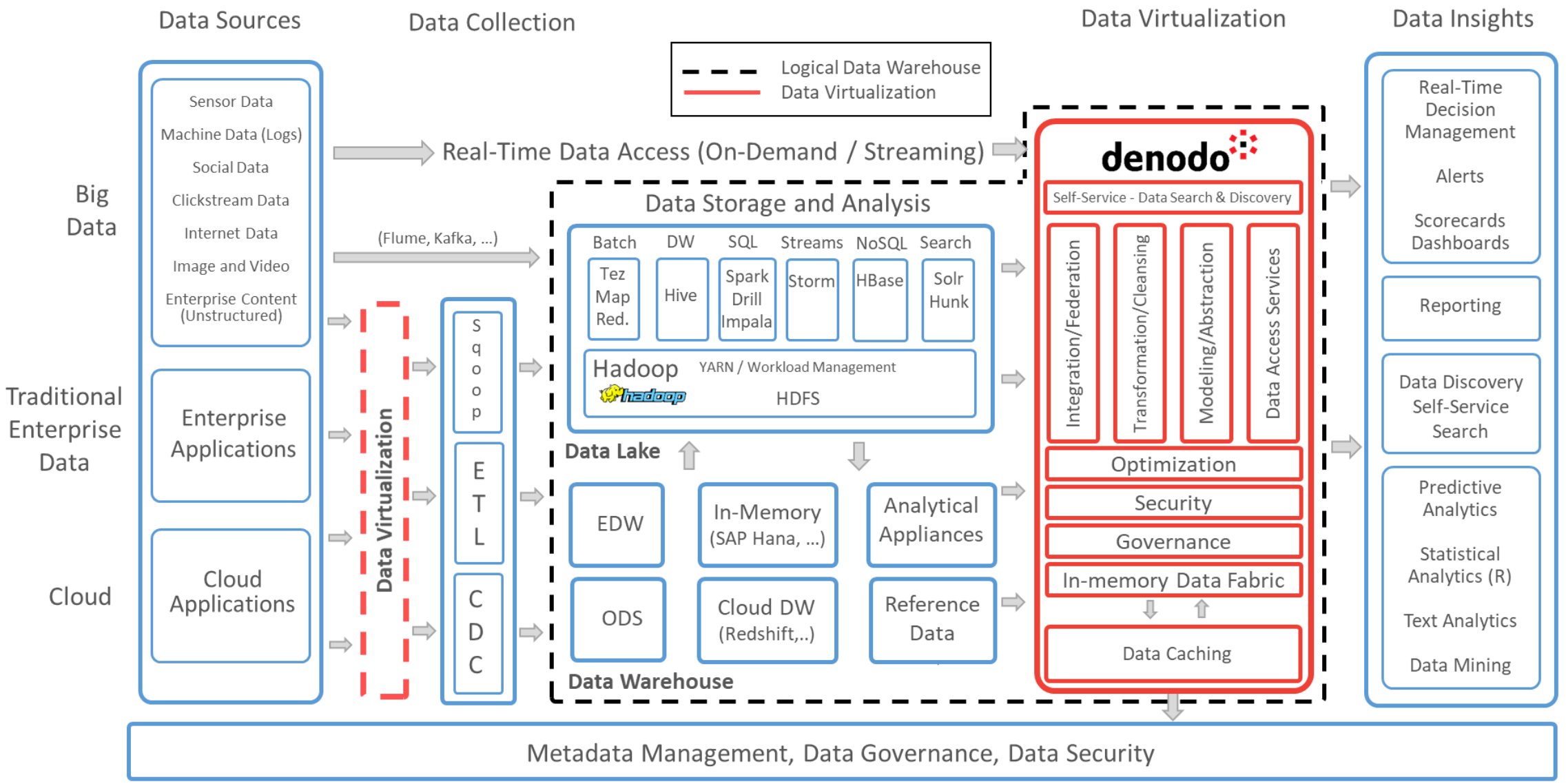


Source: Forrester's Business Technographics® Global Data And Analytics Survey, 2017

Data Lake – The Challenges

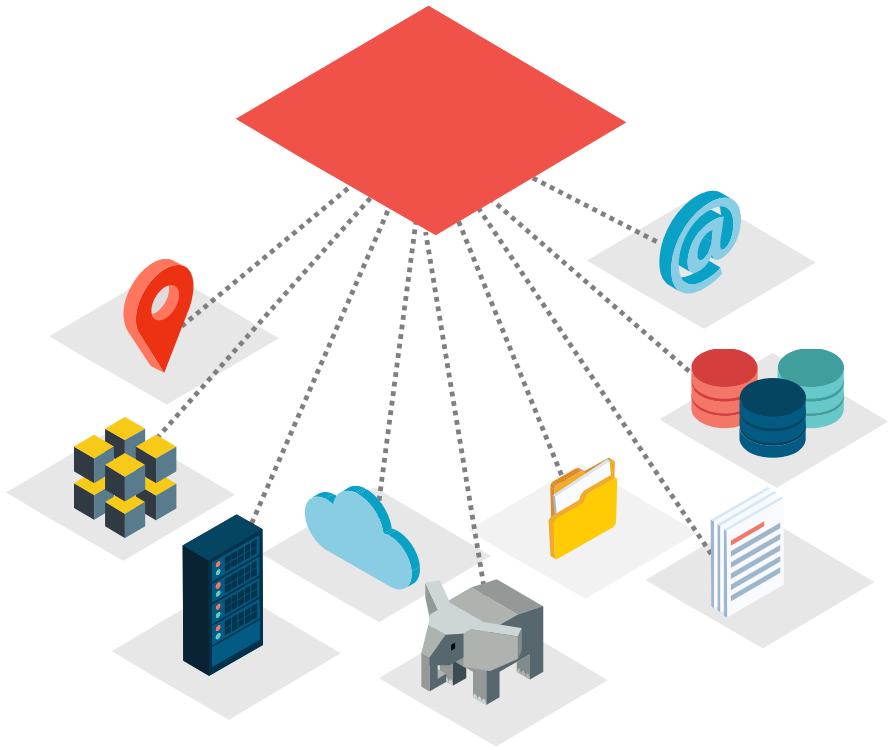


Big Data & Analytics Reference Architecture



Denodo Platform and Big Data Integrations

Hadoop as a Data Source

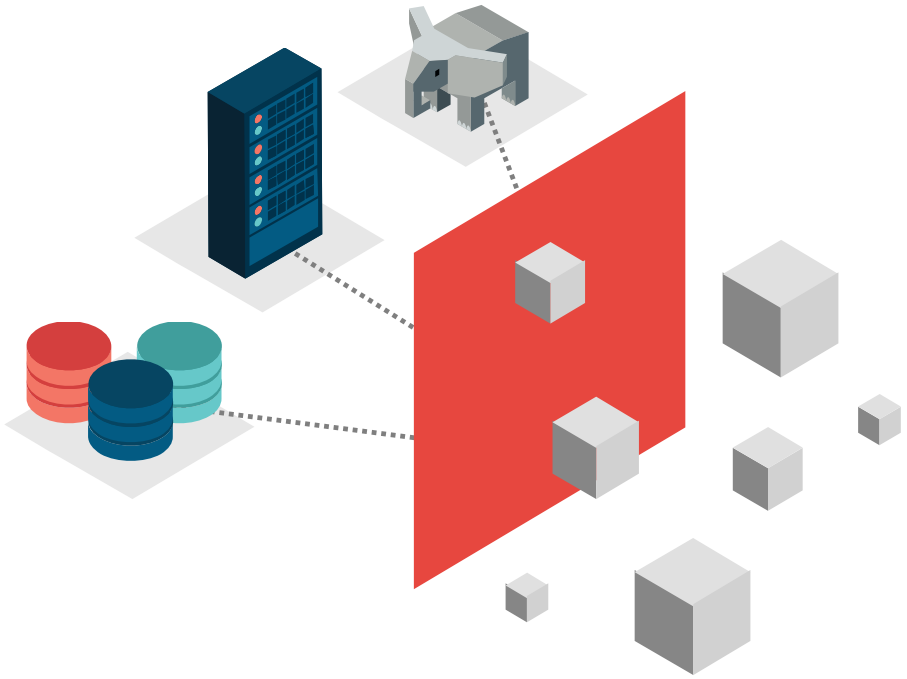


Denodo offers native connectors for all the major SQL-on-Hadoop engines:

- Hive
- Impala
- SparkSQL
- Presto

In addition, Denodo also offers connectivity for HBase and direct HDFS access to different file formats

Hadoop as a Cache



Denodo uses an external RDBMS of your choice to persist copies of the result sets to improve execution times

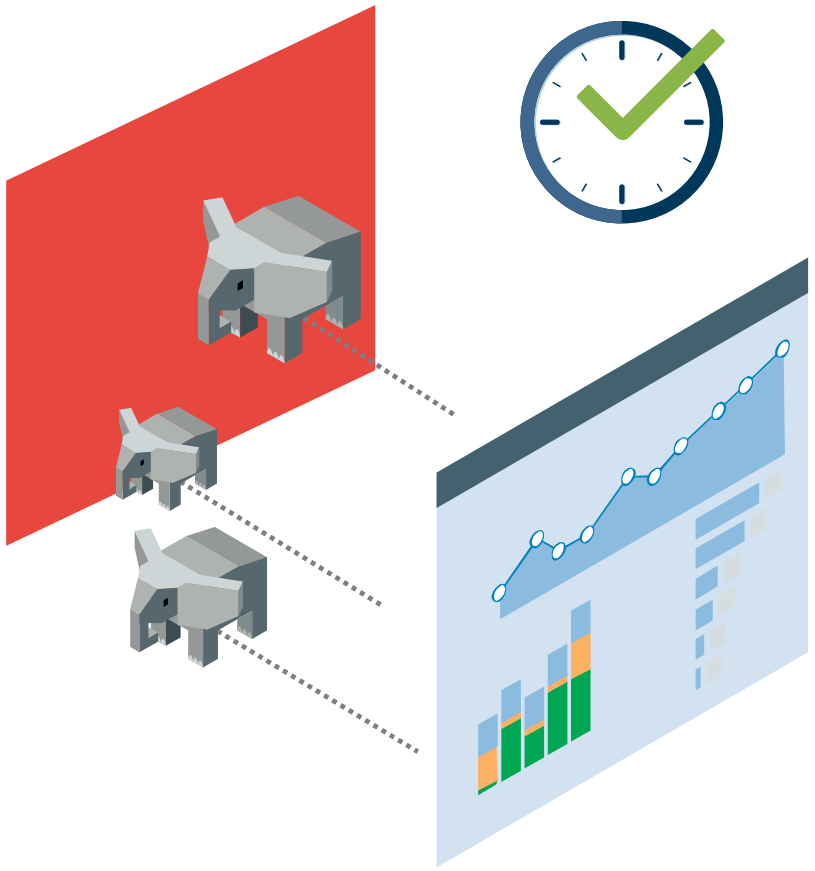
- Since data is persisted in an RDBMS, Denodo can push down relational operations, like JOINS with other tables, to the database used for cache

SQL-on-Hadoop systems can also be used as Denodo's cache

Cache load process based on direct load to HDFS:

1. Creation of the target table in Cache system
2. Generation of Parquet files (in chunks) with Snappy compression in the local machine
3. Upload in parallel of Parquet files to HDFS

Hadoop as a Processing Engine



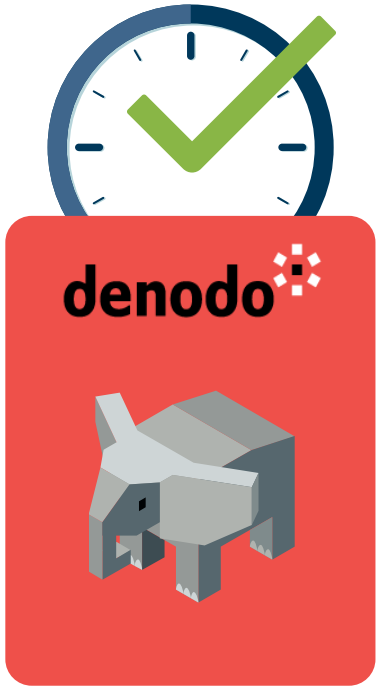
Denodo optimizer provides native integration with MPP systems to provide one extra key capability: **Query Acceleration**

Denodo can move, on demand, processing to the MPP during execution of a query

- Parallel power for calculations in the virtual layer
- Avoids slow processing in-disk when processing buffers don’t fit into Denodo’s memory (swapped data)

Demo

Combining Denodo's Optimizer with a Hadoop MPP



Denodo provides the most advanced optimizer in the market, with techniques focused on data virtualization scenarios with large data volumes

In addition to traditional Cost Based Optimizations (CBO), Denodo's optimizer applies innovative optimization strategies, designed specifically for virtualized scenarios, beyond traditional RDBMS optimizatios.

Combined with the tight integration with SQL-on-Hadoop MPP databases, it creates a very powerful combo

Example Scenario

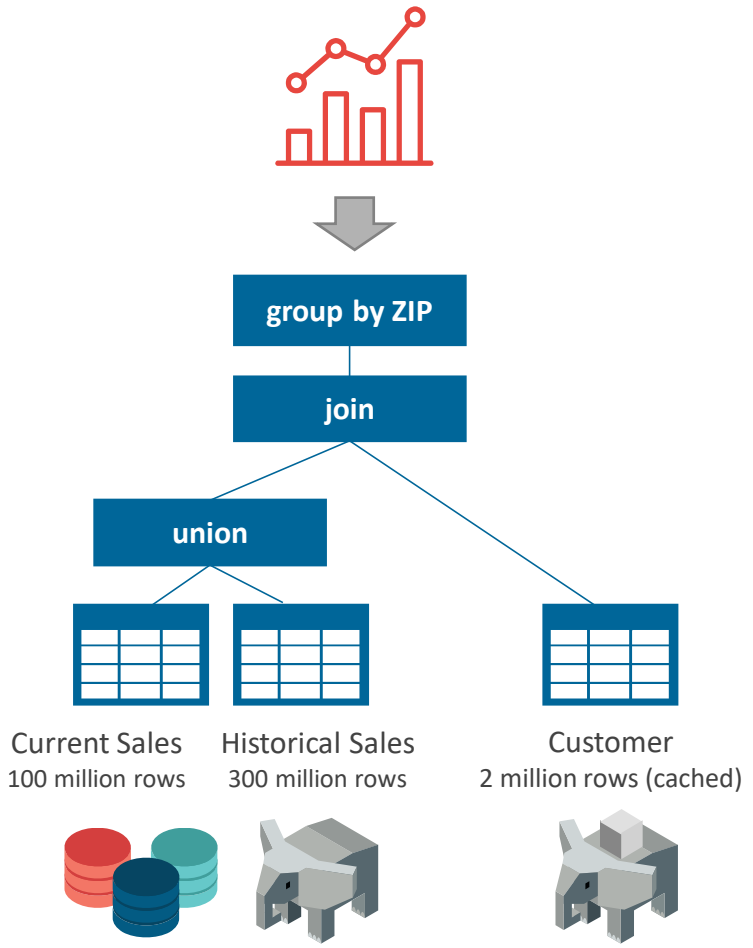
Trend of sales by zip code over the previous years.

Scenario:

- Current data (last 12 months) in EDW
- Historical data offloaded to Hadoop cluster for cheaper storage
- Customer master data is used often, so it is cached in the Hadoop cluster

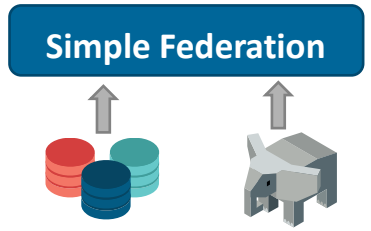
Very large data volumes:

- Sales tables have hundreds of millions of rows

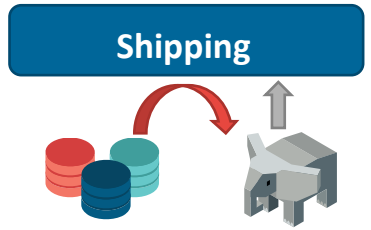


Example: What are the options?

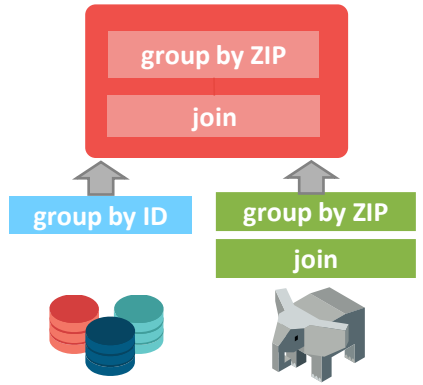
1. Option A: Simple Federation in Virtual Layer
   - Move hundreds of millions of rows for processing in the virtual layer



2. Option B: Data Shipping
   - Move “Current sales” to Hadoop and process content in the cluster
   - Moves 100 million rows

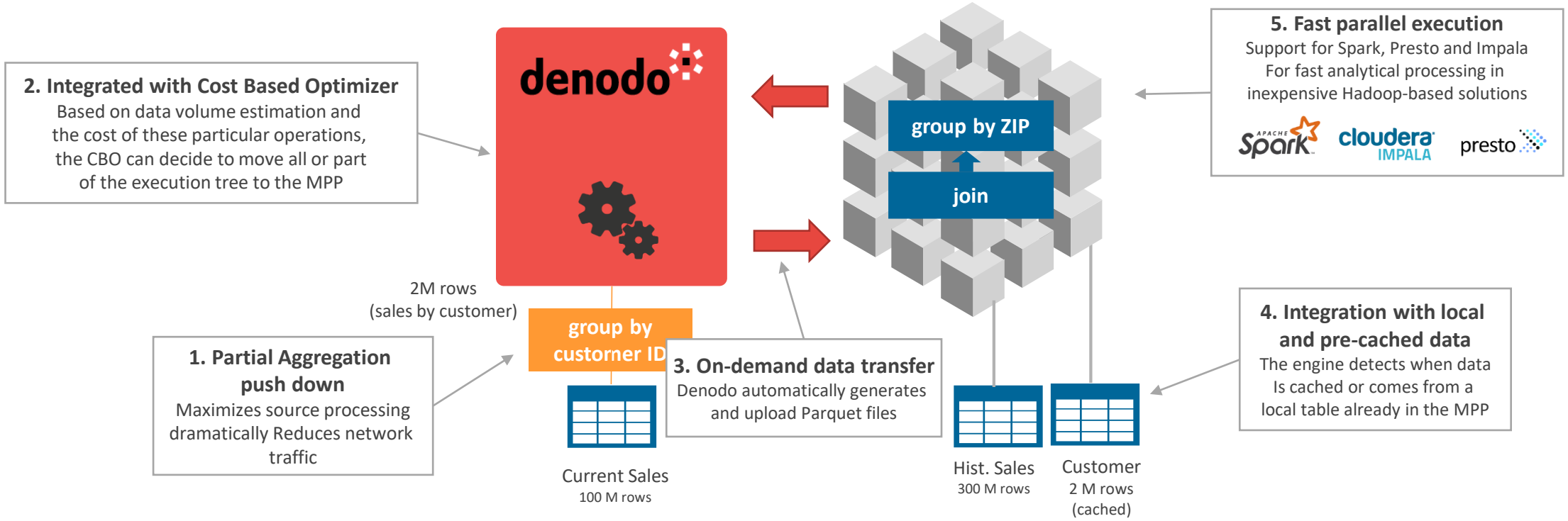


3. Option C: Partial Aggregation Pushdown (Denodo 6)
   - Modifies the execution tree to split the aggregation in two steps:
     1. First by Customer ID for the JOIN (pushed down to source)
     2. Second by zip code for the final results (in virtual layer)
   - Reduces significantly network traffic but processing of large amount of data in the virtual layer (aggregation by zip code) becomes the bottleneck



4. Denodo’s MPP Integration (Denodo 7)

Example: Denodo's Integration with the Hadoop Ecosystem

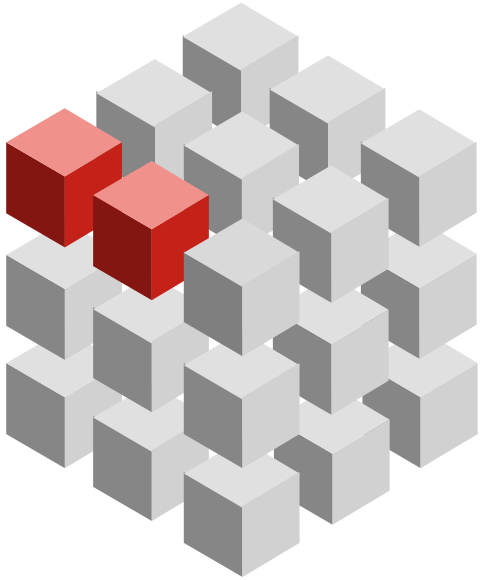


System	Execution Time	Optimization Techniques
Others	~ 10 min	Simple federation
No MPP	43 sec	Aggregation push-down
With MPP	11 sec	Aggregation push-down + MPP integration (Impala 8 nodes)

Demo

Putting It All Together

Putting all the pieces together



These three techniques (Hadoop as a data source, cache and processing engine) can be combined to successfully approach complex scenarios with big data volumes in an efficient way:

- Surfaces all the company data without the need to replicate all data to the Hadoop lake
- Improves governance and metadata management to avoid “data swamps”
- Allows for on-demand combination of real-time (from the original sources) with historical data (in the cluster)
- Leverages the processing power of the existing cluster controlled by Denodo’s optimizer

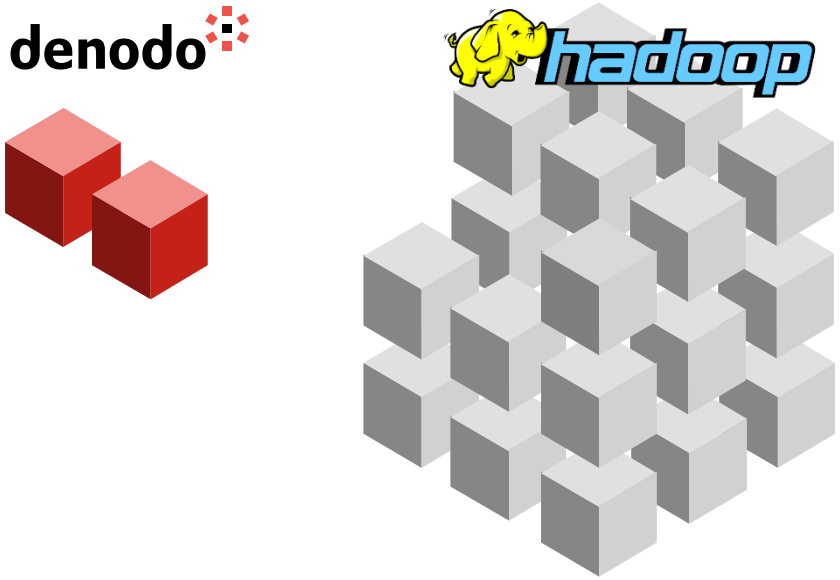
Architecture

Denodo Cluster

- Multiple nodes behind a load balancer for HA
- Running on Hadoop Edge nodes

Hadoop Cluster

- Processing and Storage nodes
- Same subnet as Denodo cluster



To benefit from this architecture, **Denodo servers should run in edge nodes** of the Hadoop cluster

This will ensure:

- Faster uploads to HDFS
- Faster data retrieval from the MPP
- Better compatibility with the Hadoop configuration and versions of the libraries

Q&A

Next steps

Download Denodo Express:
www.denodoexpress.com

Access Denodo Platform in the Cloud!
30 day FREE **trial** available!

Denodo for Azure:
www.denodo.com/TrialAzure/PackedLunch
Denodo for AWS: www.denodo.com/TrialAWS/PackedLunch

denodo
webinar series

Next session

Self-Service Information Consumption Using Data Catalog

Thursday, March 15th, 2018 at 11:00am (PST)

Paul Moxon

VP Data Architectures and Chief Evangelist

denodo
webinar series

Thank you!