Dell - Data Engineering Training

Uday Kumar – Data Platform Architect

IntelliPaat

Day 3

Data Engineering Training

Agenda

1. Introduction to Data Virtualization
2. Understanding the concept of data virtualization
3. Advantages and use cases of data virtualization.
4. Key components of data virtualization
5. Architecture of data virtualization
6. Demo & Hands-on

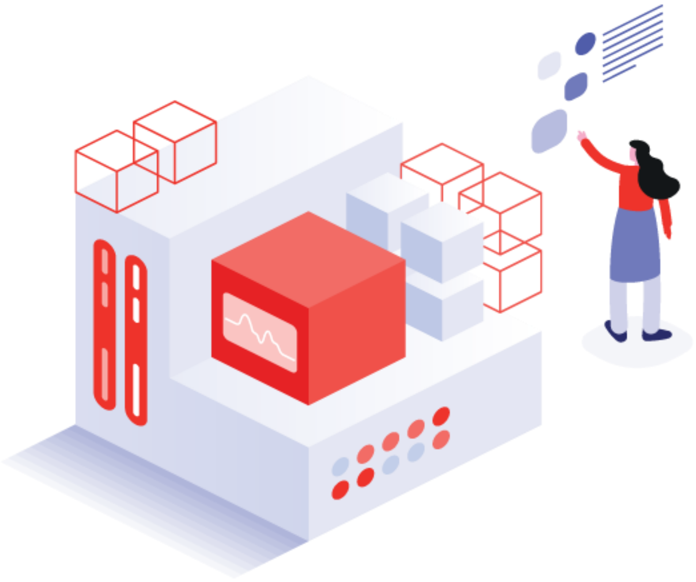
Data Virtualization

INTRODUCTION TO DATA VIRTUALIZATION

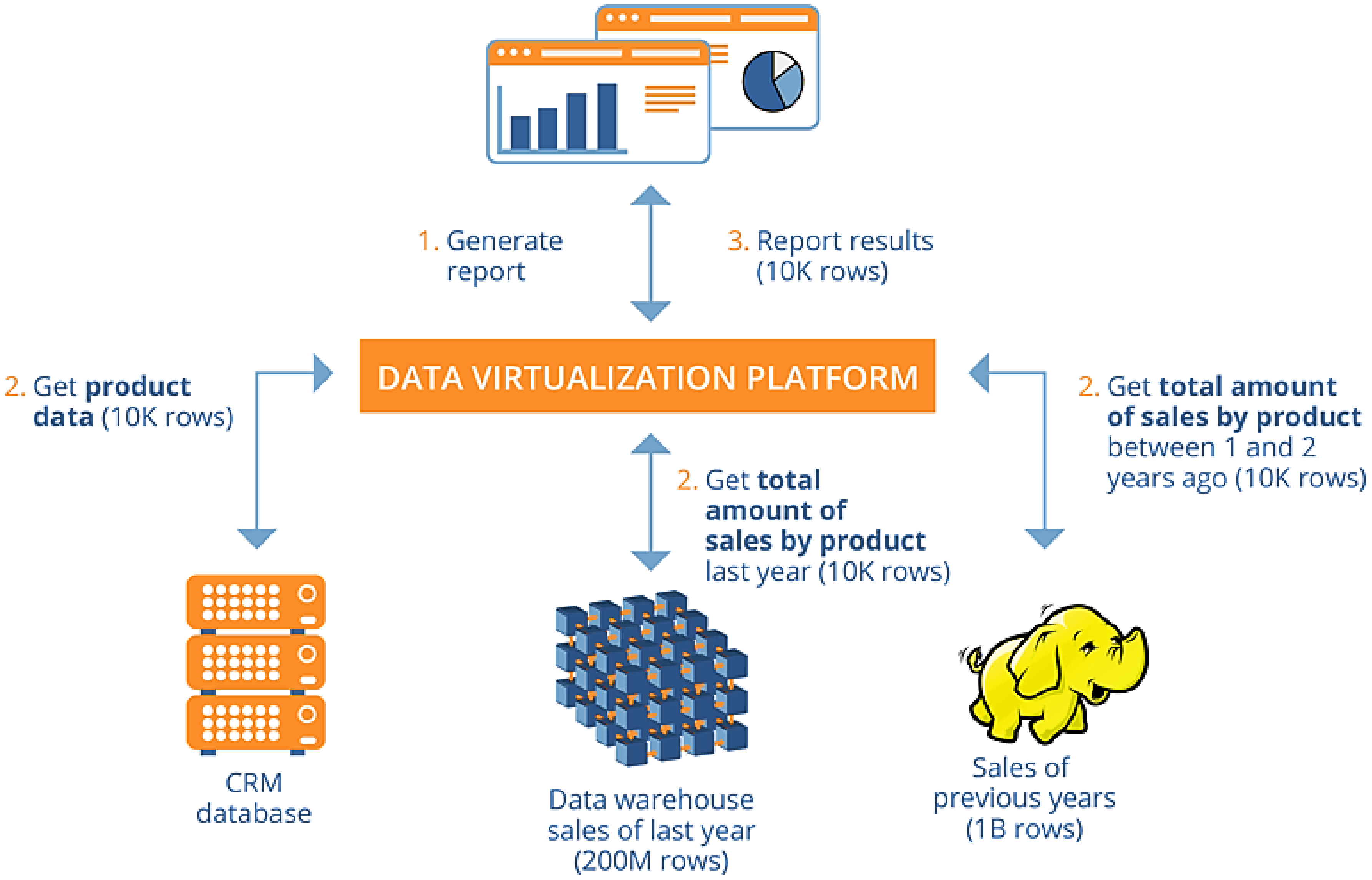
Data virtualization is an approach to integrating data from **multiple sources of different types into a holistic, logical view without moving it physically**. In simple terms, **data remains in original sources** while users can access and analyze it virtually via special middleware.

Data virtualization is a **method for accessing enterprise data through a semantic data layer** that hides the underlying complexities of the data sources, while allowing for centralized data governance.

Data virtualization uses a simple **three-step process—connect, combine, consume**—to deliver a holistic view of enterprise information to business users across all of the underlying source systems.



DATA VIRTUALIZATION – HOW IT WORKS



1. Generate report

3. Report results (10K rows)

DATA VIRTUALIZATION PLATFORM

2. Get **product data** (10K rows)

2. Get **total amount of sales by product** last year (10K rows)

2. Get **total amount of sales by product** between 1 and 2 years ago (10K rows)

CRM database

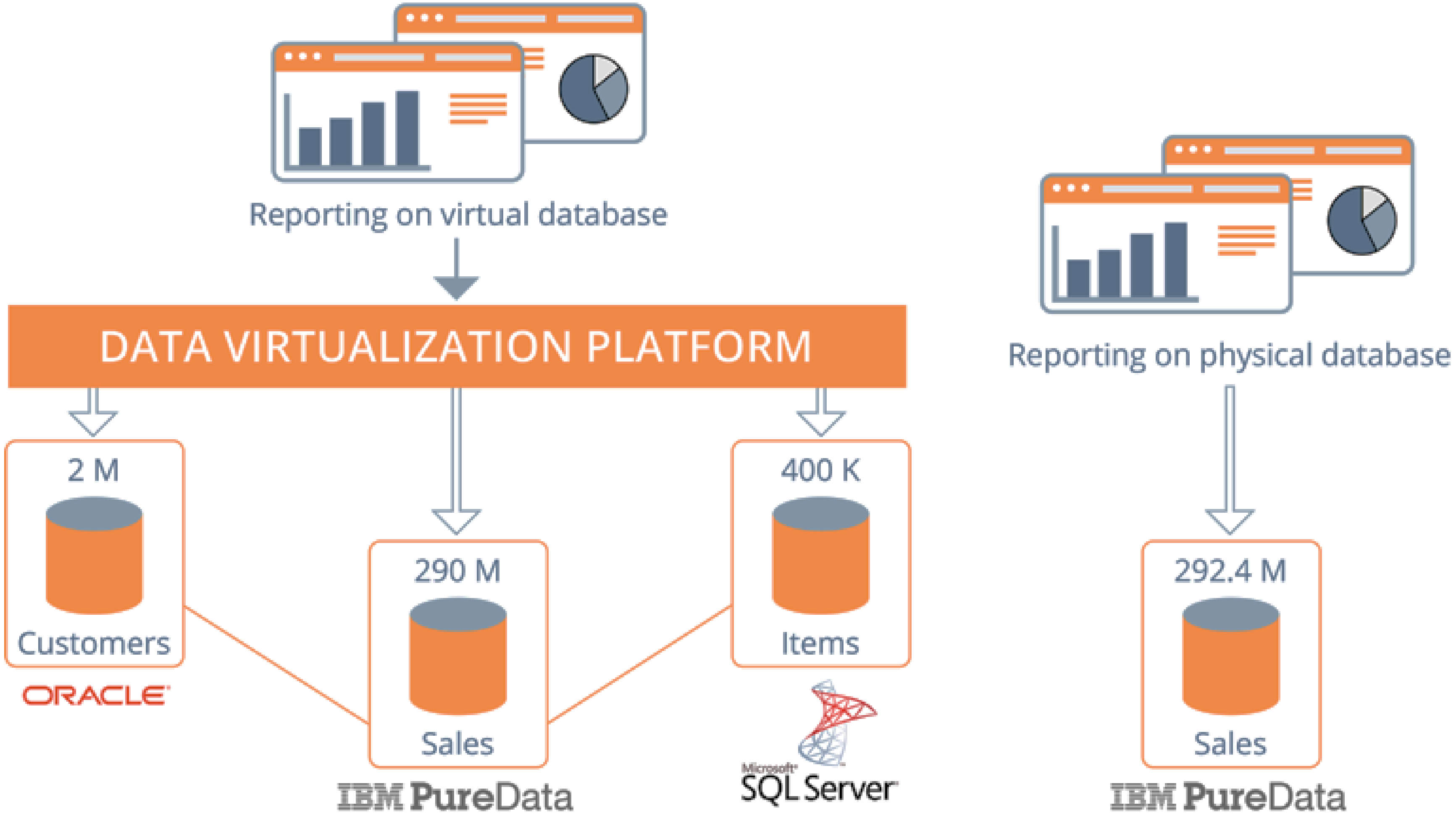
Data warehouse sales of last year (200M rows)

Sales of previous years (1B rows)

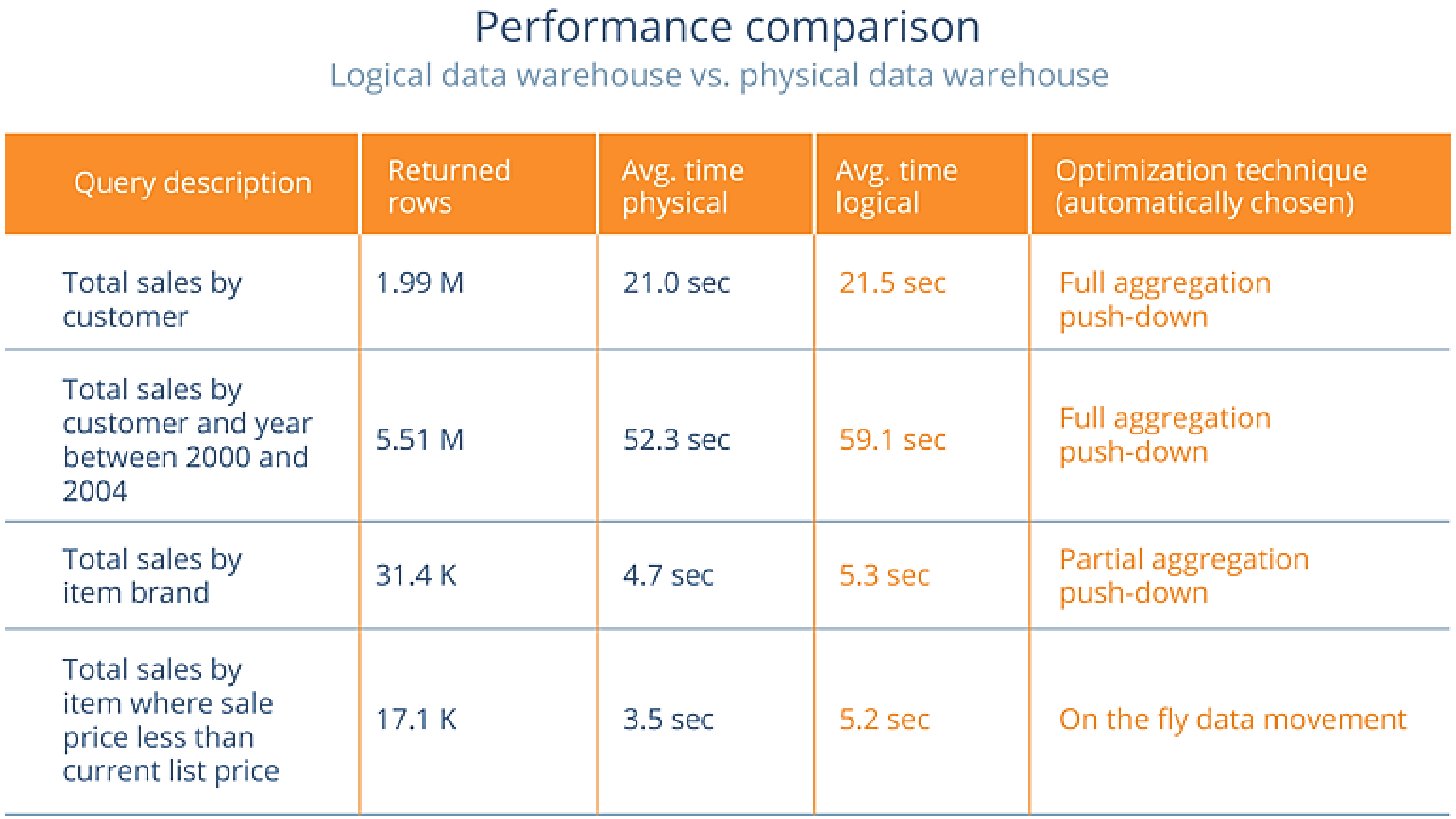
DATA VIRTUALIZATION – PRACTICAL EXAMPLE

IntelliPaat

Logical data warehouse vs. physical data warehouse



DATA VIRTUALIZATION

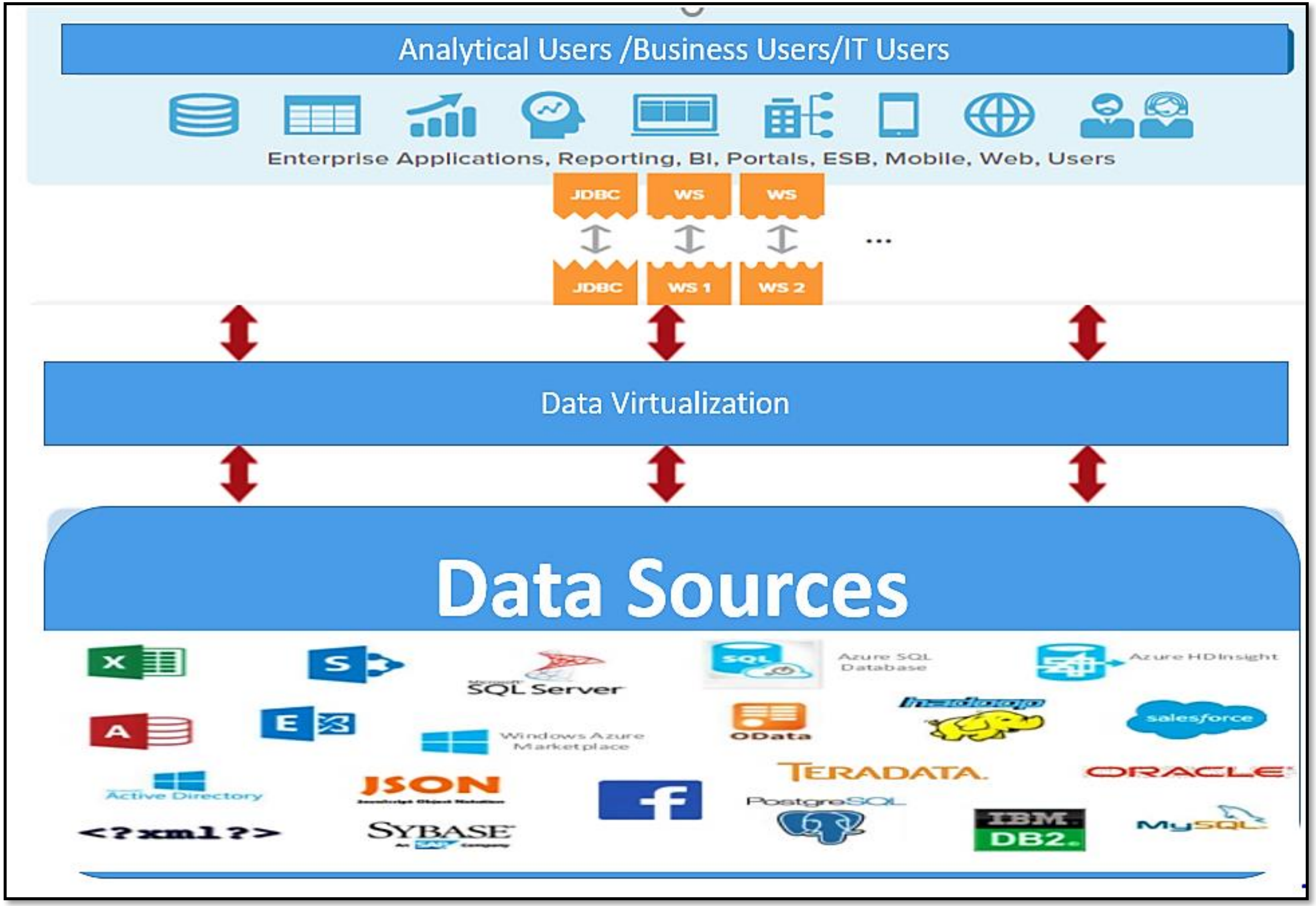


Performance comparison

Logical data warehouse vs. physical data warehouse

Query description	Returned rows	Avg. time physical	Avg. time logical	Optimization technique (automatically chosen)
Total sales by customer	1.99 M	21.0 sec	21.5 sec	Full aggregation push-down
Total sales by customer and year between 2000 and 2004	5.51 M	52.3 sec	59.1 sec	Full aggregation push-down
Total sales by item brand	31.4 K	4.7 sec	5.3 sec	Partial aggregation push-down
Total sales by item where sale price less than current list price	17.1 K	3.5 sec	5.2 sec	On the fly data movement

DATA VIRTUALIZATION



DATA VIRTUALIZATION CRITICAL FOR DATA MODERNIZATION

The Problem

Organizations recognize that to make smarter decisions, delight their customers, and outcompete their rivals, they need to exploit their data assets more effectively. This trend towards data-driven business is nothing new, but given the impact of Covid-19, the pace of transformation has dramatically increased.

Exploiting the power of data analytics/business intelligence and workflow automation is one way for companies to accelerate new revenue streams while reducing costs by streamlining and improving the performance of data services.

But here lies the challenge – enterprise data is stored in disparate locations with rapidly evolving formats such as:

- Relational and non-relational databases like MySQL, Amazon Redshift or MongoDB
- Cloud/Software-as-a-Service applications like Netsuite, Salesforce or Mailchimp
- Social Media or Website data like Facebook, Twitter or Google Analytics
- CRM/ERP data like SAP, Oracle or Microsoft Dynamics
- Data lakes and Enterprise Data Warehouses
- Flat files like XML, CSV or JSON
- Big Data

DATA VIRTUALIZATION CRITICAL FOR DATA MODERNIZATION

The demand for faster and higher volumes of increasingly complex data leads to further challenges such as:

- Delivering self-service capabilities for data users
- Creating time efficiency in data management
- Achieving trusted data quality
- To address these challenges, organizations recognize the need to move from silos of disparate data and isolated technologies to a business-focused strategy where data and analytics are simply a part of everyday working life for business users.

DATA VIRTUALIZATION CRITICAL FOR DATA MODERNIZATION

The Solution

Data virtualization (DV) overcomes these challenges by exploiting the full potential of enterprise data. It breaks free from the requirement of knowing every technical detail of the data and **ultimately aggregating data into one single 'view**', without the need to move data into a central storage.

While all data remains in the source systems, **data virtualization creates a virtual/logical layer that delivers real-time data access** with the possibility to manipulate and transform the data in virtual views. This virtual layer delivers a simpler and more time-efficient data management approach.

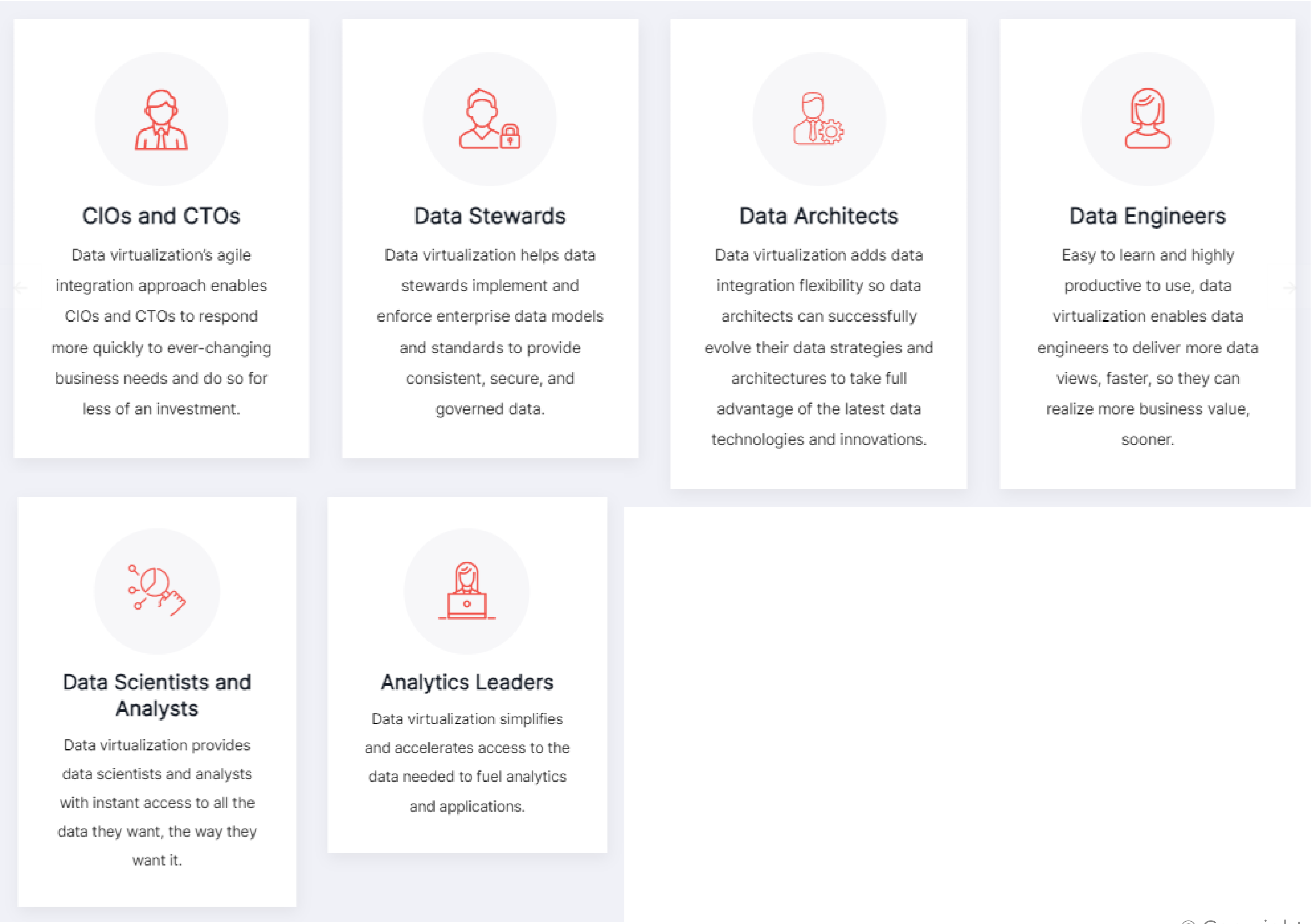
DV tools can **make data accessible with SQL, REST,** or other common data query methods, regardless of source format, further simplifying data management efforts.

DATA VIRTUALIZATION ADVANTAGES

IntelliPaat

- Faster Time-to-Solution
- Flexibility and Simplicity
- Cost-Effectiveness
- Consistent and Secure Data Governance

DATA VIRTUALIZATION HELP ORGANIZATION TO SUCCEED



CIOs and CTOs

Data virtualization's agile integration approach enables CIOs and CTOs to respond more quickly to ever-changing business needs and do so for less of an investment.

Data Stewards

Data virtualization helps data stewards implement and enforce enterprise data models and standards to provide consistent, secure, and governed data.

Data Architects

Data virtualization adds data integration flexibility so data architects can successfully evolve their data strategies and architectures to take full advantage of the latest data technologies and innovations.

Data Engineers

Easy to learn and highly productive to use, data virtualization enables data engineers to deliver more data views, faster, so they can realize more business value, sooner.

Data Scientists and Analysts

Data virtualization provides data scientists and analysts with instant access to all the data they want, the way they want it.

Analytics Leaders

Data virtualization simplifies and accelerates access to the data needed to fuel analytics and applications.

DATA VIRTUALIZATION COMMON USE CASES

Data Integration: This is the most likely case you will encounter, since virtually every company has data from many different data sources. That **means bridging an old data source, housed in a client/server setup, with new digital systems like social media. You use connections, like Java DAO, ODBC, SOAP, or other APIs**, and search your data with the data catalog.

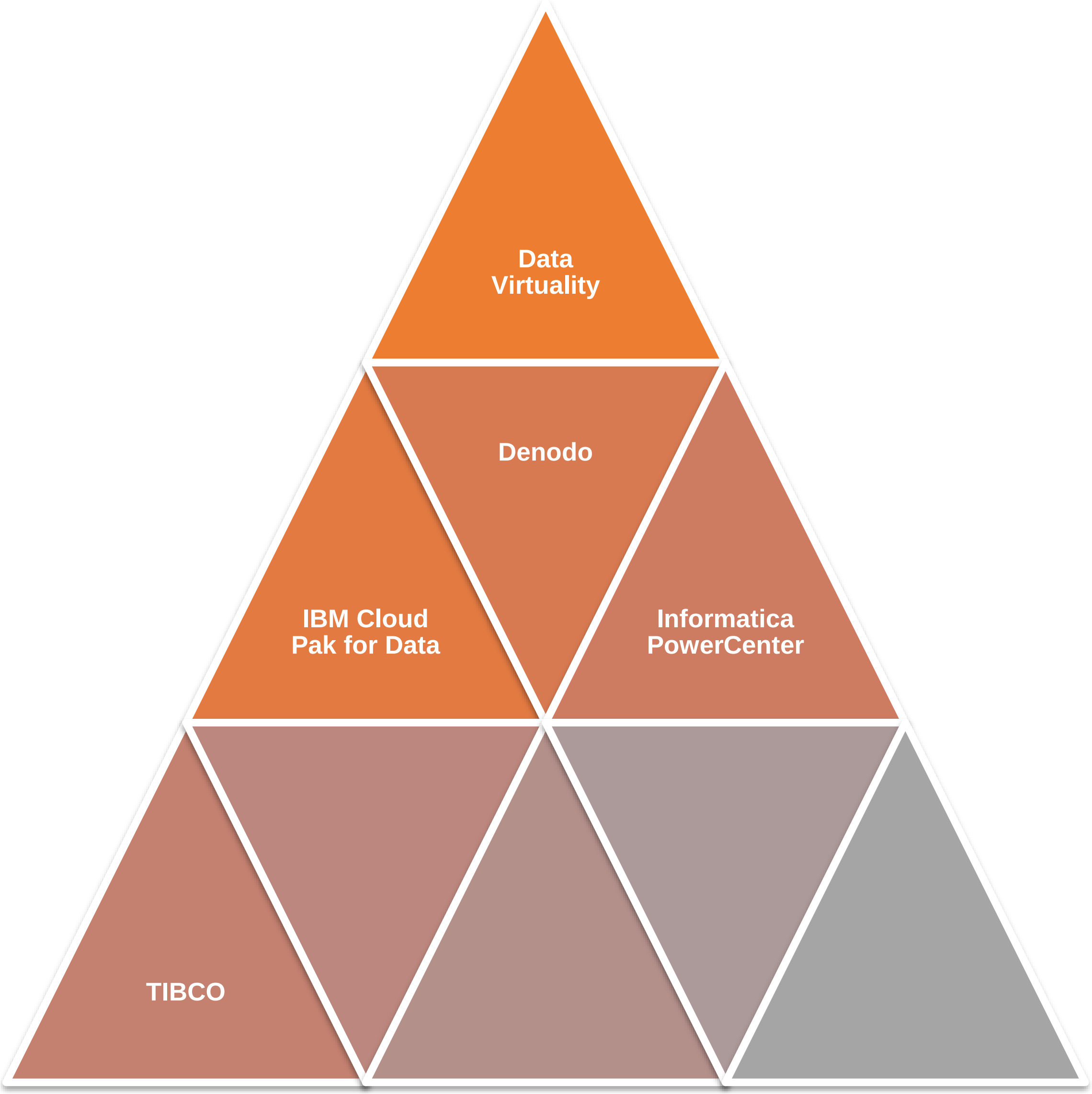
Big Data and Predictive Analytics: The nature of data virtualization works well here because Big Data and predictive analytics are built on heterogeneous data sources**. It’s not just drawing from an Oracle database/RDBMS, Big Data comes from things like Web portal, Cellular Phone, social media, and email**. So data virtualization lends itself to these highly diverse methodologies.

Operational/end Usage: One of the great headaches for call centers or customer service applications is siloed data, and for a long time, it remained that way. **Dell Technologies would need a different call center for Servers than for Laptops/Desktops,** for example. With data virtualization spanning the data silos, everyone from a call center to a database manager can see the entire span of data stores from a single point of access.

Cloud Migration: Data virtualization technology can provide a secure and efficient mechanism to **replace TB-size datasets from on-premise to the cloud**, before spinning up space-efficient data environments needed for testing and cutover rehearsal.

DATA VIRTUALIZATION TOOLS

DATA VIRTUALIZATION VENDORS



The Benefits of Data Virtualization

Data virtualization (DV) is architecturally different from ETL which results in some clear differences in performance and approach:

With DV, data access is in real-time so unlike ETL, data doesn’t have to be moved. This creates much faster data delivery that requires far less infrastructure (because you don’t need intermediate servers and storage ‘depots’ for your data).

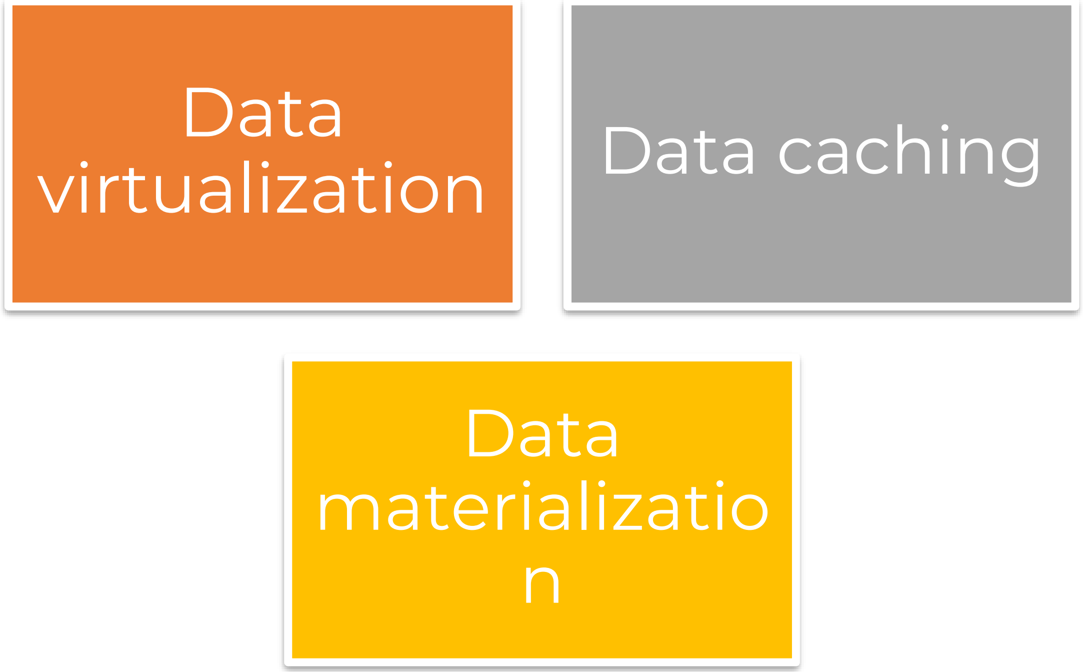
New data sources can be quickly reviewed, prototyped and added to the virtual layer without significant work.
Data virtualization can suffer from scalability constraints, but some products utilize caching features to compensate for this issue. However, even with caching some use cases present a challenge e.g. data trending/historical analysis.

Modern Data Architecture

Modern data architectures help data-driven organizations to quickly adapt to the **ever changing business needs** by dealing with the increasing complexity of the data landscape, enabling a growing number of diverse use cases, and ensuring flexibility and agility.

In most of today's cases, data comes from many different places and needs to be **integrated efficiently while considering data quality**, metadata management and data lineage, to name a few.

A modern data architecture **doesn't rely on one data integration** capability but combines several elements to deliver a breakthrough in flexibility and performance:

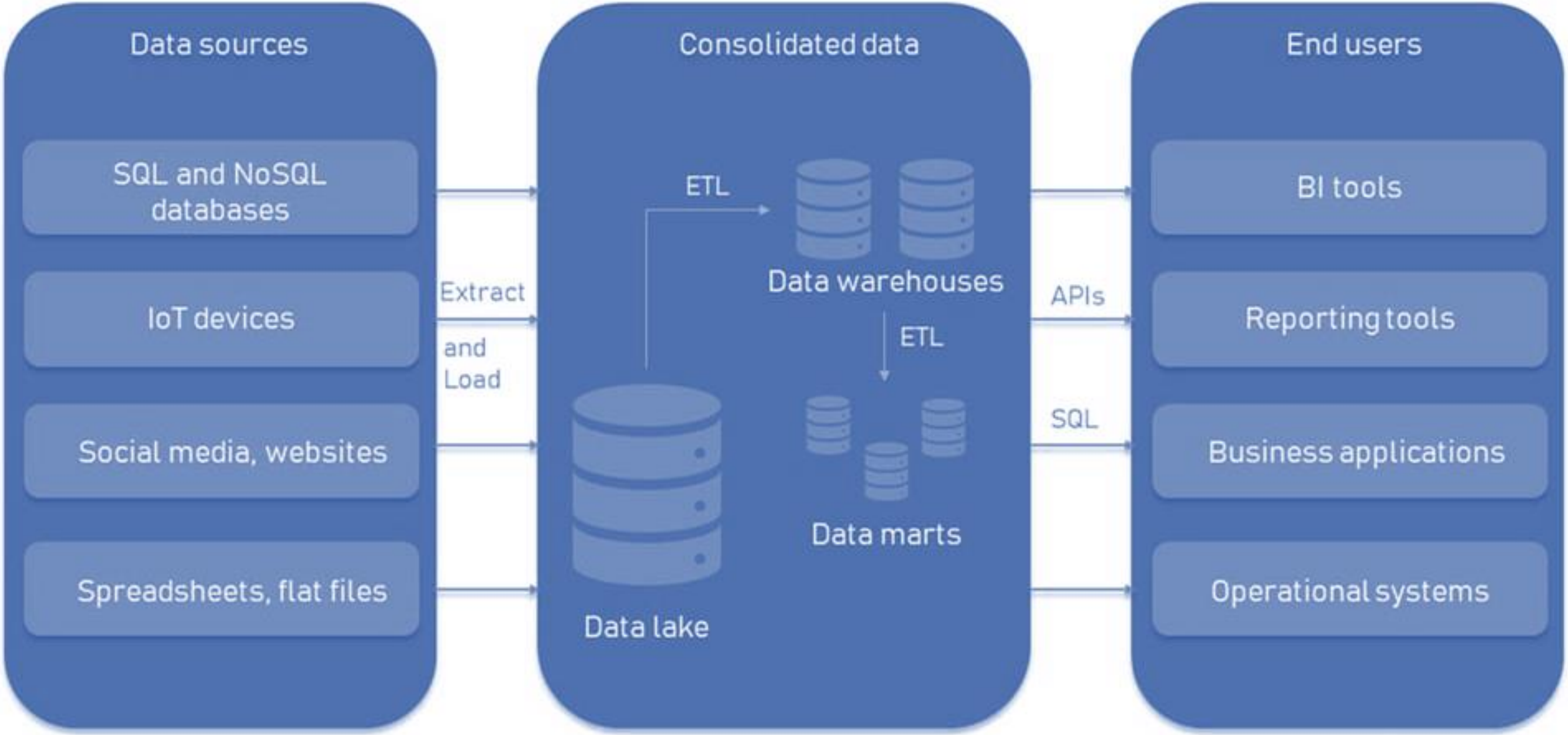


Data virtualization

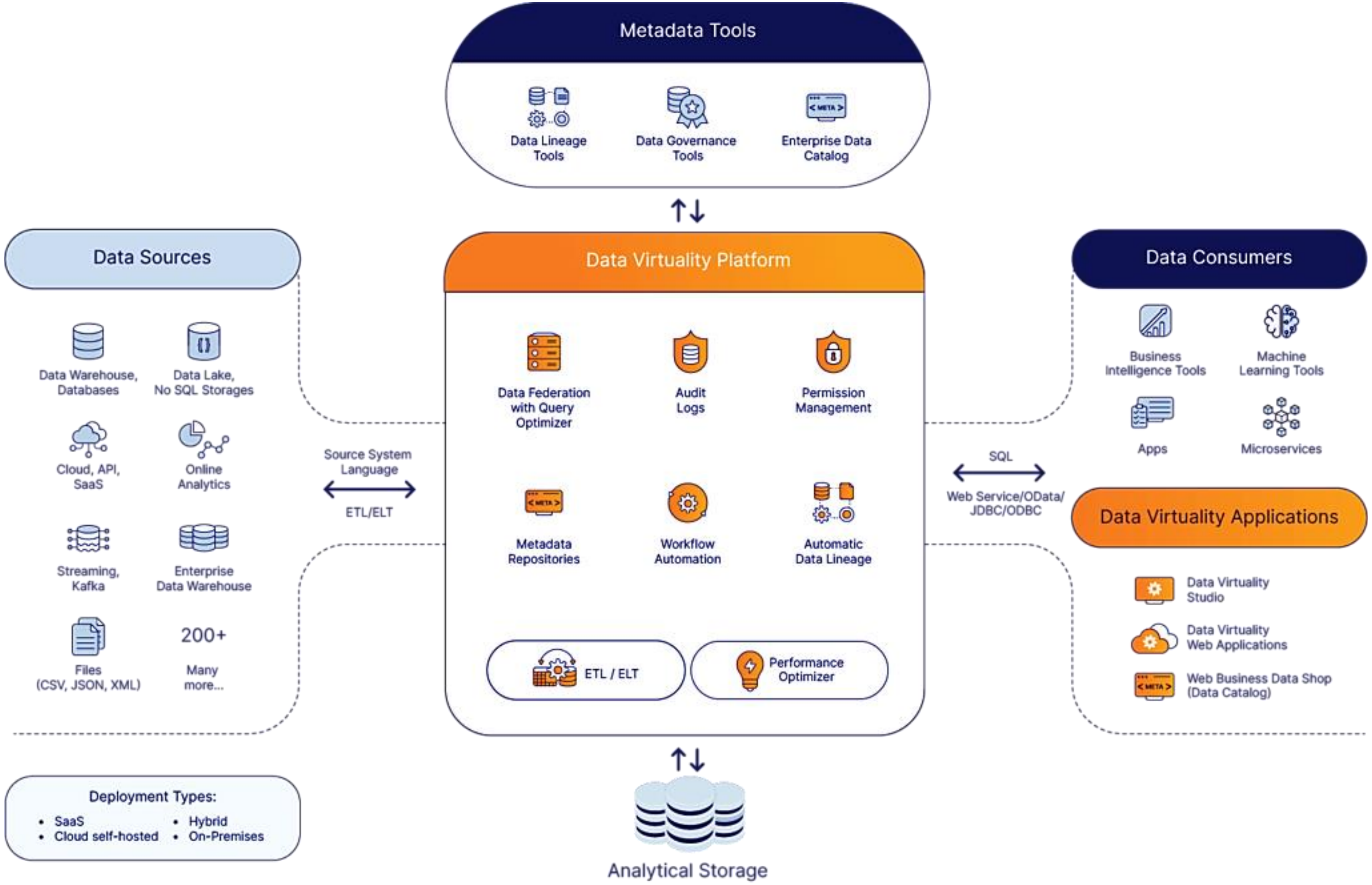
Data caching

Data materialization

Modern Data Architecture - Common



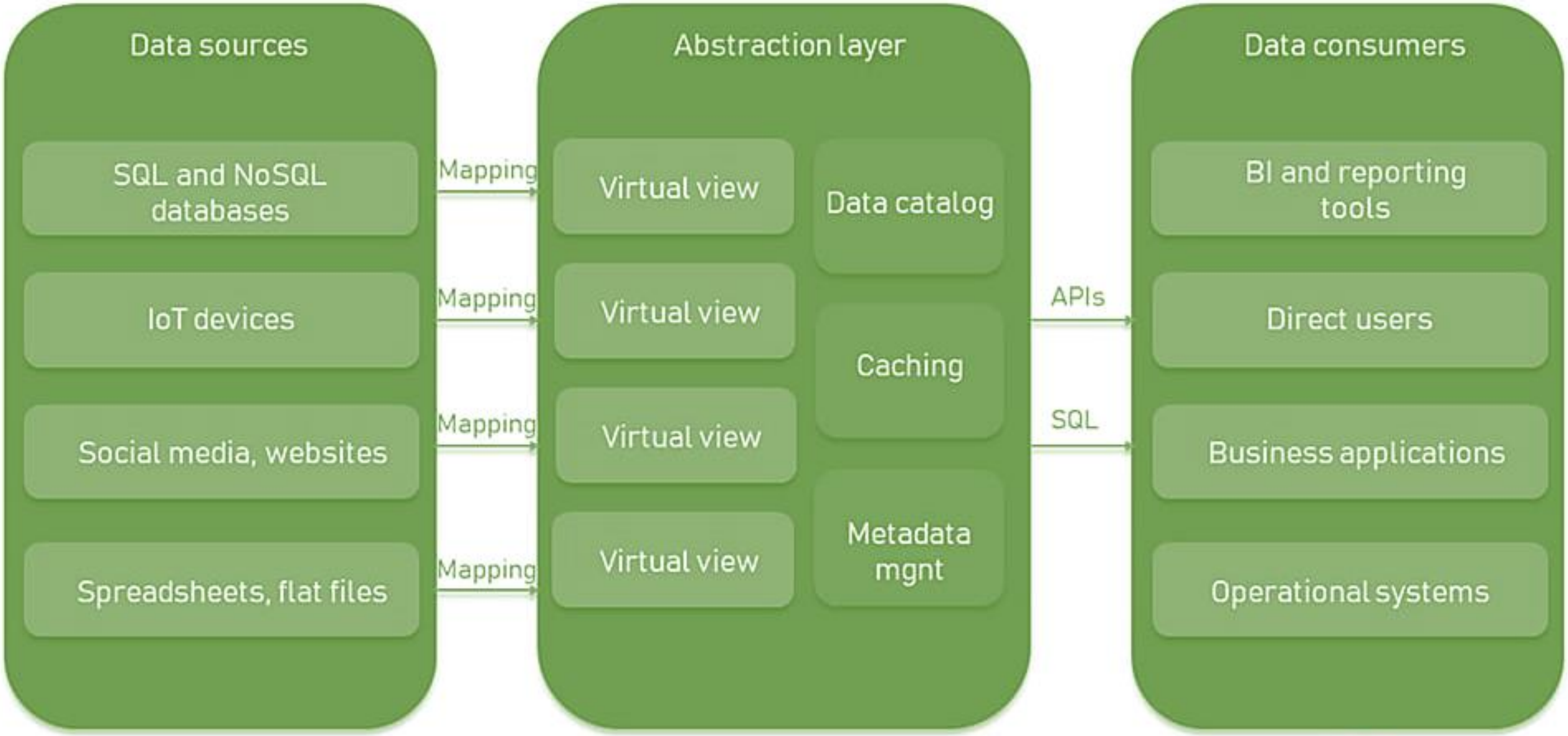
Modern Data Architecture – Enterprise Level



Deployment Types:

- SaaS
- Cloud self-hosted
- Hybrid
- On-Premises

Data Virtualization Architecture



DATA VIRTUALIZATION IN DELL TECHNOLOGIES

The goal of this solution is to **demonstrate the advantages of using data virtualization with Oracle Big Data SQL**. We show how to efficiently access data from multiple sources without introducing extract and load overhead. Our focus is to demonstrate the **functionality of Oracle Big Data SQL using moderately sized sample datasets**. Showing how Oracle Big Data SQL can scale to petabytes of source data is beyond the scope of this solution.

Dell produced an elastic server and storage solution that can be used for data virtualization and to consolidate disparate data sources. To facilitate an **elastic private cloud solution**, **Dell Technologies used the combination of compute nodes with VMware vSphere virtualization and PowerFlex software-defined storage**. This elastic private cloud solution is designed to **provision resources based on the needs of data virtualization**, meaning that growth can be addressed incrementally.

DATA VIRTUALIZATION IN DELL TECHNOLOGIES

The Dell and Starburst partnership marries a **fast query engine** with **high-Performance compute and storage platforms** on 2023

DATA VIRTUALIZATION IN DELL TECHNOLOGIES

The partnership with Starburst allows Dell **to offer customers the ability to use data virtualization across their multi-cloud environments**. Specifically, the **new data virtualization solution uses Dell Technologies hardware and software-driven storage designed to manage data at scale**, including, respectively, PowerEdge servers and ECS object storage.

The solution makes use **of Starburst's ability to query data across any database**, making it instantly actionable for data-driven organizations. Specifically, **Starburst provides a fast and efficient analytics engine for data warehouses**, **data lakes, or data mesh**. It unlocks the value of distributed data by making it fast and easy to access, no matter where it lives.

Starburst is built **on top of Trino, the open-source, high-performance distributed** SQL engine that's known for running fast analytic queries against data sources ranging in size from GBs to PBs. (Trino was formerly called PrestoSQL.)

DATA VIRTUALIZATION - DEMO

Contact Us

☎ 080-4524-9465

✉ support@intellipaat.com