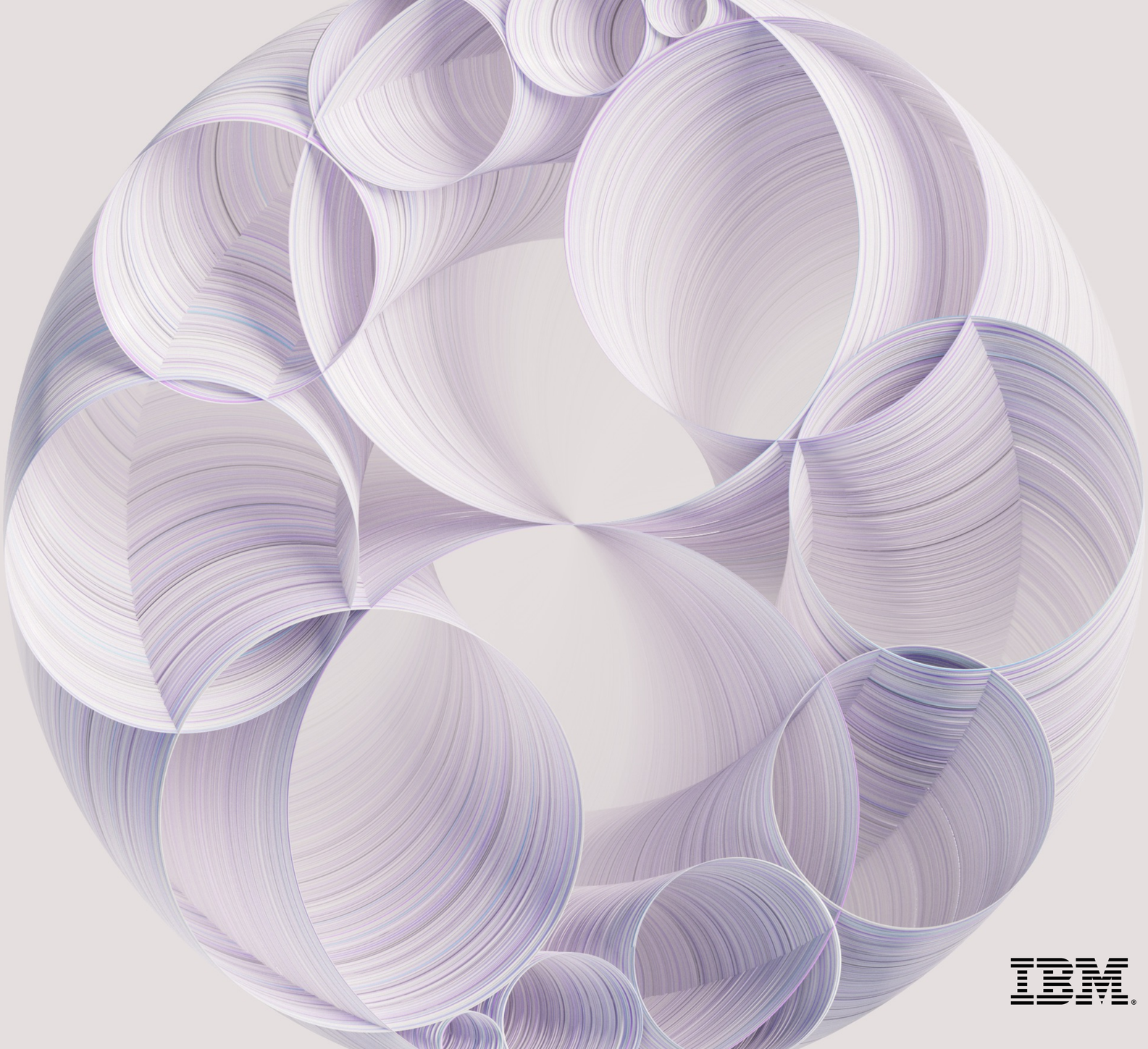
Tech Bites Watsonx.data Use Cases



Kevin Shen
Lead Product Manager watsonx.data
July 2023

watsonx.data™

IBM

Warehouse Offloading for Cost Optimization

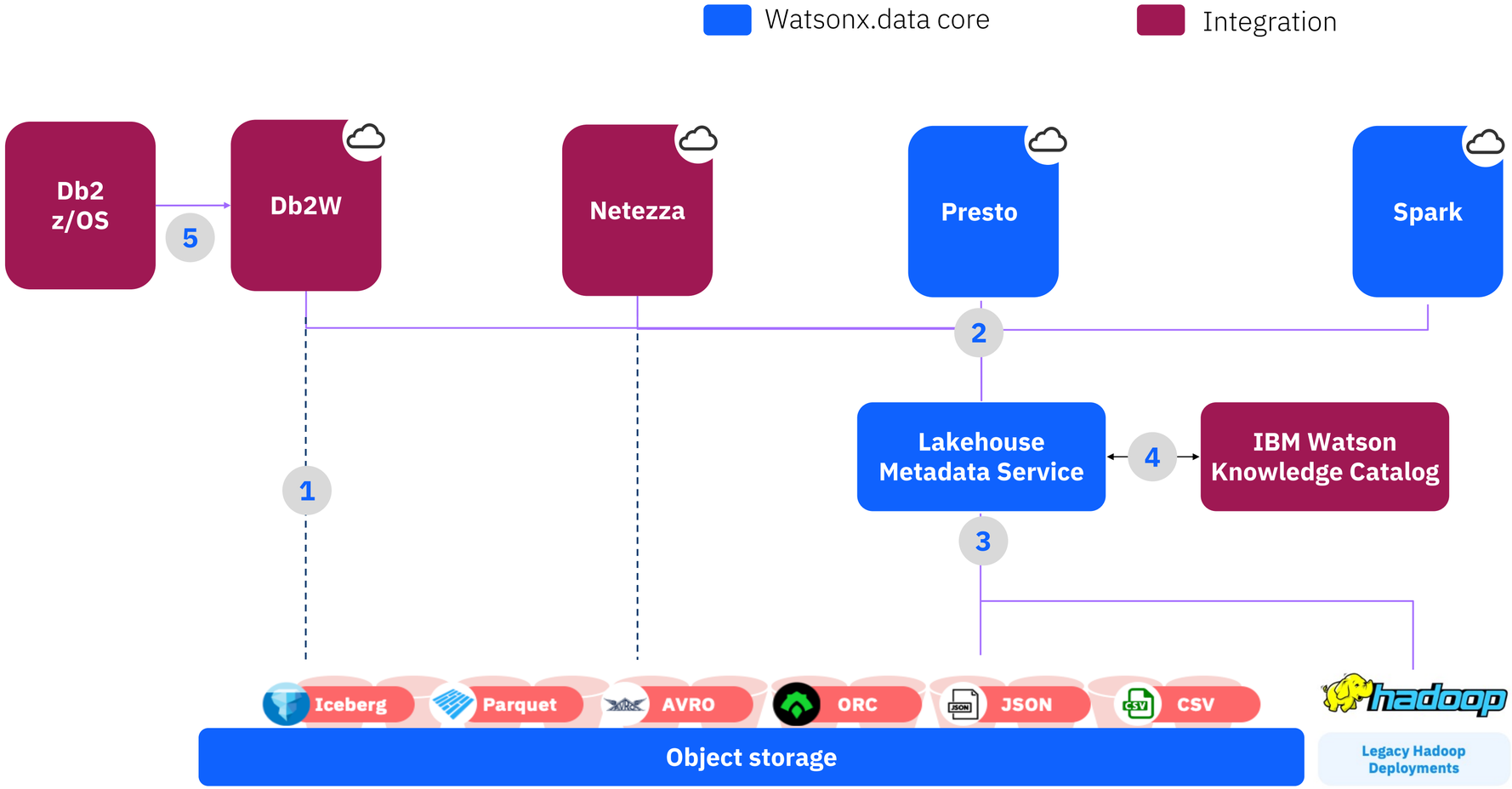
Warehouse offload narrative

- Competitive and IBM (Netezza + Db2)
- Talk track – Client is concerned with the spend on traditional warehouse today – looking to optimize for both performance and cost
- **Value prop:** Cost optimization and openness through the shared meta layer and fit-for-purpose engines
- Example: Snowflake write-intensive workloads moving to Spark and/or Presto, thus educing cost of Snowflake virtual data warehouses

Key Information to Gather

- Understanding current customer warehouse workload
  - A workload analyzer is also being developed to help understand the customer's current workload
  - What workload can be offloaded to Presto or Spark
  - What associated data need to be offloaded as well?

The integrated watsonx.data ecosystem for maximum workload coverage and optimal price-performance



1. Warehouses can **access data in the Lakehouse**
2. Multiple engines can **access same data lake data**
3. The Lakehouse can **access data residing in Db2/Netezza**
4. WKC policies **enforced by the Lakehouse** via metadata service
5. Analyze z data **easily and securely** with “Z Data Ingest”

Unlocking Data Lake Data

Modernizing data lake narrative

- Modernizing storage architecture to facilitate shared metadata and fit-for-purpose engines
- Talk track – Converting legacy file storage structures into open-file structures and assigning those into the shared meta layer, thus facilitating fit-for-purpose engines
- Value prop:
  - Cost optimization and openness through the shared meta layer and fit-for-purpose engines
  - Faster time to value with fewer data movement and transformation
  - Improving the quality of data over time with table formats that brings transactional guarantee
  - Example: Accessing data lake data with data from other sources at the same time

Key Information to Gather

- The type of data lake employed and the access pattern
- Data movement patterns, do data stay in the lake? Do they move between lakes and warehouses?

Data Store for AI + BI Workloads

Data Store for AI narrative

- Generative AI and BI have distinctly different data store requirements
- Talk track – Leveraging watsonx.data as the data store for AI and BI
  - Ability to persist files of various type in object store
  - Ability to process ultra wide tables, distributed queries
  - Unlimited snapshotable storage with Iceberg table format
  - Spark as an ai training engine

- Value prop:
  - Consolidation of data stores. One copy of data for multiple uses
  - Ability to handle the different modal of data

- Focusing on improving this in the future with items such as vector databases

Key Information to Gather

- Where are they preparing data for AI?
- Where are they training and deploying their models?
- Overlap in the data used for their AI training vs BI needs?

Governed Data Access with Data Fabric

Governed Data Access and Sharing with WKC

- Governing disparate data sources is difficult and providing self serviced access while being secure is challenging
- Talk track – Enable your organization to share data freely without concerns over access or governance
- Value prop:
  - Global governance policy that are enforced locally to reduce time and effort sent managing governance policies

Key Information to Gather

- Are they a current WKC customer or looking to adopt?
- What are their current governance solution today

Use Cases we are not supporting fully today, but may in the future

Logical Data Warehouse

What is a Logical Data Warehouse?

- A data management architecture that creates an virtual layer on top of existing data repositories to access data in place. Essentially data virtualization

Why may it come up as a watsonx.data use case?

- The Presto engine operates as a virtualization engine and can perform data virtualization over different sources
- Starburst a major, a vender that supports a branch of Presto called Trino has been strategically focused on virtualization and connector enhancements for the virtualization and logical data warehouse use case

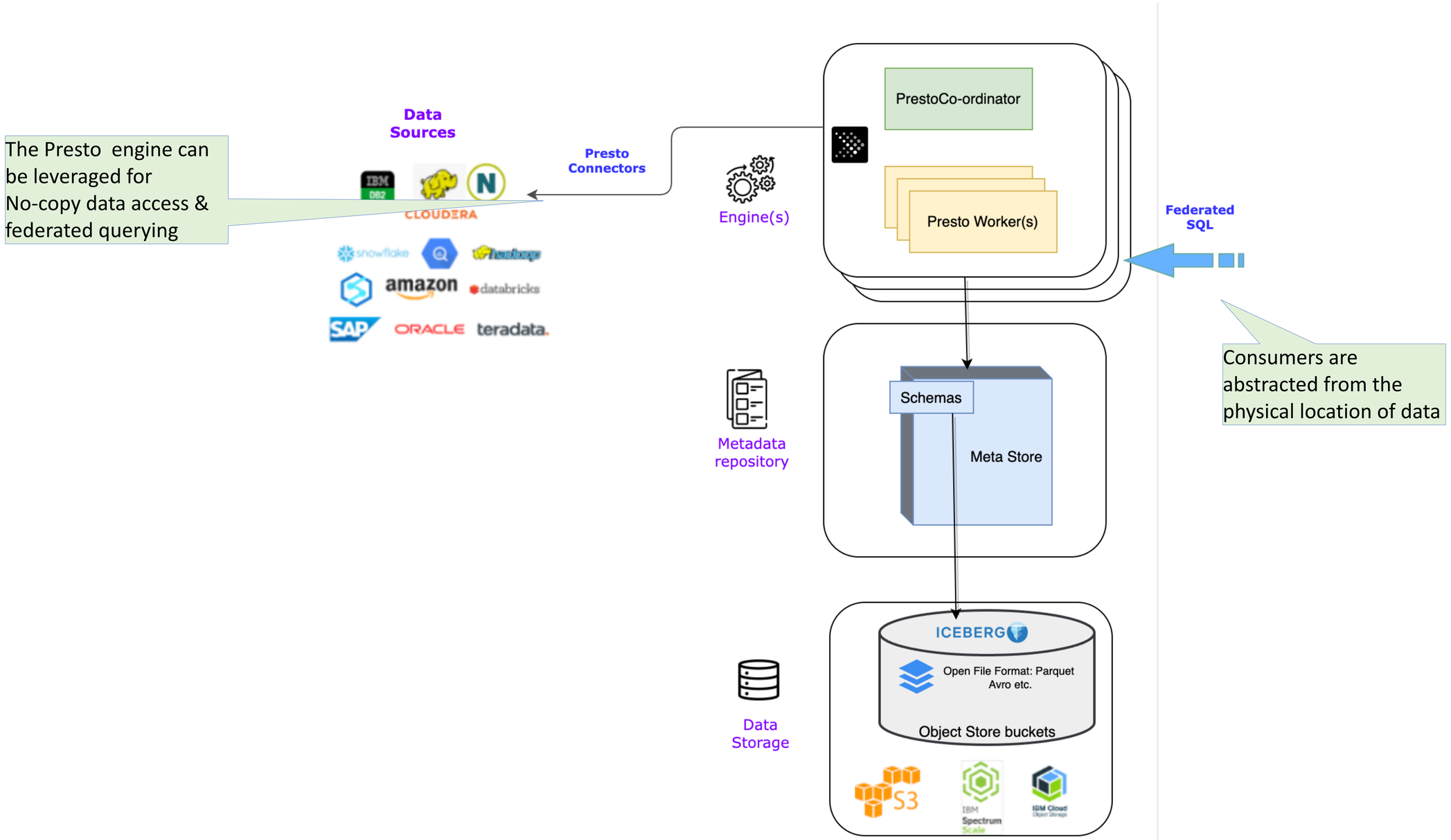
Why are we not currently targeting specifically for this use case?

- Strategically, we are focused on direct data lake access first, with multiple engines accessing the same data source
- We have a story of consolidation as it enables us to provide more than one engine

Future

- The team is exploring plans to bring enhancements to our connectors in the future so we can look to support requirements for this use case

Presto Connectors - Access Data in-place



Replacing Warehouse or Hadoop

Why are we not replacing?

- Watsonx.data offers engines that are great at supporting large queries over modern data formats and it will not offer the same SLA as a traditional warehouse
- Customers may have petabytes of data in their data lake and will unlikely move all of it in a short time frame
- Workload migration may be more challenging and difficult than the data migration

Why may it come up as a watsonx.data use case?

- Other data lakehouse vendors such as Dremio and Databricks will message how they can replace a data lake and a warehouse with their solution
- Customer have a desire to consolidate and reduce cost

Future

- The team is actively working on the Velox engine, which will bring warehouse “like” performance to the presto engine
- Customers may slowly shift their data from their existing data lake to watsonx.data once we prove out data access to their lake and the value our integrations brings

Tech Bites

July 25th

Solution Engineering
Pradeep Kutty
Rob Wilson

watsonx.data – Use Case Positioning

Use Cases	Modernization / optimization path for existing Db2 Clients	Modernization / optimization path for existing Netezza Clients	Optimization path for existing Cloud Data Warehouses (i.e., Snowflake)	Modernization path for existing Hadoop Data Lakes
AI / ML at Scale , Share Data Responsibly *Large data processing workloads for Machine learning and AI*	*Workloads often require large volumes of data and significant computational resources – **Share data with watsonx.data***	*Workloads often require large volumes of data and significant computational resources – **Share data with watsonx.data***	*Reduce storage and compute cost - **Move data processing to watsonx.data and keep consumption layer on top of Snowflake.***	*If client's intent is to modernize legacy workloads and/or migrate to the cloud, **leverage watsonx.data** and platform level Data & AI capabilities as a cost-effective solution.*
Real Time Analytics *Analytics and reporting uses cases*	*Workload should stay on Db2*	*Workload should stay on Netezza*	***If cost is a driver, consider watsonx platform**. Augment NZ with **watsonx.data**- Such as move infrequent data to watsonx.data*	
Real Time Analytics *Operational analytics (ODS)*	*Workload should stay on Db2*	*For ODS requirements, augment with **watsonx.data** leveraging Db2 as the fit for purpose engine*	*Augment with **watsonx.data** leveraging Db2 as the fit for purpose engine*	
Streamline Data Engineering *Data Transformation and ELT workloads (Write Intensive)*	*Reduce storage and compute cost – **Perform ELT operations in watsonx.data and promote as needed into warehouse.***	*Reduce storage and compute cost – **Perform ELT operations in watsonx.data and promote as needed into warehouse.***	*Reduce storage & compute cost. **Leverage watsonx.data to transform and filter data before it is loaded into Snowflake.***	
BI , Share Data Responsibly *Data Exploration and Visualization*	*Reduce storage and compute cost – **Gain novel insights by joining real time data from watsonx.data with your proprietary warehouse data.***	*Reduce storage and compute cost – **Gain novel insights by joining real time data from watsonx.data with your proprietary warehouse data.***	*Reduce storage and compute cost - **Move data processing to watsonx.data and keep consumption layer on top of Snowflake.***	

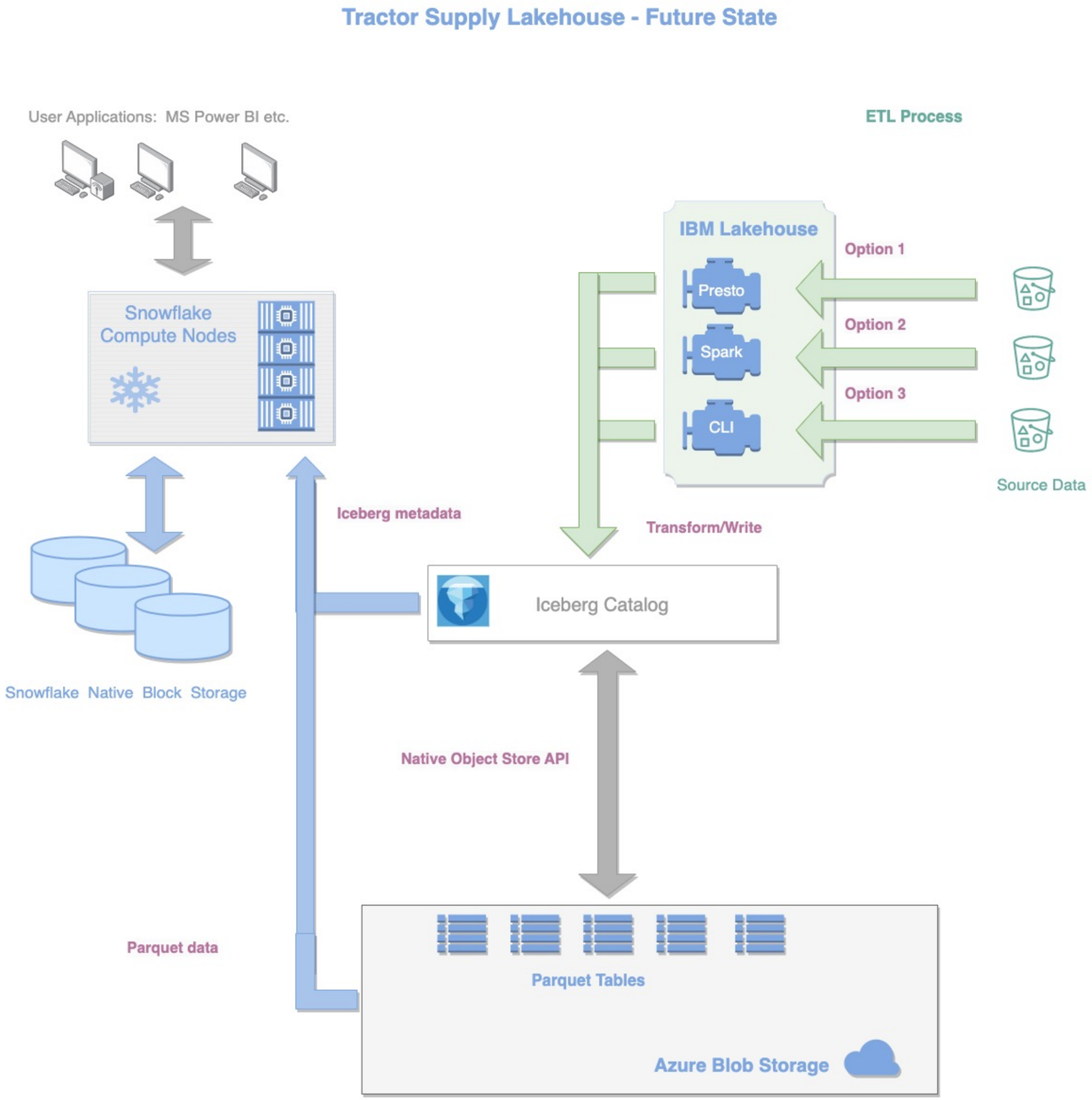
Client Engagements – watsonx.data Use Case Patterns

Integration CPD + Other IBM Solutions	AI Solutions	EDW Augmentation	Takeout's	BI / Analytics	Hadoop Migration / Modernize
➢ US Coast Guard ➢ Lockheed Martin ➢ IBM Semiconductor ➢ Bank of America ➢ Lumen ➢ Honda ➢ E&Y (UKI)	➢ IBM Semiconductor ➢ Samsung ➢ Comparus ➢ UBS ➢ Wipro	➢ Tractor Supply (Snowflake) ➢ Toronto Hydro (Netezza) ➢ HSBC (Teradata) ➢ Toyota (Netezza)	➢ AmeriSource Bergin (Databricks) ➢ CITI (Starburst)	➢ AMC Networks	➢ Wandisco (P) ➢ nFolks (P) ➢ NucleusTeq (P) ➢ Etisalat

What's working?

- Lakehouse **messaging around augmentation strategy** is resonating.
- Clients **loved the UI.**
- Clients liked **versatility of deployments** – stand alone and development images.
- **Positive install experience** – Takes hours v/s days (Comparing to CPD installs)
- **Semantic enrichment** was a huge hit in the demos !

Tractor Supply: Snowflake Augmentation with Watsonx.data



Data Reads

Current Solution
- Snowflake on Azure with Native Block Storage
- MS Power BI

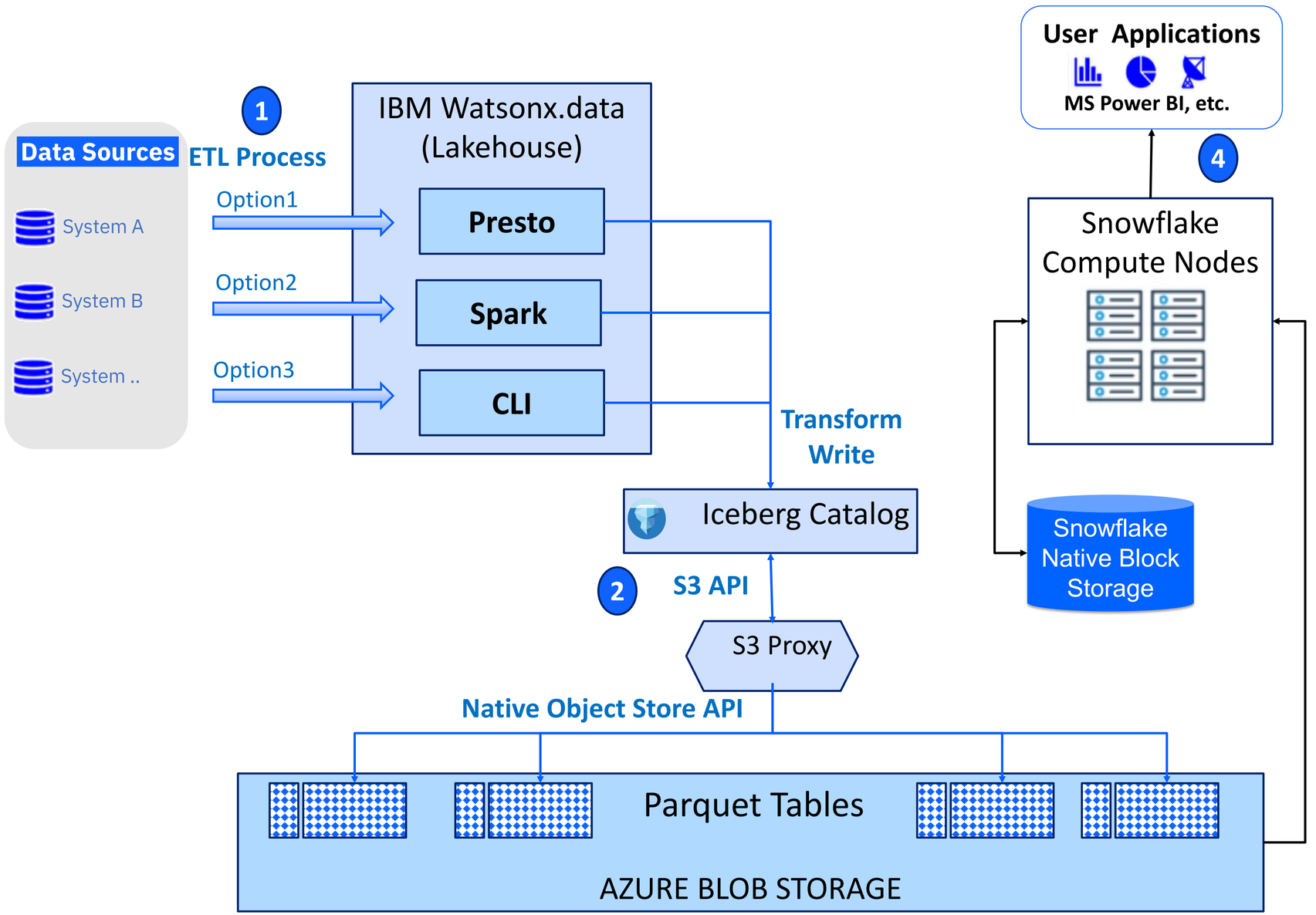
Challenges:
- ETL workloads (write intensive) are driving high computing and storage costs

Proposed Solution (in Evaluation)
- Keep Power BI workloads on Snowflake
- **Offload** write intensive workloads (ETL) to **Watsonx.data**
- Azure Object Storage with Parquet files

Benefits:
- Reduce computing and storage costs

***Status**: Evaluation using Beta code completed. Client waiting for GA version which will address some performance requirements*

Tractor Supply: Snowflake Augmentation with Watsonx.data



Current Solution
- Snowflake on Azure with Native Block Storage
- MS Power BI

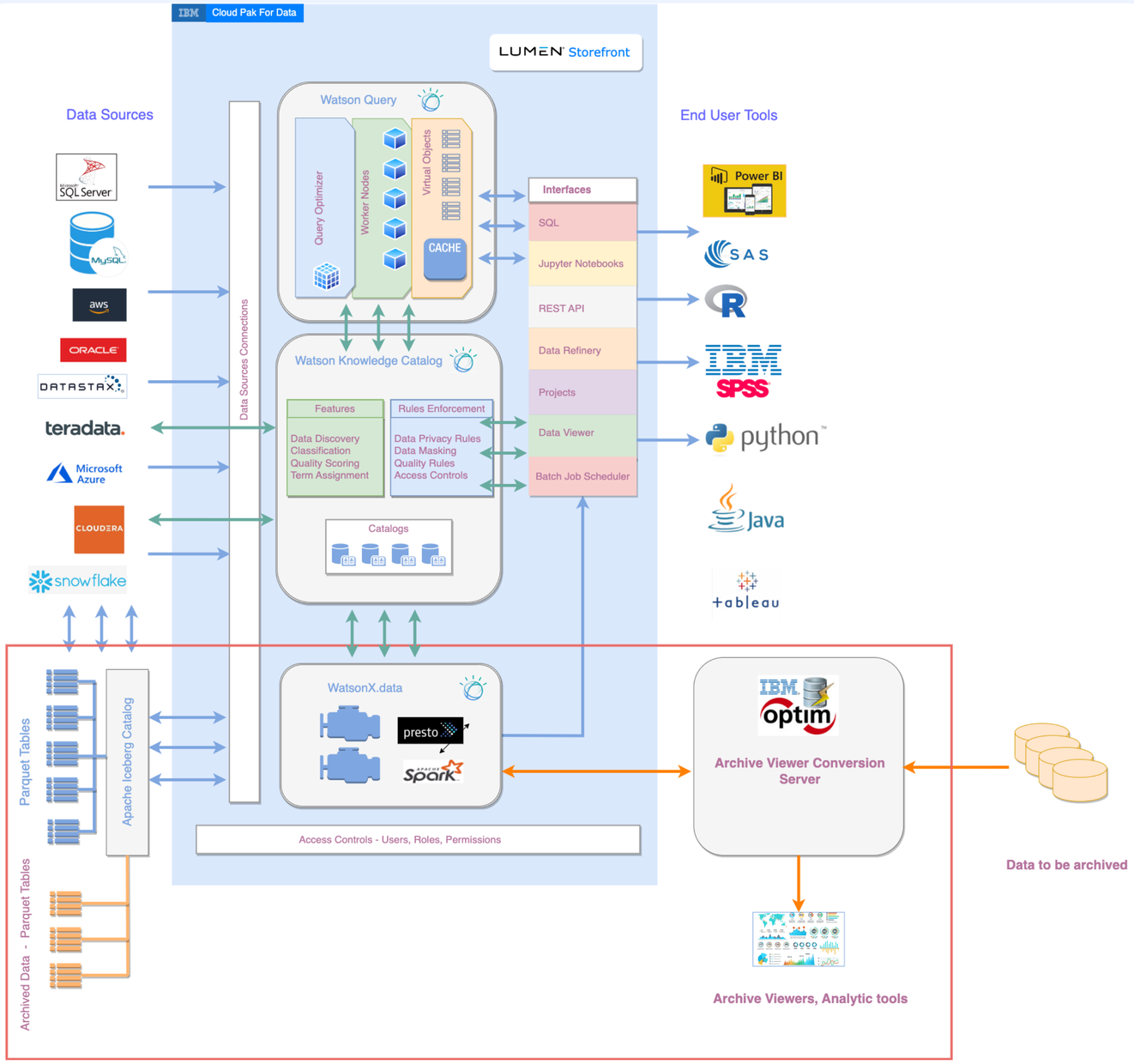
Challenges:
- ETL workloads (write intensive) are driving high computing and storage costs

Proposed Solution (in Evaluation)
- Keep Power BI workloads on Snowflake
- **Offload** write intensive workloads (ETL) to **Watsonx.data**
- Azure Object Storage with Parquet files

Benefits:
- Reduce computing and storage costs

***Status**: Evaluation using Beta code completed. Client waiting for GA version which will address some performance requirements*

Lumen - Storefront Platform



watsonx.data:

- Introducing new Compute and Storage capabilities
- Presto and Spark engines tightly integrated with the rest of the platform
- S3 storage with open-source data format – Parquet
- Apache Iceberg catalog, - provides consistency layer for multi-engine access to parquet tables
- Integrated with Optim Data Archive tools
- Cloud agnostic platform – same capabilities available in AWS, IBM Cloud, Azure, on-prem.
- Hybrid options!