

//* IBM Z Day
//* open | on | secure

October 1, 2024

IBM TechXchange Virtual Event



IBM Z data, analytics and mastering the rare art of AI model deployment

Eric Siegel
Jonathan Sloan

Agenda

Predictive AI's value

Predictive AI's deployment challenge

The **bizML** playbook

Mainframe data

A step back in time

William Francis Sutton Jr. was an American bank robber. During his forty-year robbery career he stole an estimated \$2 million, and he eventually spent more than half of his adult life in prison and escaped three times. For his talent at executing robberies in disguises, he gained two nicknames, "Willie the Actor" and "Slick Willie".



Willie Sutton

Why is this interesting?

Willie Sutton wasn't necessarily as famous for his bank robbing as much as his ability to get to the point. When Willie was finally captured, a reporter asked him, "Mr. Sutton, why do you rob banks?" He replied ...

Because that's where the money is!

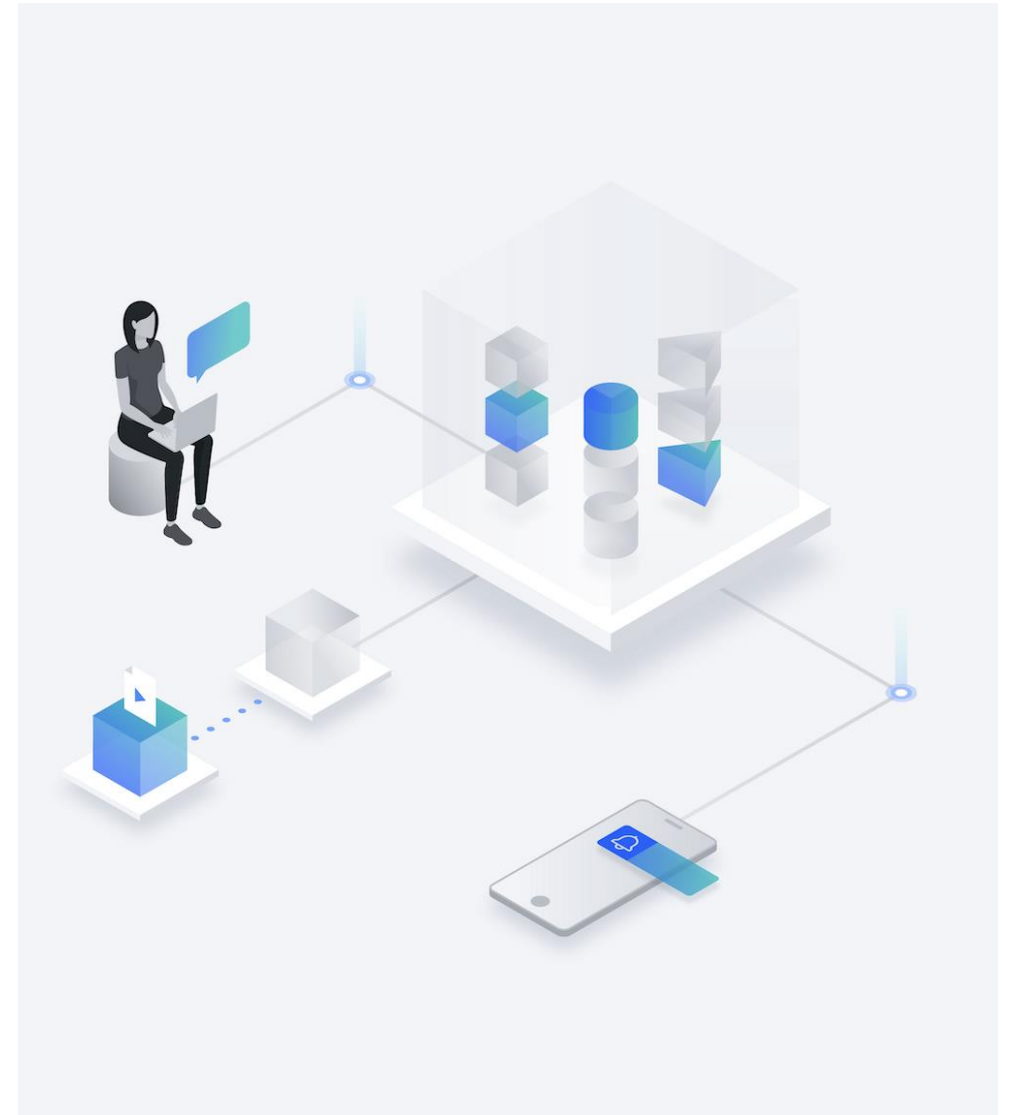


Willie Sutton

Why talk about the mainframe?
Because that's where the money is....

Transactional application data is the *definitive source of truth*, represents *the current state* of the business and provides *unique predictive value* for business outcomes.

With this data, analytics and AI applications can help drive greater revenue, improve productivity, and better manage cost.



For most enterprise size organizations, transactional data originates on IBM Z

IBM Telum Processor

Powering IBM z16 &

Featuring a dedicated on-chip AI accelerator



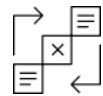
With IBM z16, process up to 3.5million inference requests per second with 1ms response time using a Credit Card Fraud Detection model.



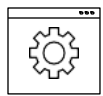
Very low and consistent inference latency and scalable capacity



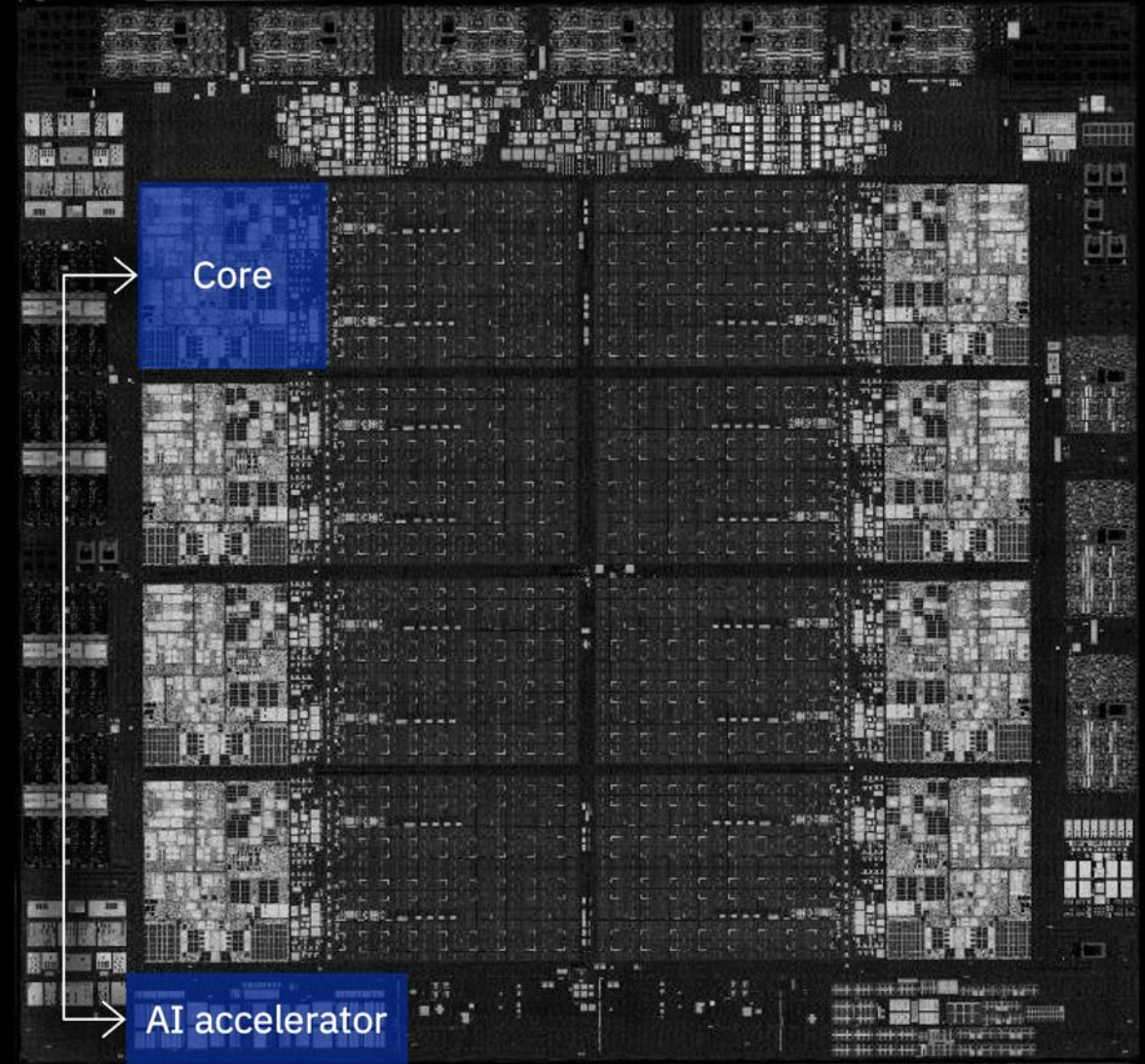
Fast direct storage access through new cache design



Security – enterprise-grade memory virtualization and protection



On-chip AI-accelerator that works with a standard z-Core general processor



IBM z16 with on-chip AI accelerator

228K z/OS CICS

credit card transactions per second with 6ms latency with in-transaction fraud detection inferences using DL model.

< 8% increase

in latency for CICS credit card transactions adding in-transaction fraud detection with avg response time of 30ms

20x lower response time

when co-locating apps and inferencing requests compared to x86 server in same data center with avg latency of 5ms

19x higher throughput

when co-locating apps and inferencing requests compared to x86 server in same data center with avg latency of 5ms.

41x lower energy consumption

Utilized by z16 compared to remote x86 server using NVIDIA GPU



IBM Z OPTIMIZED OPEN-SOURCE ECOSYSTEM

65%

of organizations to use opensource
for their AI Implementations, up
from 48% in 2021

The State of Enterprise Open Source, RedHat Report

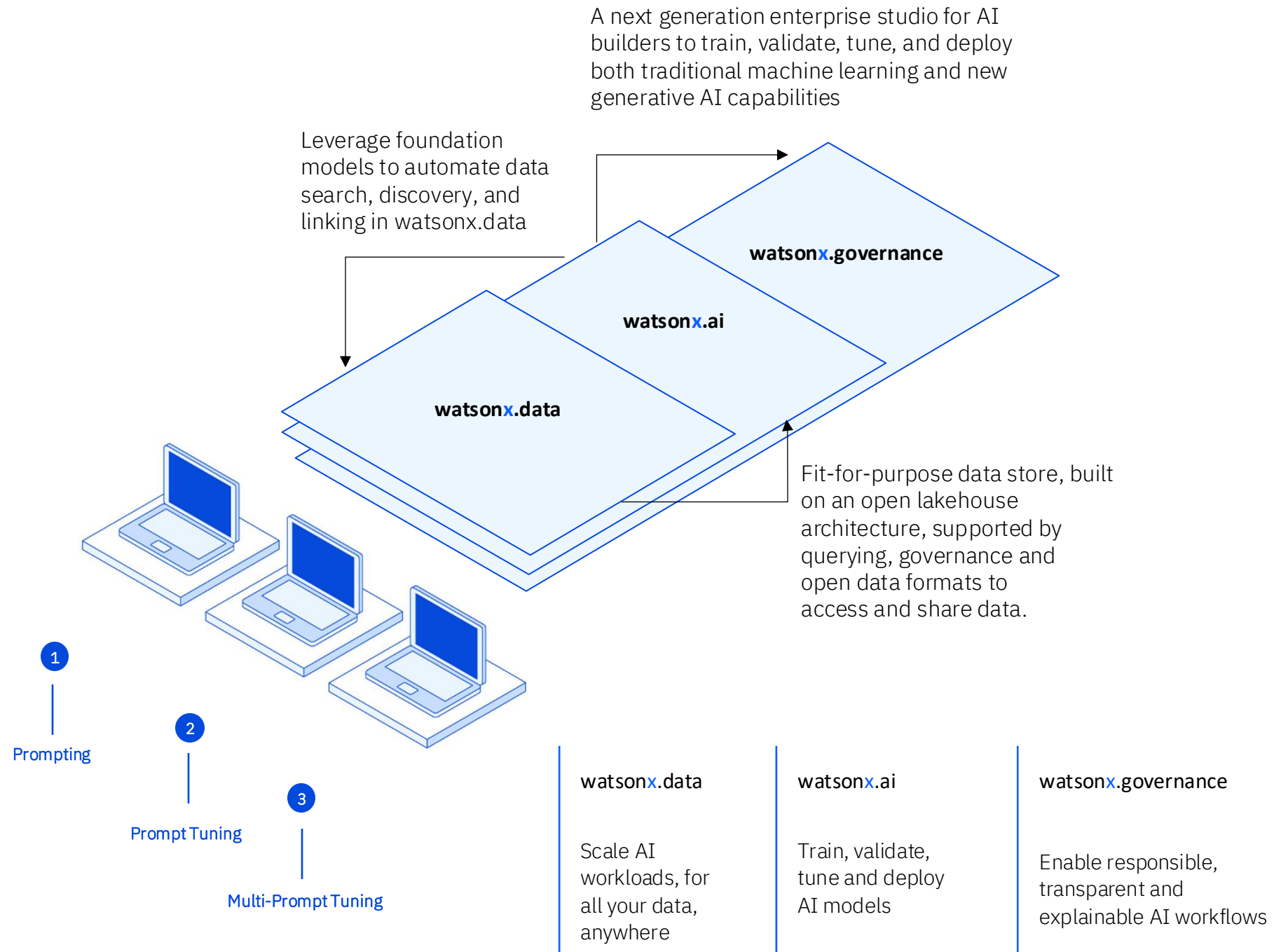


- Optimized to leverage IBM Z hardware innovations
- Delivered securely
- Utilized in IBM Z and LinuxONE offerings, including:
 - Machine Learning for z/OS
 - Cloud Pak for Data
 - AI Toolkit for IBM Z and LinuxONE
 - Python AI Toolkit for IBM z/OS

The platform for AI and data |

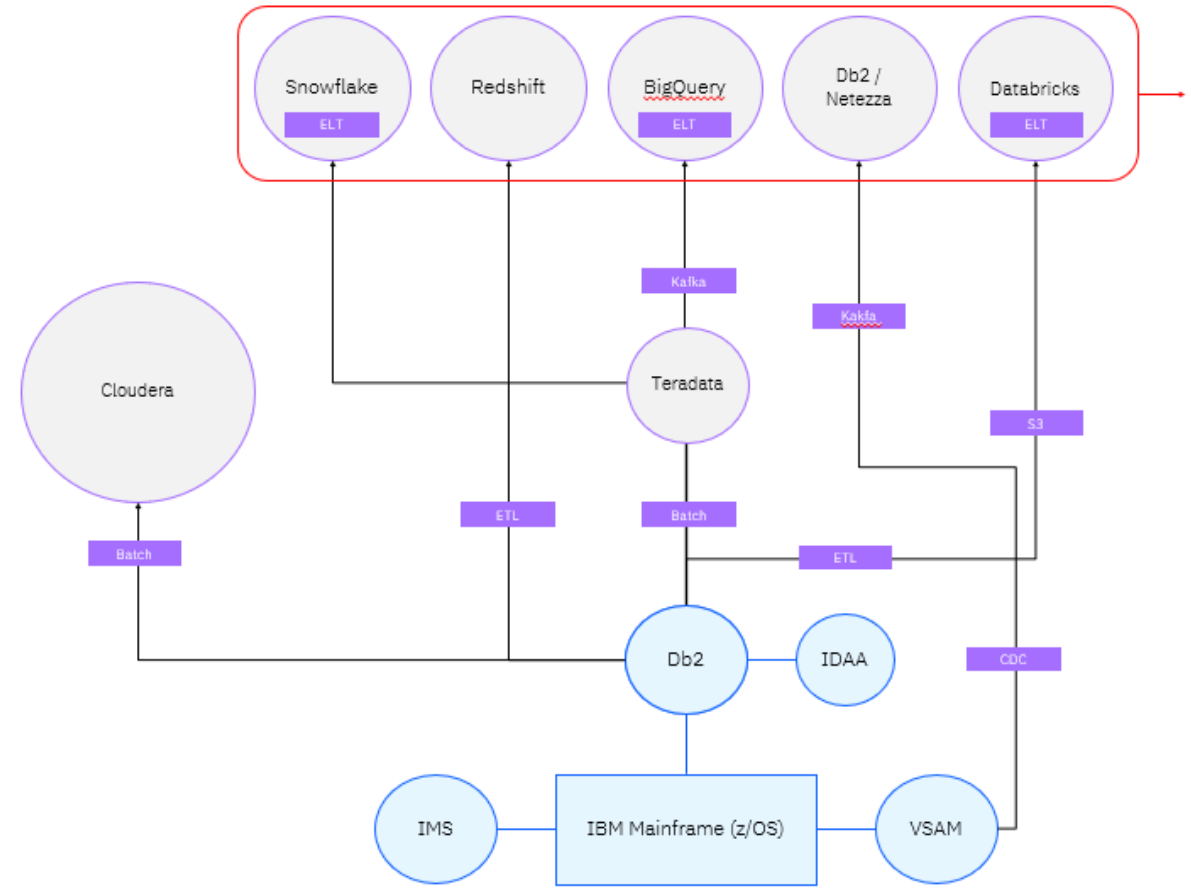
watsonx

IBM's next
generation data,
AI and
governance
platform



Historically, it has been difficult and expensive to move data from the mainframe to other platforms for analysis and model building

- Data can be complex to access
- Data resides in unfamiliar data sources
- Every unique application has its own specific data transformation requirements
- Data replication/extraction competes for resources with operational applications
- IBM studies have shown that extract, transform and moving processes can consume as much as 1/3 of mainframe processing





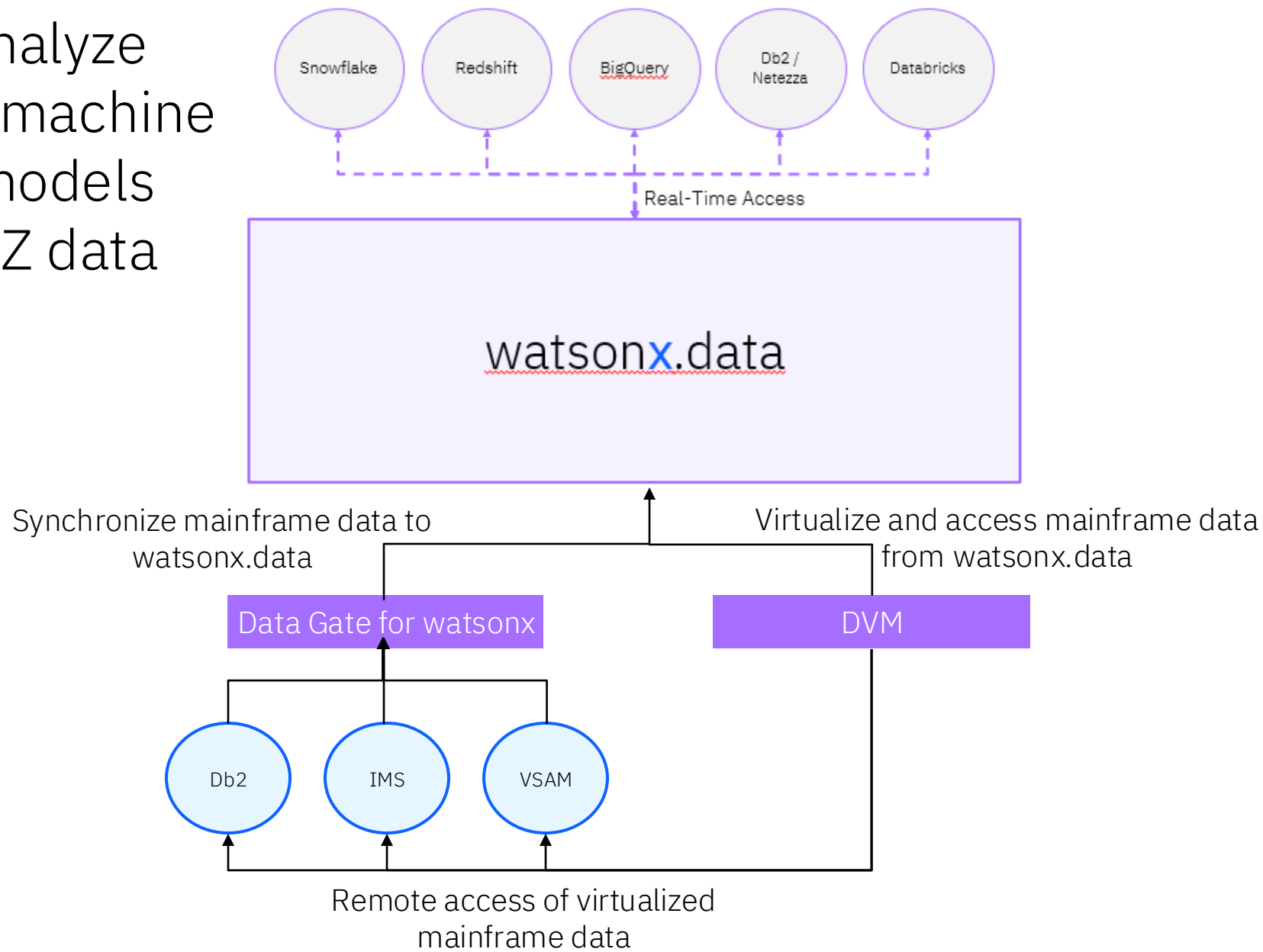
Why follow a decades old architectural pattern of moving data?

You can build and deploy models on the mainframe!

And now there are more modern, effective, lower cost and better ways of providing mainframe data for analytics and AI.

IBM Data Gate for watsonx and
IBM Data Virtualization Manager for z/OS

Access, analyze
and build machine
learning models
from IBM Z data



IBM® Data Gate for watsonx™

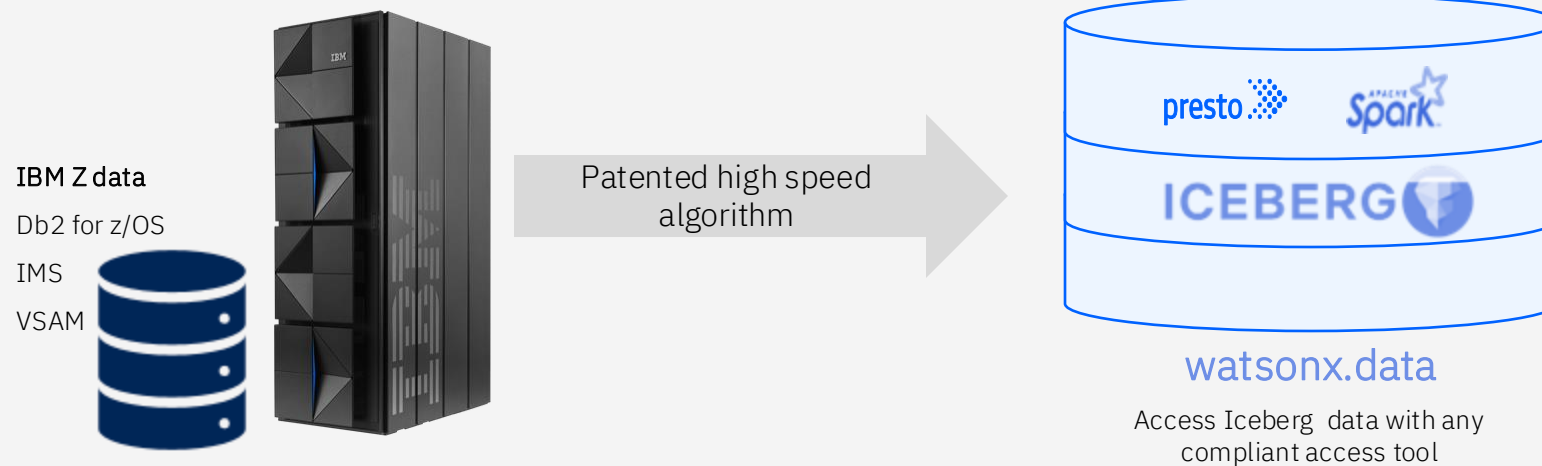


Bring the most up-to-date critical,
IBM Z transactional Db2 for z/OS,
IMS and VSAM data to watsonx

IBM Data Gate for watsonx

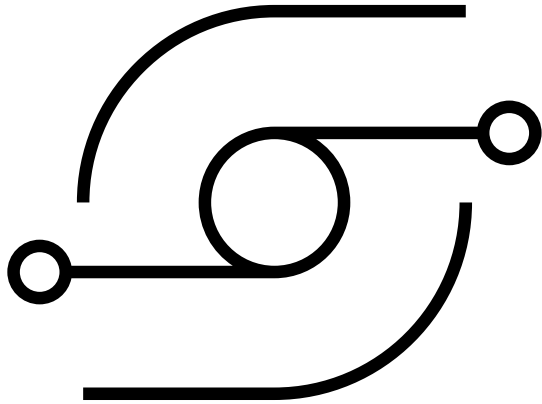
Synchronize Db2 for z/OS, IMS and VSAM IBM Z data sources to the Iceberg open table data format

- Readily incorporate IBM Z data within watsonx.data
- Revolutionarily simple, efficient data synchronization
- Deliver up-to-date data for analytics and AI models
- Apply as much compute for complex analysis as you need without impacting mainframe processing



Integrated Synchronization

A cost effective, highly optimized, purpose built, unidirectional log data provider capturing Db2 log changes. Through Integrated Synchronization, Data Gate can provide high-speed integration of Db2 for z/OS data with watsonx.data. It allows up-to-date data to be consumed in the lakehouse in a seamless and efficient way.



- 1 Simplicity - A radically simplified user experience for synchronizing Db2 for z/OS tables and keeping them transactionally consistent, drastically reducing the number of processing steps as compared to full-featured data replication solutions. Data capture is built directly into Db2 for z/OS, so there is nothing to install or administer!



-
- 2 Isolation – Access Db2 for z/OS data without any impact to transactional workloads running on IBM Z



-
- 3 Low latency – An integrated, patented synchronization protocol maintains data currency, with customers today experiencing rates of 1.9M rows/sec and 11TB/h. Optional protocol to ensure that most recent data has been written before executing query against Db2 (“replication-aware query execution”)



IBM Data Gate for watsonx - streamline access to mainframe data sources

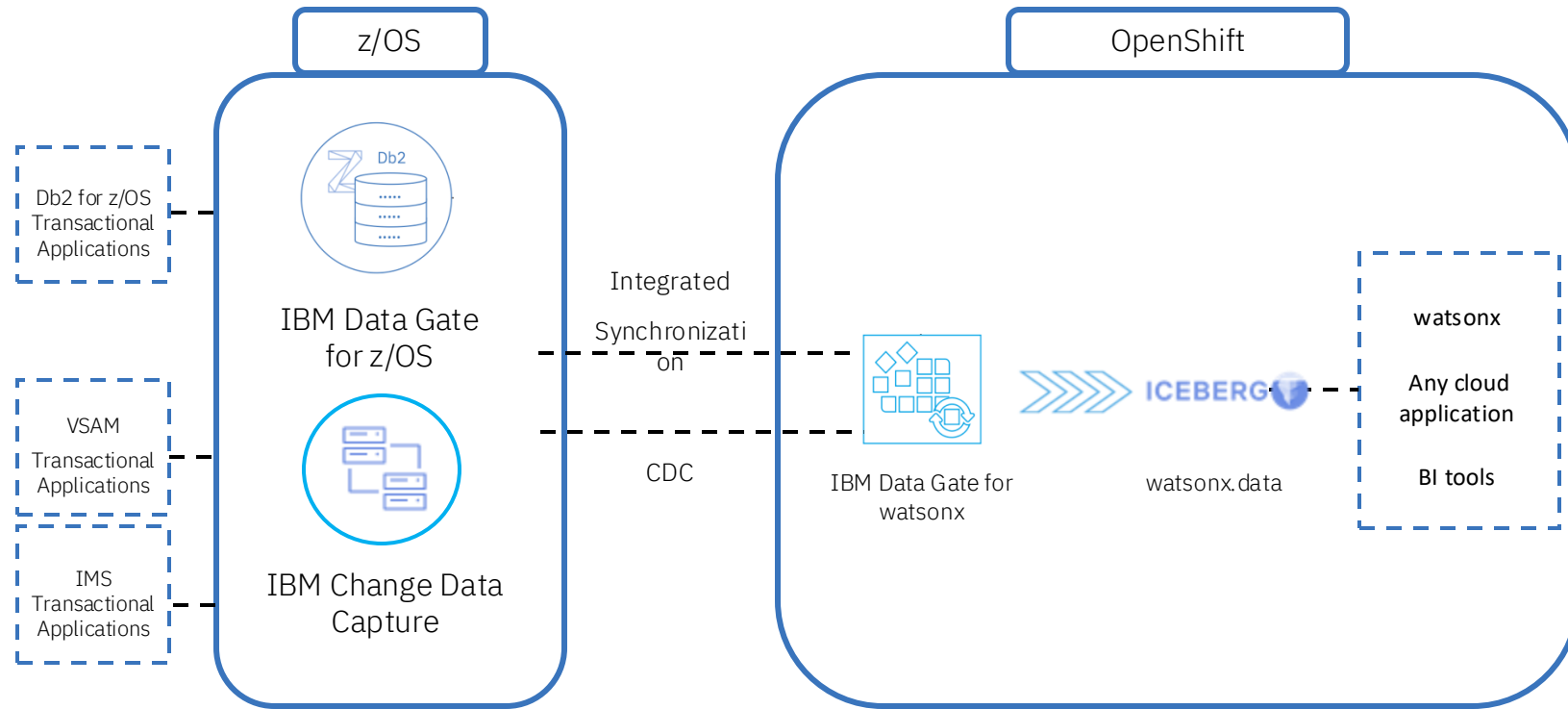
- Reduce the cost of mainframe data movement
- Simplify access to unfamiliar mainframe data sources
- Improve the efficiency of data transformations
- Isolate transformations from operational processing
- Free up skilled resources to work on operational application development and management



IBM Data Gate for watsonx

- Synchronize Db2 for z/OS, IMS and VSAM data:
 - Write directly to the Iceberg open format data files
 - Ingest into IBM watsonx.data lakehouse via Presto or Spark

Generally available as of:
June 2024



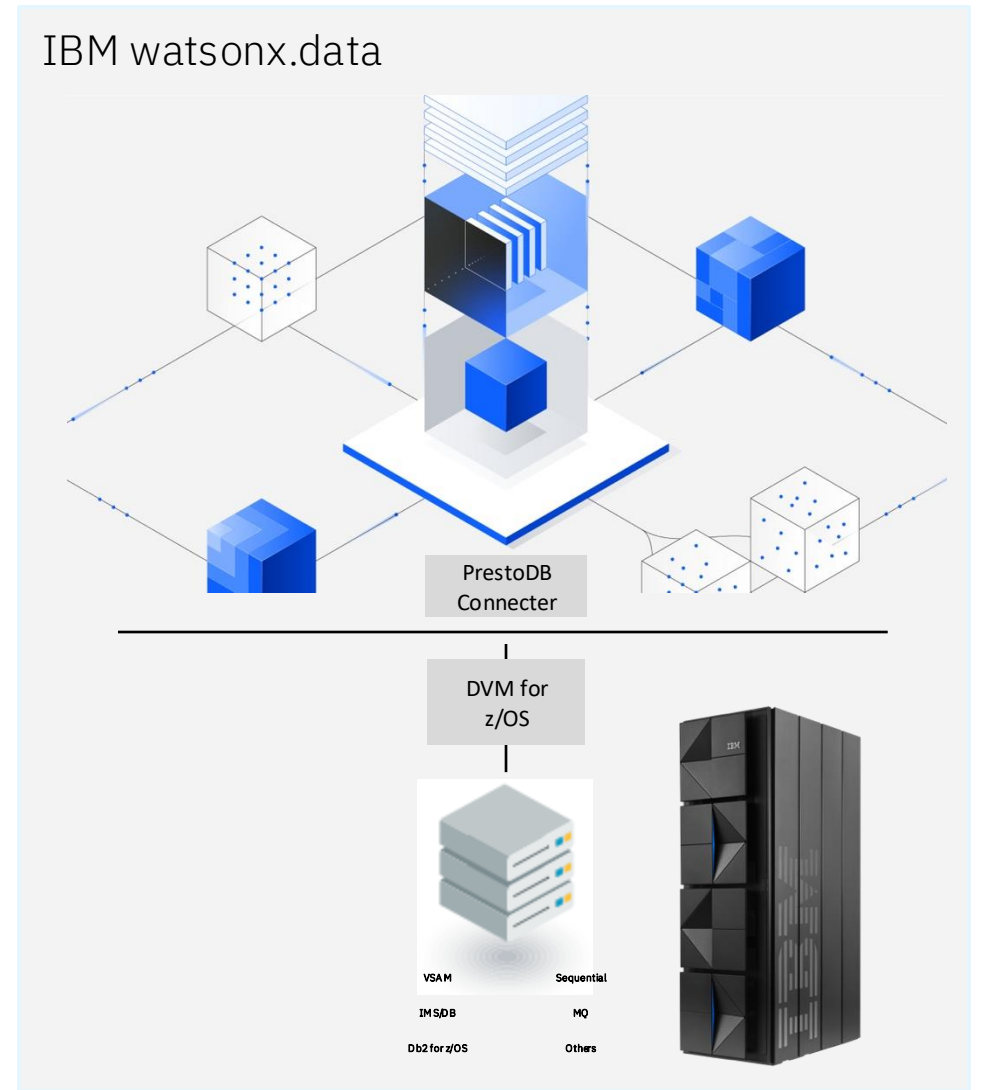
IBM Data Virtualization Manager for z/OS

Granular, virtualized, real-time access to
mainframe data

AI initiatives with IBM watsonx.data

Integrate IBM Z data within your data lakehouse

- Seamlessly access IBM Z data via PrestoDB Connector
- Consume less resources with bulk data moves as compared to DIY ETL approaches
- No recovery logging necessary
- Record selectivity and point in time access to data
- Reduce the time for users to access and analyze data from mainframe systems
- Apply significant compute resources to a complex analysis without impacting mainframe operational processing



A data lakehouse combines the high-performance characteristics of a data warehouse with the cost-efficiency, flexibility, and scalability of a data lake.

