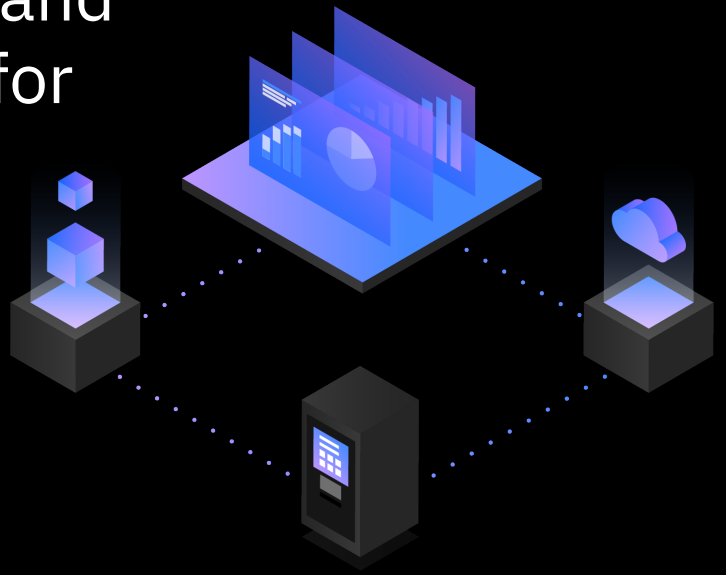
Data Gravity, Data Virtualization and Data Fabric – What It All Means for z/OS and for the Enterprise



IBM Z Council – West Coast

March 17, 2022

Robert Catterall, IBM
Senior Consulting Db2 for z/OS Specialist

IBM

First concepts, then context

- The concepts of data gravity, data virtualization and data fabric are important to understand in a non-platform-specific way
  - Why? Because they apply to all manner of data-hosting platforms, not just IBM Z
- With an understanding of the cross-platform importance of the concepts, we can look at them specifically in a Z context, and highlight enabling technologies

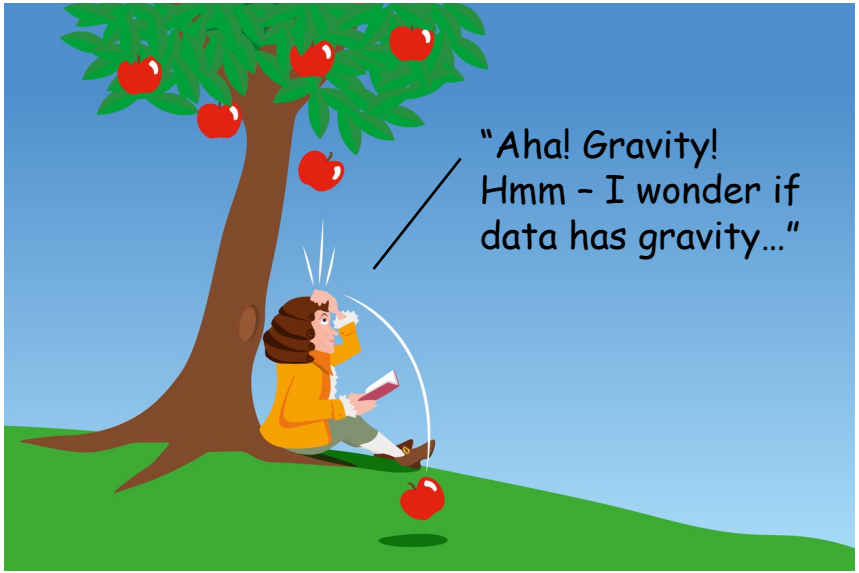
Agenda

- Data gravity, data virtualization and data fabric – a non-platform-specific view
- Making the concepts real for IBM Z and z/OS

Data gravity, data virtualization and data fabric – a non-platform-specific view

Data gravity

- In the physical world, gravity is related to mass – the greater the mass, the greater the gravitational pull
- The data gravity concept is similar: the greater the amount of data on a "system of origin," the greater the "pull" of that data
- What that means:
  - Instead of fighting gravity by copying the data to other platforms for various types of processing (reporting, ad-hoc query, machine learning, specialized applications – whatever), *bring the processing to the data*



The costs of fighting data gravity…

- Copying data from its system of origin to other platforms so that it can be accessed for various purposes has associated costs:
  - From a data security perspective, it "increases the threat area" – the greater the number of places in which a set of data is stored, the greater the chances that the data could be accessed in non-approved and potentially harmful ways
  - Data consistency challenges: multiple copies of a set of data can become inconsistent, and inconsistency can lead users to lose trust in the data, and data that is not trusted tends not to be used
  - Data latency challenges: data copied to another system could be hours or more behind currency with respect to the source data, and increasingly users want access to data that is up-to-the-second current

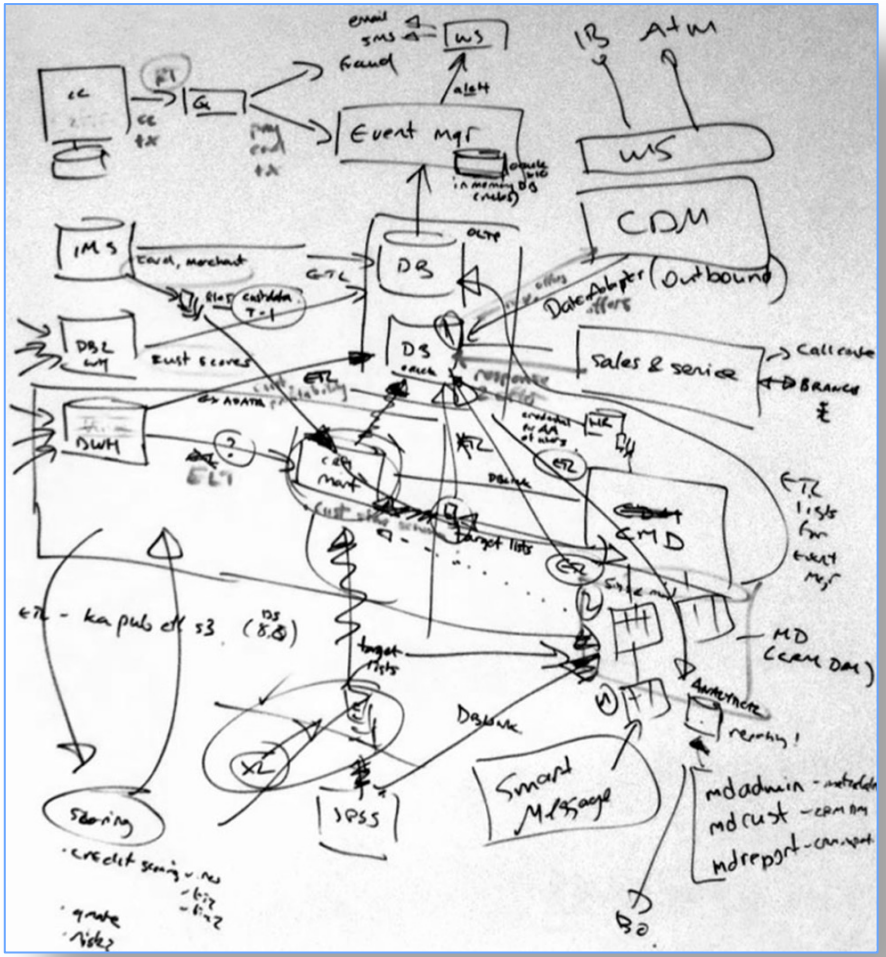
The costs of fighting data gravity (2)

- Data governance challenges: with multiple copies of a set of data, will there be consistency with regard to data cataloging, data discovery, data classification, and so on?
- "Hard" costs: copying data from systems of origin to other systems to support various uses has very real resource costs:
  - Hardware expenses (compute, storage, network)
  - Software expenses (replication and ETL, and OS and DBMS software on target systems)
  - Personnel expenses, to manage replication/ETL processes and to manage target systems
- And, of course…

Complexity

Does this look familiar?

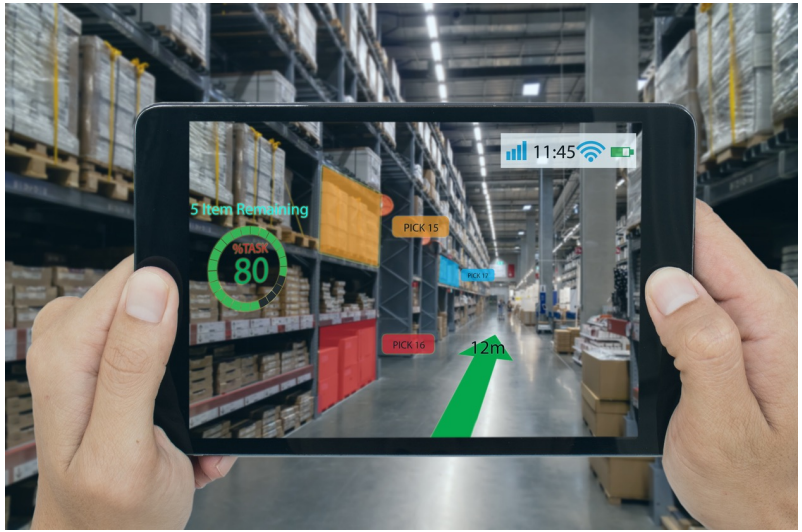
In-place access to system-of-origin data can significantly simplify and streamline an enterprise's IT infrastructure



Data virtualization – the augmented reality analogy

- In this picture, the room is real, but the furniture in the room is not
  - Furniture itself could be real, but pieces are in another location – augmented reality makes it appear that the real (but remote) furniture is in the real room
- So it is with data virtualization – the technology can make it appear that data actually stored at location A is present alongside other data at location B
  - Because the two sets of data appear to be co-located, users can perform joins of data in the two sets, and other things that would be otherwise not do-able

Another data virtualization effect



- In this example, augmented reality is changing the appearance of what is really there – enriching real on-the-shelves items with informational color
- So it is with data virtualization – the technology can change the way a data source appears to a data-consuming user or application
  - For example, data virtualization technology can make data in a file appear to be data in a relational database management system, and therefore accessible using a SQL interface such as JDBC or ODBC

Data virtualization and data gravity

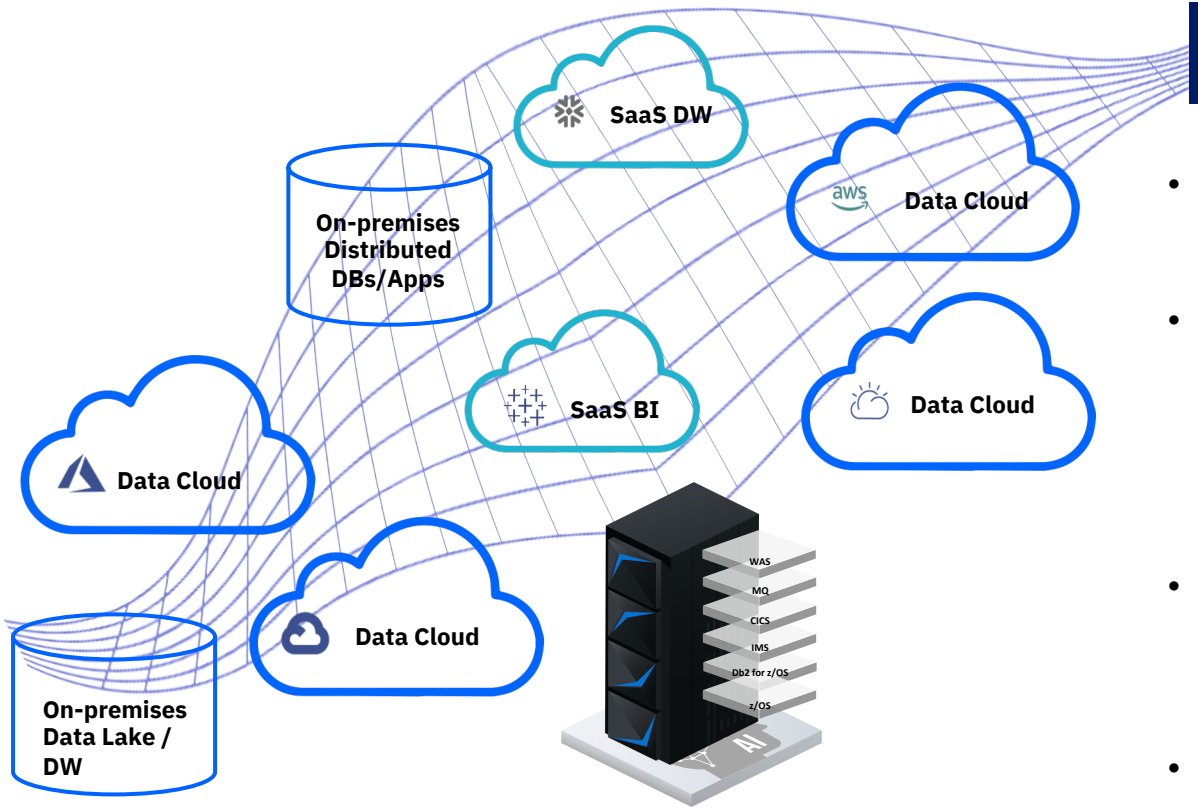
- Data virtualization is very important when it comes to working with data gravity and not against it
- With effective data virtualization, in-place access to data on the system of origin becomes much more feasible from a user's (or application's) perspective:
  - Data can be accessed in-place at location A and at location B, but can appear co-located to a user or an application
  - Data that can be in a not-easily-accessible form on its system of origin can be made to appear in a much more easily-accessible form, without having to be copied to another system

Data fabric – like a smart universal remote for your data

- Not long ago, you needed separate remotes (i.e., separate user interfaces) for various devices and systems in your home
  - For your TV, your stereo, your cable converter, your air conditioner/heater, etc.
  - And, these remotes were dumb, and often a pain to use

- Today, a single smart universal remote (a “virtual assistant”) can manage all kinds of devices and systems, and it has a good bit of intelligence (it learns), and it has a really user-friendly interface (you can talk to it)
- So it is with data fabric…

Uniform, intelligent, user-friendly



Data fabric architecture

- An abstraction layer that brings uniformity and consistency to a disparate collection of data sources
- Uniformity and consistency *not just for user access* – also for:
  - *Data discovery*
  - *Data cataloging*
  - *Data protection*
  - *Data governance*
- A *smart* fabric – AI and machine learning technology are leveraged for intelligent automation of data management tasks
- Data sources can be a mix of on-prem and in-cloud (public and/or private)

Plenty of analyst's views on data fabric – here's one

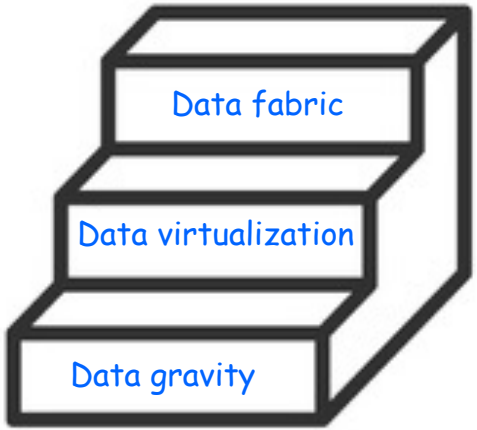
Data fabric is about…

Dynamically orchestrating disparate data sources intelligently and securely in a self-service manner and leveraging various data platforms to deliver integrated and trusted data to support various applications, analytics and use cases

Data-driven organizations use data virtualization and data fabric architectures to get value from data quickly and to support new business requirements such as real-time and integrated insights

Data gravity, data virtualization and data fabric

- Data gravity and data virtualization are important for a data fabric
  - In-place access to data on systems of origin (i.e., working with data gravity, not against it) is key to data fabric efficiency and performance
  - Data virtualization enhances the data fabric user experience – for business users, application developers, data scientists – by abstracting particulars of different data sources
    - Users are more productive because they can focus on the data itself and not on particularities of data organization and data-serving platforms



Making the concepts real for IBM Z and z/OS

Data gravity – Db2 Analytics Accelerator for z/OS

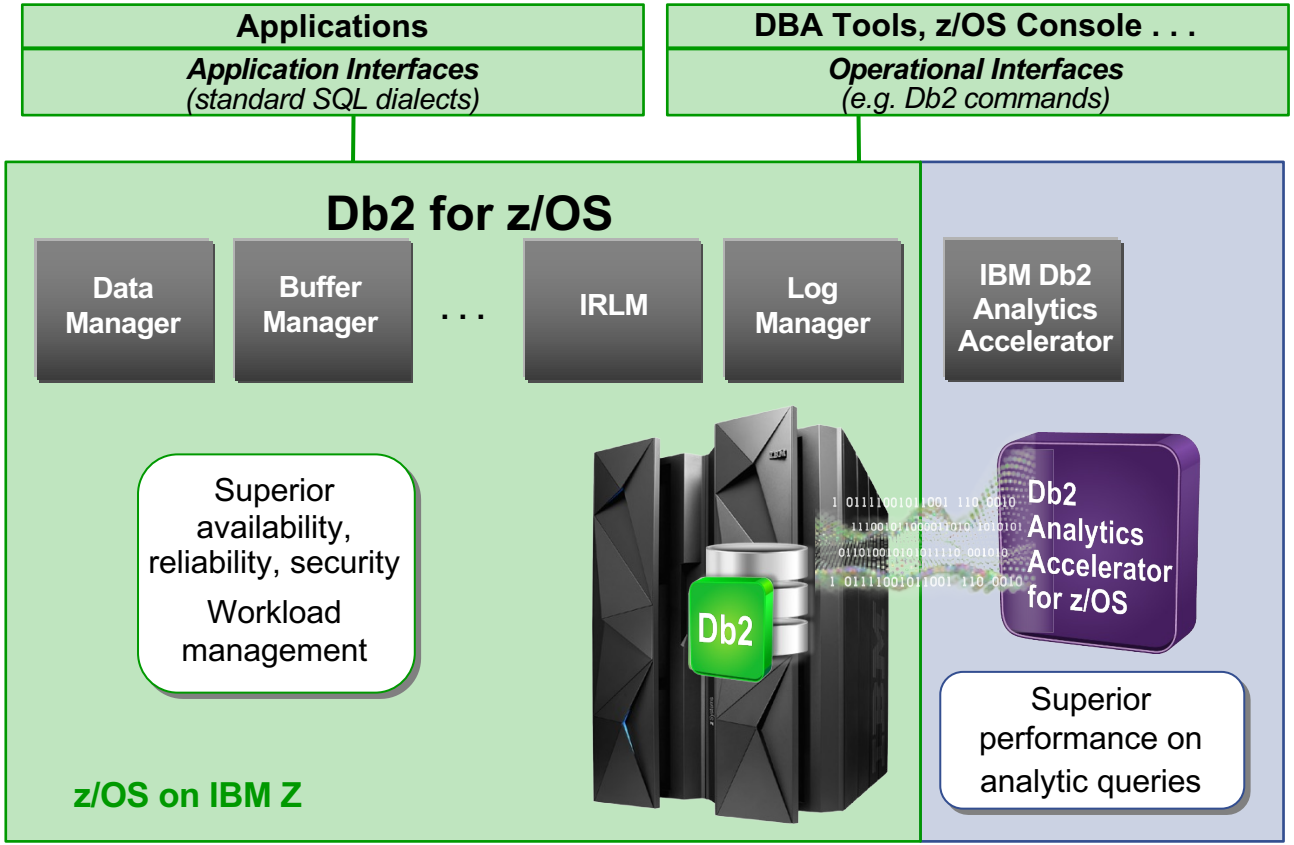
- Data gravity is about “bringing the processing to the data”
- If the processing in question is high-volume batch or transactional work, *no problem for z/OS*
  - z/OS systems, with Db2, IMS or even just VSAM files have long been processing huge “run the business workloads”
    - Thousands of transactions per second for a single LPAR, and “n” times that volume when running n-way Db2 or IMS data sharing on a Parallel Sysplex cluster
- But what if we’re talking about an analytics workload, characterized by complex, data-intensive queries? Can we bring that to z/OS-based data?
  - **YES** – that’s where the Accelerator comes in

Accelerator architecture

- The Accelerator has been around for 10+ years
- Formerly, implemented as an outboard analytics server connected to front-end Db2 for z/OS system
- The modern Accelerator is a virtual server, running in a containerized form on mainframe IFL engines – front-end Db2 for z/OS system can be in same mainframe
- Within Accelerator container: Db2 for Linux
  - Because of containerized form, no need for Linux admin
  - Accelerator's Db2 is Db2 Warehouse, with BLU Acceleration – an in-memory, column-oriented data organization
  - Complex, data-intensive queries can run 100s of times faster versus front-end Db2 for z/OS system

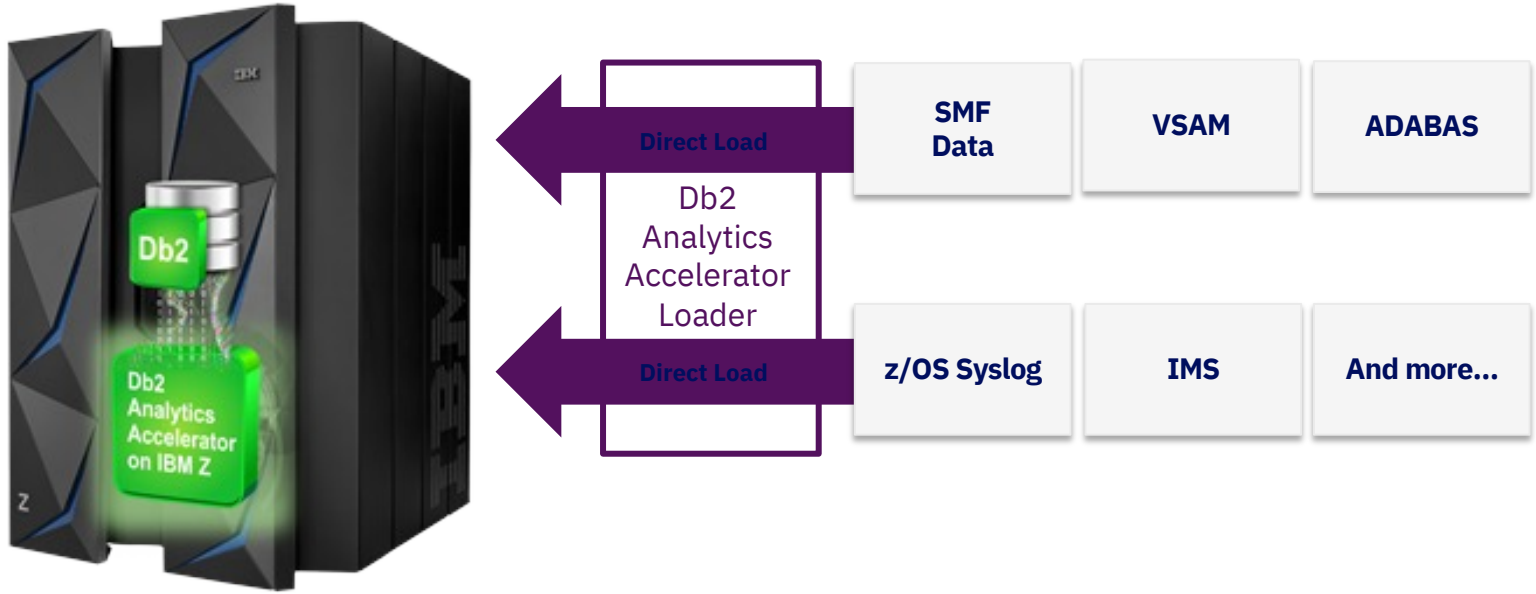
True hybrid transaction/analytical processing

- Logically invisible – queries are directed to the front-end Db2 for z/OS system
- Db2 for z/OS optimizer automatically determines if query would run faster on front-end or on Accelerator and routes query accordingly
- Query result returned to user or application as usual
- Integrated synchronization keeps copies of tables on Accelerator within 1-2 seconds of currency, with almost no use of general-purpose engines



What about z/OS-based data outside of Db2?

- That's where the Db2 Analytics Accelerator Loader comes in
- The Loader can load non-Db2 data simultaneously into a front-end Db2 for z/OS table and an Accelerator table, or into an Accelerator-only table



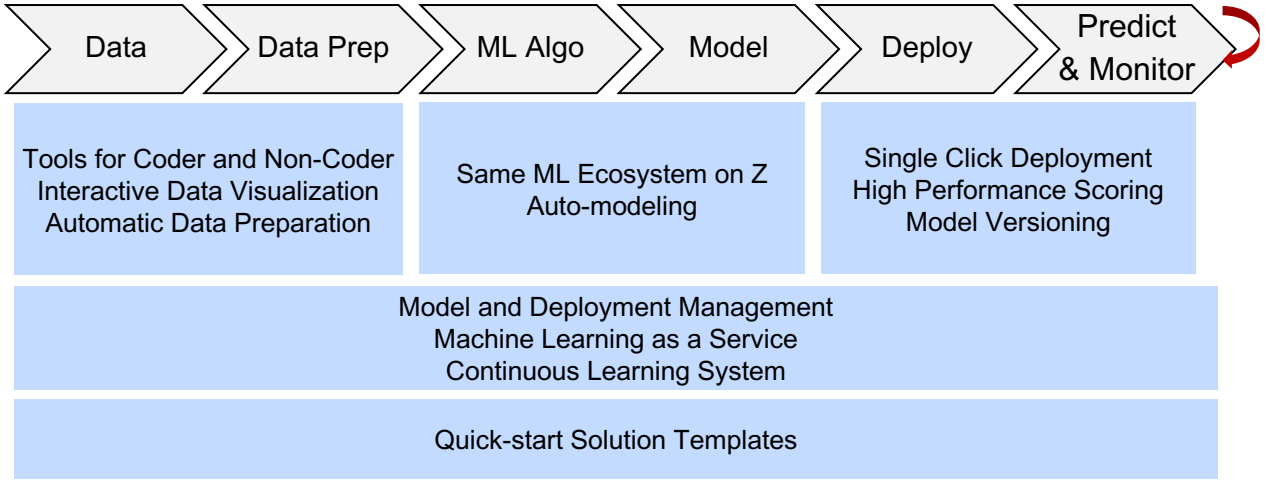
Data gravity – Watson Machine Learning for z/OS

- Predictive models developed using machine learning technology offer the opportunity for real-time insights that can be infused into operational applications, for purposes such as…
  - Fraud detection
  - Cross-selling and up-selling for online shoppers
  - Customer care
  - And much more…
- But, is real-time scoring feasible for high-volume, response-time-sensitive z/OS-based transactional and batch systems?
- **Yes** – that's where Watson Machine Learning for z/OS comes in

IBM Watson Machine Learning for z/OS (WMLz)

Provides an end-to-end machine learning platform for AI on z/OS

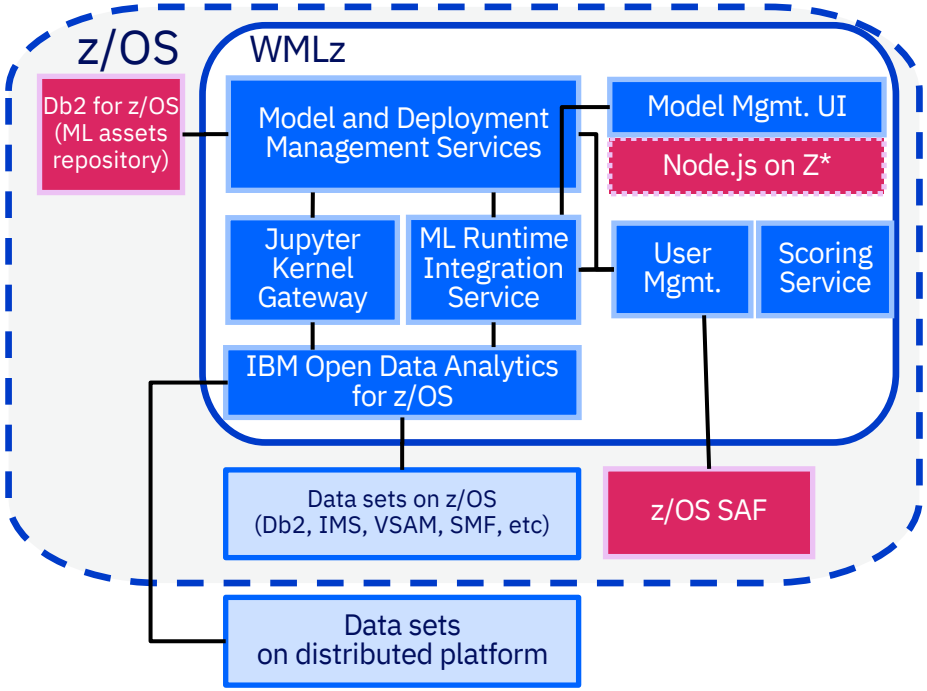
- Delivers predictive analytics capabilities on IBM Z to generate real-time insights at the source system
- Delivers essential model versioning, auditing and monitoring
- Supports high availability, high performance (score can be generated in < 1 ms), low latency and ML model automation
- Offers quick-start solution templates for common business requirements to kick-start machine learning projects



Works with data gravity, and integrates with existing application environment

- Leverage current data for improved insights
- In-place data access for efficiency and more-frequent model refresh
- Models can be developed by data scientists on Linux systems, imported and deployed in z/OS
- Optimized deployment at point-of-transaction

Watson Machine Learning for z/OS – architecture

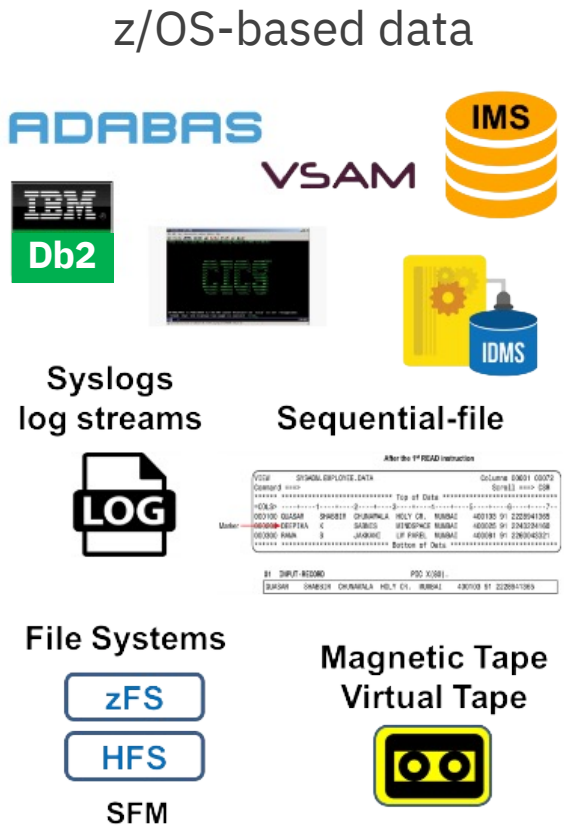


- IBM Open Data Analytics(IzODA) provides the base data access and Spark/Python libraries
- WMLz provides machine learning libraries and full life cycle model management and deployment capabilities (the services are REST-callable)
- Jupyter Kernel Gateway provides access to ML resources from Linux Services
- Db2 used as metadata and machine learning asset repository
- SAF provides authentication services
  - Support both RACF keyring keystore and file-based keystore

WMLz components and bundled components

WMLz prerequisites

Data Virtualization Manager for z/OS

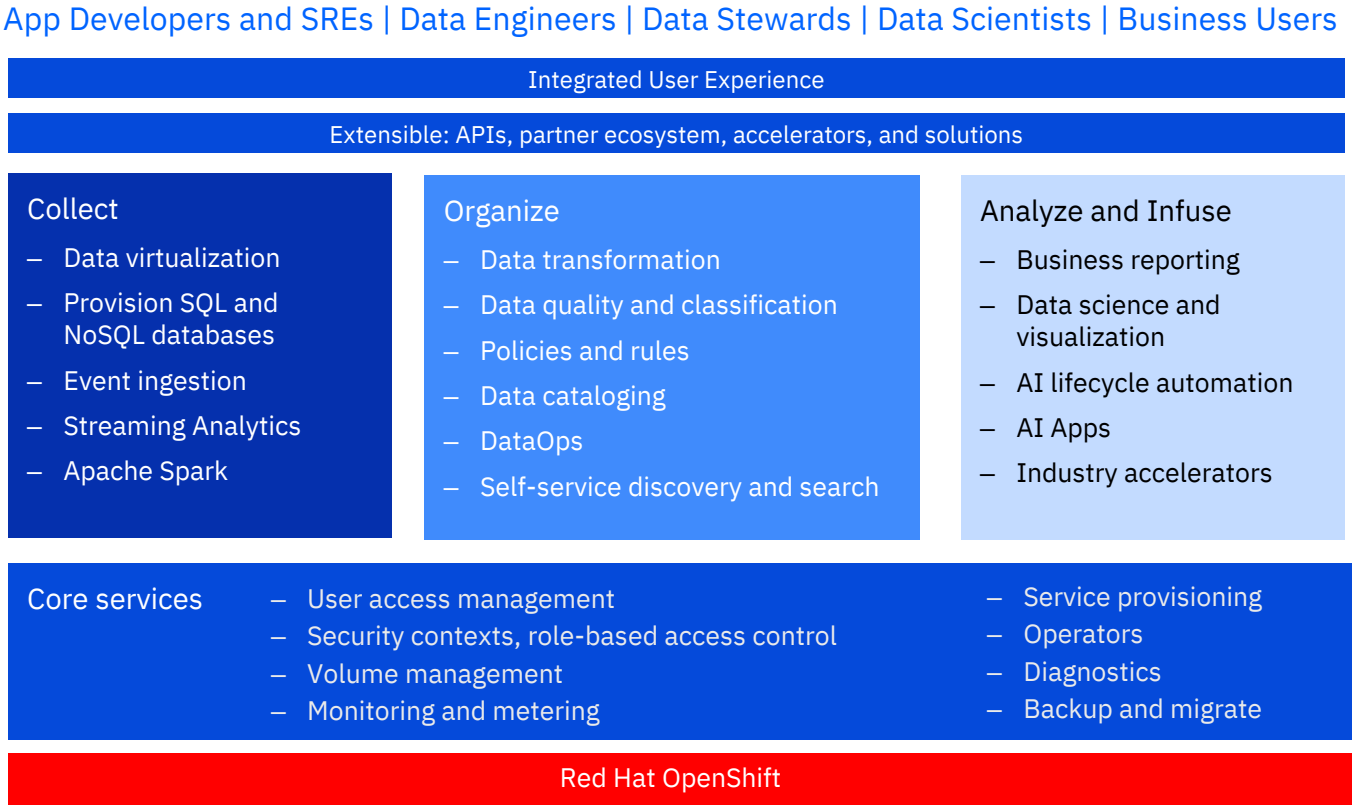


- Enables applications and users to access to non-relational z/OS-based data sources via modern interfaces such as JDBC, ODBC and REST (latter in conjunction with z/OS Connect)
- Runs in z/OS – substantially better performance versus IBM InfoSphere Classic Federation Server for z/OS (an older virtualization solution)
- Virtualization processing done by DVM is essentially 100% zIIP-eligible
- DVM also allows joins to be performed for different data sources (and that includes joins data in z/OS and on remote servers)

Cloud Pak for Data

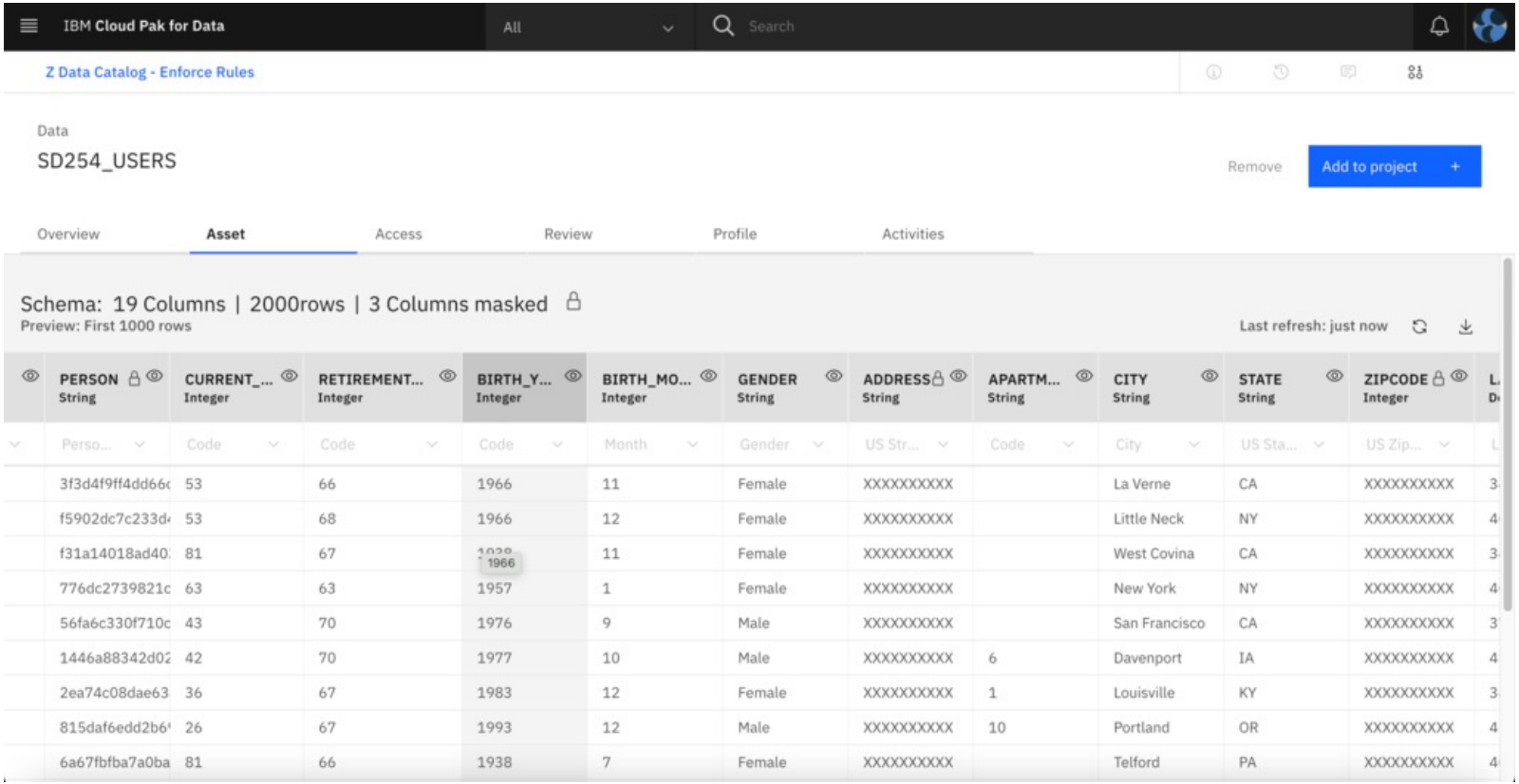
It's not called "Cloud Pak for Data for z/OS" because it is applicable to all of an enterprise's data platforms

- *IBM's premier data fabric-enabling technology*
- *Unified, modular, deployable anywhere – on-prem, private cloud, public cloud (IBM, Microsoft Azure or AWS Cloud)*
- *Also available in as-a-service form, fully managed on IBM Cloud*
- *Data virtualization works with DVM for access to non-relational z/OS-based data sources*



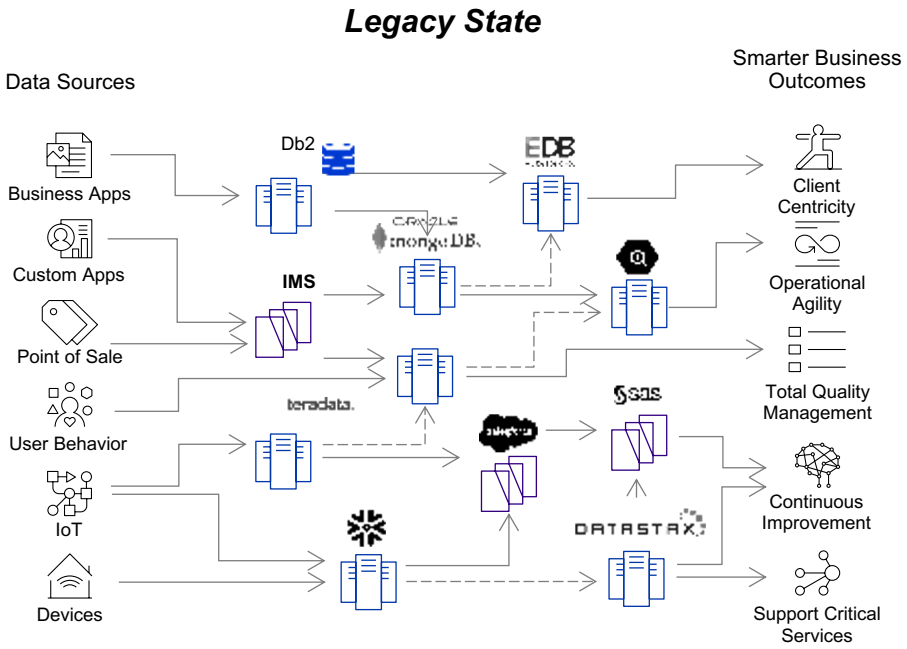
Cloud Pak for Data – intelligent data management

Example: automatic and dynamic masking of address information, based on security rule and user role

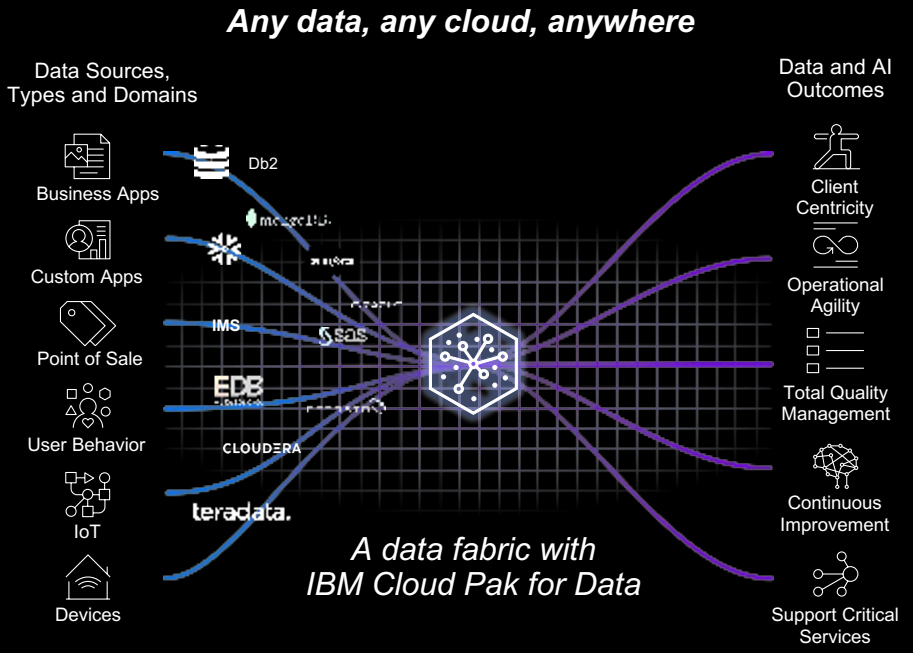


Not just IBM technology – an expanding CP4D ecosystem

A data fabric enabled by Cloud Pak for Data can turn this…



Into this…



To recap…

- Data gravity, data virtualization and data fabric are concepts, but they are important concepts – when implemented, they can deliver real value
- These concepts are not platform-specific – they apply across data-serving platforms
- IBM has technologies that can make these concepts real for z/OS systems – among those technologies are…
  - *Data gravity enablement:* Db2 Analytics Accelerator for z/OS, Watson Machine Learning for z/OS
  - *Data virtualization enablement:* Data Virtualization Manager for z/OS
  - *Data fabric enablement:* Cloud Pak for Data

Robert Catterall

rfcatter@us.ibm.com