Leverage and Monetize Mainframe data through Hybrid Cloud

Yichong Yu, STSM, IBM Cloud for Financial Services
Surya V Duggirala, IBM Distinguished Engineer and CTO, Enterprise Cloud Architecture

Many of the world's leading businesses run on the mainframe, including worldwide banks, largest insurers, airlines, and many of the US's largest retailers. Mainframes process around 29 billion annual ATM transactions and 12.6 billion transactions per day. It is estimated that about 80 percent of the world's corporate data resides or originates on mainframes.

According to IDC's research, the mainframe has evolved from a siloed system into a connected and now a transformative platform, and with that the mainframe platform is transforming from a revenue-supporting machine into a revenue-generating machine and is increasingly playing a central role in organization's digital transformation journey.

To retain the core strengths and attributes of IBM Z platform (IBM Z and IBM LinuxONE) and to leverage extensive cloud services, security and regulatory compliance programs of IBM Cloud, IBM recommends hybrid cloud approach to mainframe application modernization.

IBM's point of view on mainframe application modernization with hybrid cloud starts with seven entry points including how to simplify data access. Data security and compliance certification is a major focus area for clients before mainframe data is made available for access for multiple use cases to derive more value from it. IBM Cloud is designed for enterprise and regulatory workloads with differentiated cloud security, confidential computing services, and native IBM Z dev/test support. Together, IBM Cloud and IBM Z can help accelerate mainframe application modernization by leveraging and monetizing the mainframe data in a secure and scalable way.

Data integration strategy

Leveraging and exploiting the potential of huge transactional data residing on IBM Z by applications deployed on IBM Cloud opens lots of opportunities and helps in application modernization. Hybrid cloud data integration in support of analytics to make better decisions and to unleash the power of data to support existing applications and new business initiatives is a growing need as enterprises adopt cloud technologies in their digital transformation journey.

Many factors go into choosing the data integration strategy to satisfy business needs. Companies should start with the business problems at hand, understand the requirements of the use cases, and evaluate different options. Here are some of the factors that need to be considered and evaluated before choosing the data integration strategy.

- **What?**
  What data is needed to solve the business problems? What are the data sources? Is the source data from databases, sequential data sets, or/and somewhere else? Is it structured or unstructured? What data should be made available for consumption? Is it for a particular purpose or for potential many consumptions by various applications?
- **Where?**
  Where are the data sources and data destinations? Are the source data all on-prem, or does it need integration with other data in cloud or from some external data sources? Where will the data be consumed, is it feeding some applications or AI models on-prem, cloud applications, or as a data product consumed by others?
- **When?**
  When does the data need to be made available, is it in real-time or in batch with some delay?
- **How?**
  How will the data be made available? Can the data be accessed in place via virtualization, or data copy needs to be maintained in cache or off-premises? What is the granularity and volume, is it transaction level or large volume data dump?
- **Why?**
  Why are there special requirements on data integration, for example, is there PII data, is there regulations to meet, can the data leave on-prem or certain region? Any other requirements on latency, performance, security, compliance, auditing, etc.?

Enterprises can evaluate different options and adopt the right data integration strategy, or a combination of strategies based on the business requirements. For example, an enterprise can leverage real-time information sharing to make new data available to on-prem applications as well as cloud native applications via APIs or Events. For other use cases, existing production datasets can be exposed for SQL queries in response/request approach via data virtualization, if the enterprise authorizes such access. They can also consume and use batch data transfer to make a big volume of historic data available for data scientists to train AI models in cloud.

Data security in hybrid cloud

To unleash the power of data to better support existing applications and to explore new business initiatives, access to transactional data on mainframe is crucial. Transactional data is highly sensitive and needs to be protected. Security incidents are extremely expensive. According to the data breach report 2022, average cost of a data breach for Financial industry is $5.97M. Data needs to be secured and governed, and yet can be easily discovered and used in hybrid cloud environment.

In hybrid cloud settings, when the data on-prem is made available for consumption in cloud, whether querying the data in place with data virtualization, synchronizing the data in real-time to feed cloud applications, or communicating large volume of data over for AI model building purpose, data needs to be protected in motion (moving over a network connection), in use (during processing or runtime), and at rest (in storage and databases). IBM confidential computing can be used to protect data in use. Encryption methods such as HTTPS, SSL, and TLS are often used to protect data in motion. IBM Cloud Direct Link and VPN can also be used between private infrastructure and IBM Cloud. Encrypting data protects the information at rest, even if that information is lost or stolen.

IBM Cloud Hyper Protect Crypto Services is an industry-leading key management and hardware security module service built with mainframe-level encryption and offered as a service in the cloud. It is built on FIPS 140-2 Level 4 certified hardware, the highest level in the industry, and can be integrated with other IBM Cloud services to protect digital assets in Cloud. Hyper Protect Crypto Service supports Keep Your Own Key (KYOK) which ensures your full control of the entire key hierarchy where no IBM Cloud administrator has access to your keys. Unified Key Orchestrator (UKO), a part of Hyper Protect Crypto Service, enables key orchestration across hybrid multicloud environments from a single point of control. Sensitive information at rest is secure. For more details on how to secure applications and data in IBM Cloud can be found in this paper.

IBM Cloud for Financial Services is designed to build trust and enable a transparent public cloud ecosystem with security, compliance, and resiliency features that financial institutions require. The reference architectures provide separation of concerns in network settings. Cloud services or ecosystem partner services can evidence compliance to the controls and become IBM Cloud for Financial Services Validated. The Financial Services Validated designation signifies that you have successfully evidenced compliance to the control requirements of the IBM Cloud Framework for Financial Services. IBM Cloud Security and Compliance Center enables continuous compliance and protect customer and application data. Super sensitive workloads can be run in Hyper Protect Virtual Servers.

With IBM Cloud for Financial Services framework to protect and monitor sensitive workloads and IBM Cloud Hyper Protect Crypto Services to support KYOK, data in motion, in use, and at rest are fully protected in IBM Cloud.

Data integration patterns

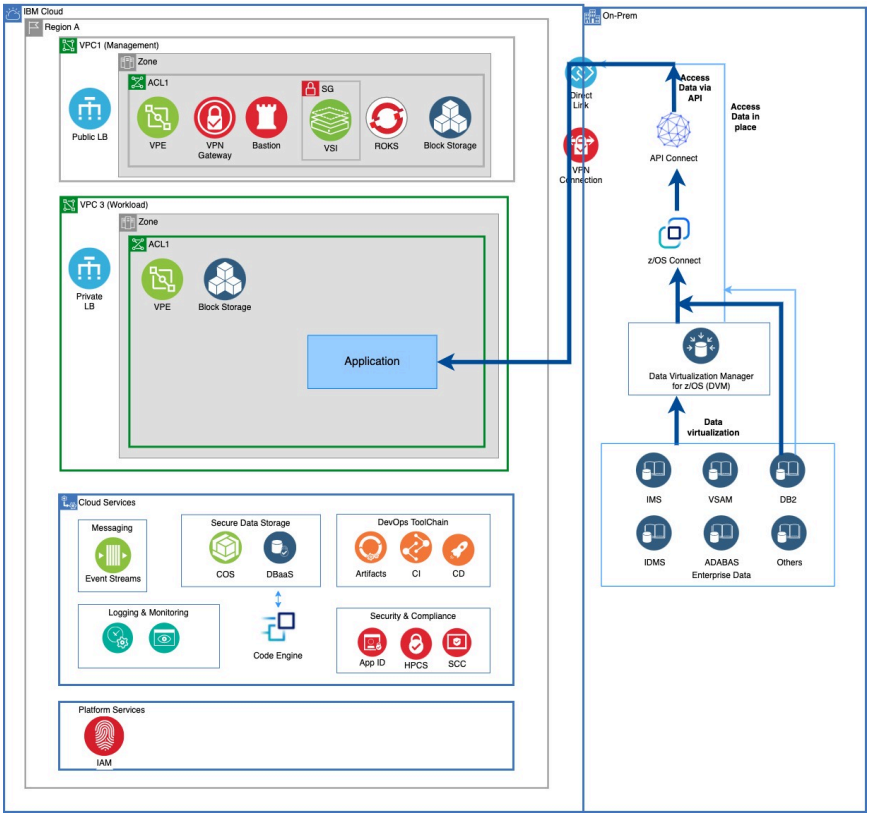
We will discuss different data integration patterns in this section.

Data exposure pattern: Unleash and expose via APIs

Enterprises are collecting data from diverse data sources. However, research shows that up to 68% of data is not analyzed in most organizations and up to 82% of enterprises are inhibited by data silos. Breaking down silos by copying disparate data for analysis into central data stores could be costly and error prone. With data virtualization, one can access data at the source without moving data.

Data virtualization is a data integration method that brings together disparate data sources to create virtualized, integrated views of the data. Data virtualization also provides a data services layer that abstracts the underlying physical implementation of the individual data sources. Data virtualization helps meet compliance requirements since there are less copies of data, and data is fresh and accurate from source.

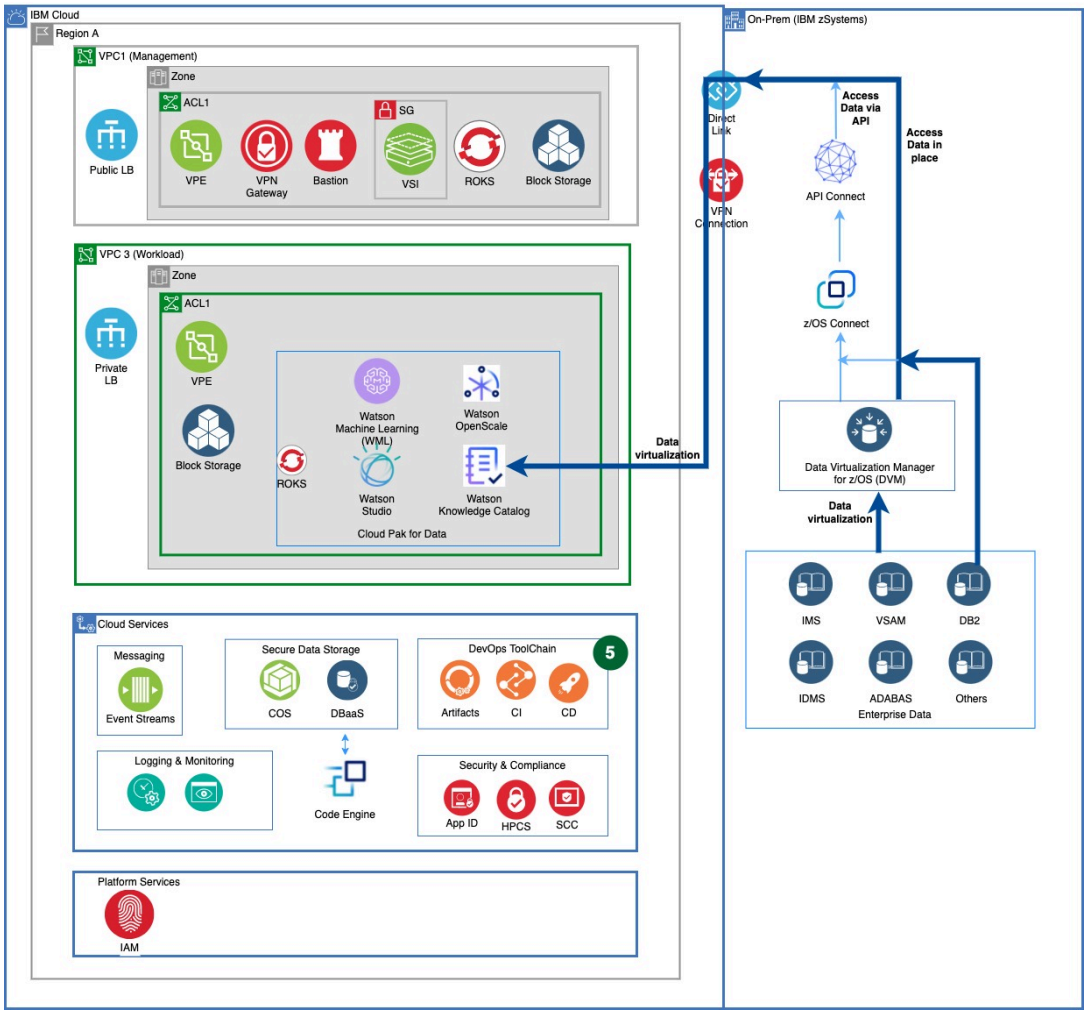
IBM Data Virtualization Manager for z/OS (DVM) provides virtual, integrated views of data residing on IBM Z. It is extremely helpful for sequential data and allows SQL queries to be run against sequential data. IBM z/OS Connect provides a simple and intuitive way to bring the power of APIs to IBM Z platform. Together, IBM DVM and z/OS Connect provide an efficient, scalable, and secure method for leveraging mainframe application and data and make them central to your hybrid cloud strategy.



Data virtualization pattern: Unify and virtualize via data fabric

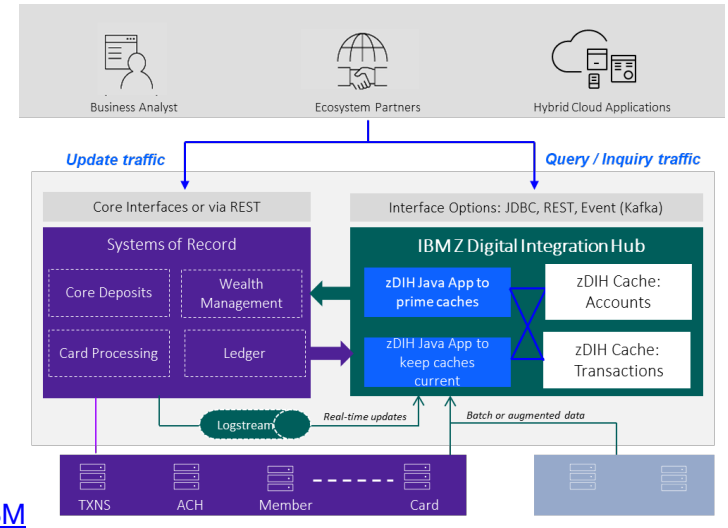
One can have another level of data virtualization in cloud by combining data from mainframe and other sources in cloud. This can break the data silos and provide a wholistic view of enterprise data.

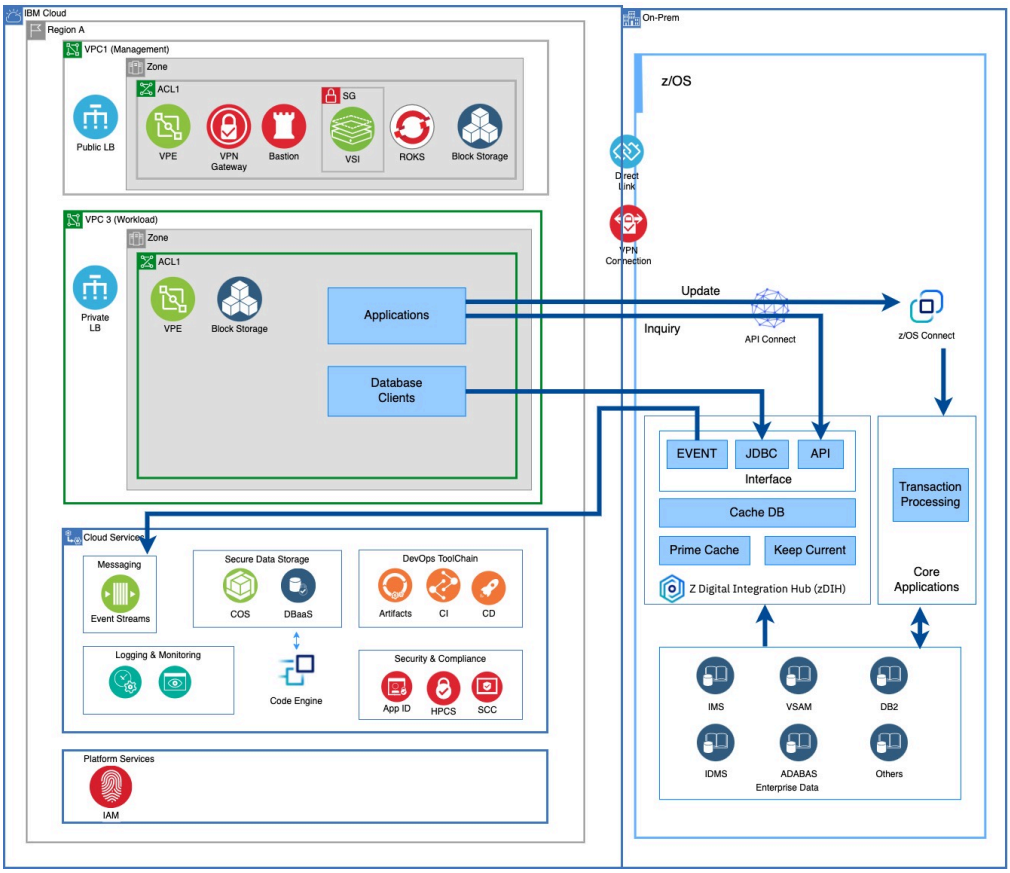
IBM Cloud Pak for Data (CP4D) simplifies and automates the process to collect, organize, analyze, and infuse data. IBM CP4D can be deployed on RedHat OpenShift cluster and can be deployed on multiple-cloud providers and on-premises. Data Fabric solutions simplify data complexity through automating data integration, data governance, and data consumption. Organizations can access and govern all data, across any cloud, to deliver trusted, business-ready data anywhere. Data Fabric weaves together producers and consumers of data allowing clients to access data wherever it is, manage its full life cycle, and deliver the data through self-service. The Data Fabric entry points are built on IBM Cloud Pak for Data and Red Hat Open Shift allowing clients to take the first step to build a data fabric as well as access key business data with AI. This provides information to make intelligent, strategic, business decisions and reap the benefits of the efficiencies.



Real-time information sharing pattern: optimize for consumability and efficiency

IBM Z Digital Integration Hub (zDIH) enables real-time information flow at scale between systems of record and hybrid cloud applications, ecosystem partners, ISV solutions, and end users.
With zDIH, you can efficiently integrate core systems of record that typically run on z/OS® and keep a set of in-memory caches on z/OS® current in real-time. These caches have a flexible set of standards-based interfaces to easily share the cached information with downstream applications and users through APIs, JDBC or Kafka based events. IBM zDIH optimizes a CQRS (command query responsibility separation) pattern by providing a mechanism to separate query traffic from update transactions and reduces impact on production applications. Since zDIH is highly zIIP-eligible, you can also benefit from cost optimization by separating query-only processing from core transactions. For more information on IBM zDIH, refer to: IBM Z Digital Integration Hub



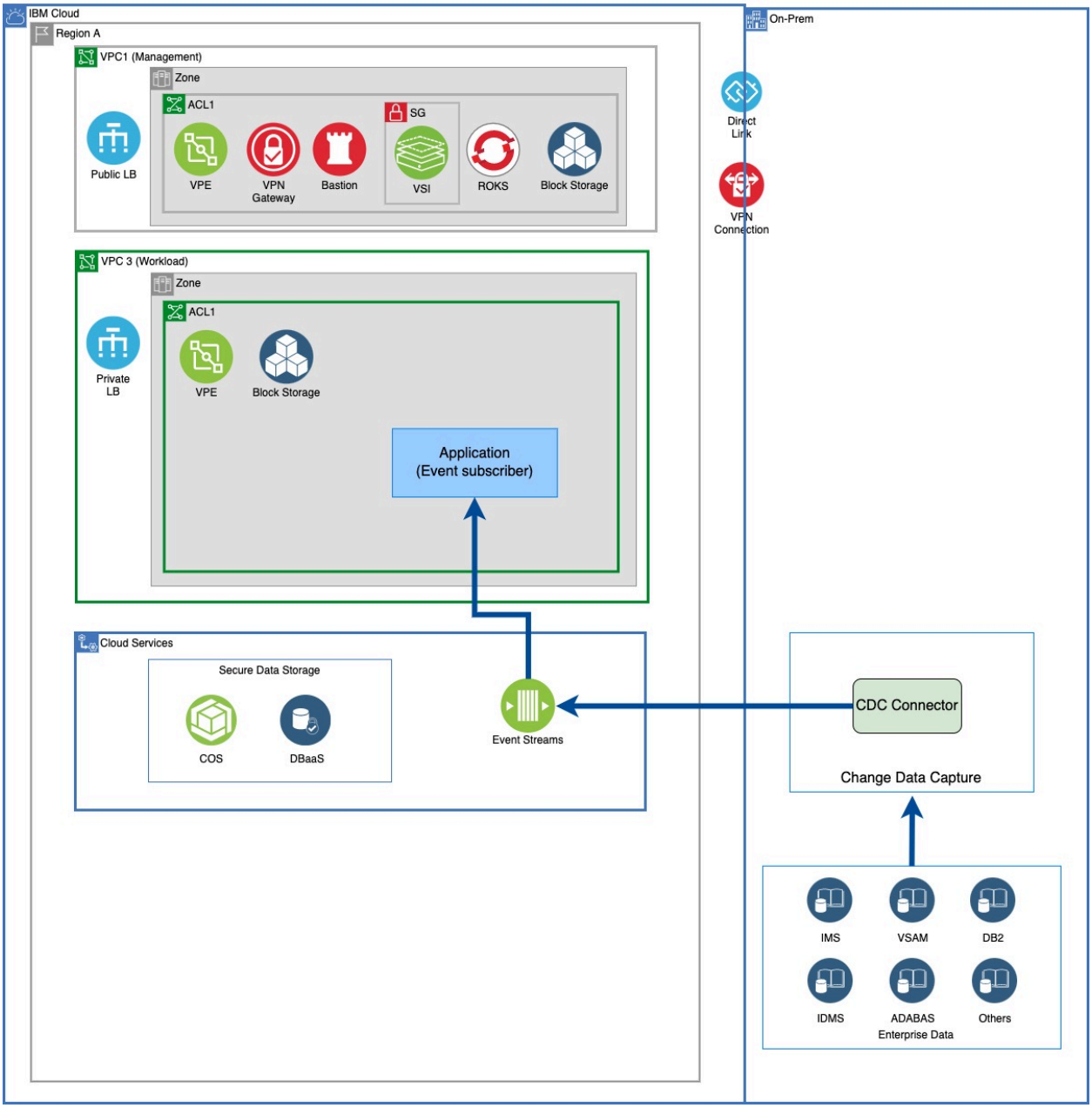


Data synchronization pattern: Integrate and feed via Event Streams

For applications and analytics that need real-time or near real-time data feed, data changes on-prem can be synchronized to cloud in real-time or near real-time via IBM Event Streams.

Data changes can be captured and replicated in Event Streams for downstream consumption. Applications on-prem or in cloud can analyze data as part of application logic that pertains to an event and respond to that event in near real time in a flexible way with newly developed response logic without introducing risks in core applications.

It is also possible to reverse the direction of the data flow and allow mainframe applications to respond to cloud events in near real time by leveraging z/OS assets.



Rather than managing the data synchronization, client can also choose IBM product to manage data synchronization automatically, for example, IBM Db2 for z/OS Data Gate (Db2 Data Gate) can be set up to manage data synchronization. IBM DB2 Data Gate facilitates the use of data originating in Db2 for z/OS within all hybrid cloud scenarios, enabling a hybrid cloud consumption model for Z data in an efficient, integrated, and easy to use way. Refer to DB2 for z/OS Data Gate for more information.

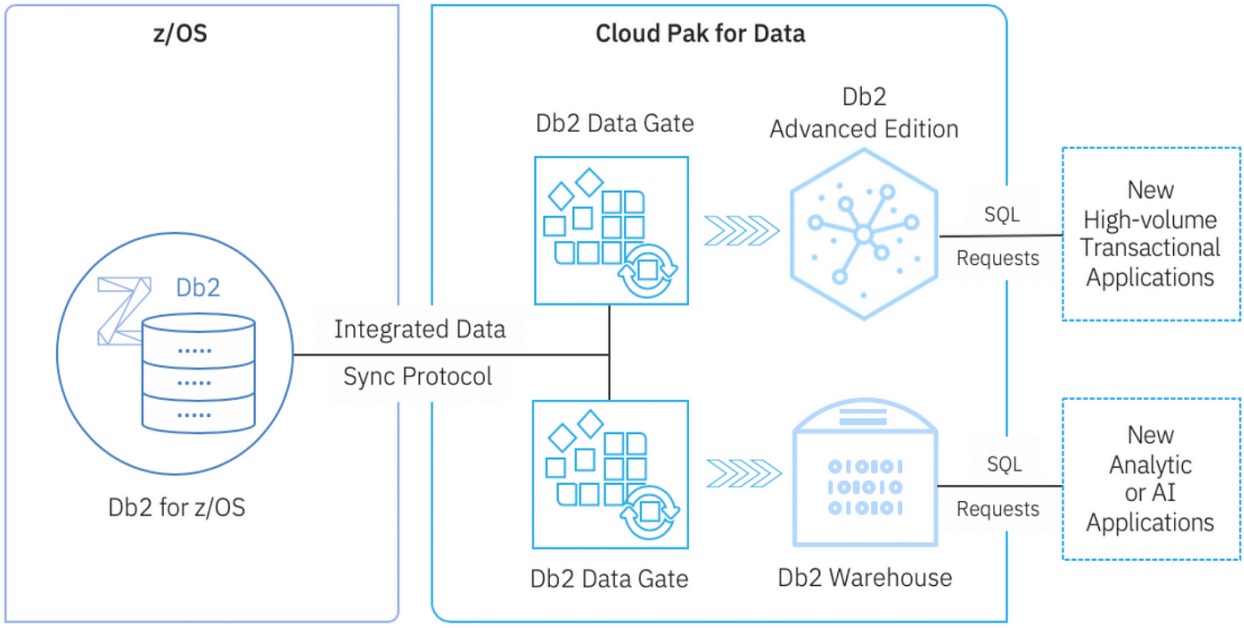


Chart source: DB2 Data Gate Blog

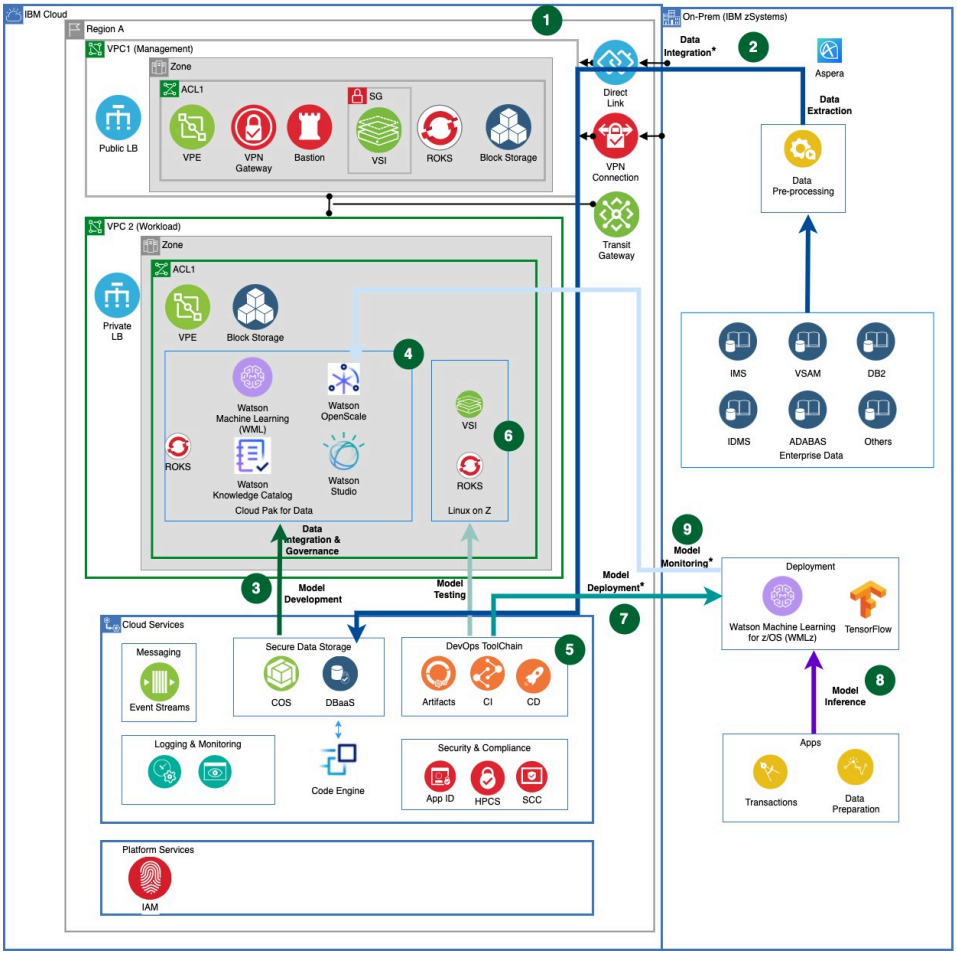
Data transfer pattern: Analyze and decide with batch

The data that is available on mainframe can be useful to derive better insights and help generate new streams of revenue. Certain use cases require large volumes of raw data to be made available, for example, training an AI model or archiving data to Cloud object storage. In this case, data can be preprocessed if needed (e.g., cleansing, filtering, anonymization, aggregation, encryption, etc.), before being transferred to Cloud.

Batch data transfer jobs can be run periodically, and large volumes of data needs to be transferred across the network efficiently. IBM Aspera offers unrivaled performance for transferring large files and large collections of files across any distance to, from or between clouds. In the cases of low or no bandwidth or for remote locations that have unstable network, big volume of data can be encrypted and copied to IBM Cloud Mass Data Migration device, and the data can be shipped and copied to cloud object storage offline.

Here is the chart for AI pipeline in hybrid cloud (refer to the AI blog for details). Step 2 captures the online batch data transfer process. IBM Cloud Direct Link and site-to-site VPN allows data flow

from on-prem to management VPC. IBM Cloud Object Storage has private endpoint enabled, which allows connection from management VPC. IBM Cloud Object Storage CLI commands or APIs can be used to write data files to IBM Cloud Object Storage buckets. IBM Cloud Code Engine jobs can subscribe to IBM Cloud Object Storage write events and process data automatically if needed. Analytics jobs or applications can consume the data in IBM Cloud Object Storage.



1. Set up the infrastructure for IBM Cloud for Financial Services.

2. Prepare data to be used for model training and store the data in Cloud Object Storage (COS) in Cloud*.

3. Understand, prepare, and manage the data and governance with tools from IBM Cloud Pak for Data, eg. Watson Knowledge Catalog.

4. Train and test AI models with tools from IBM Cloud Pak for Data, eg. Watson Studio and Watson Machine Learning.

5. IBM DevOps Toolchains can be used to manage the model source codes and tested models.

6. AI models can be tested on Linux-on-Z in IBM Cloud.

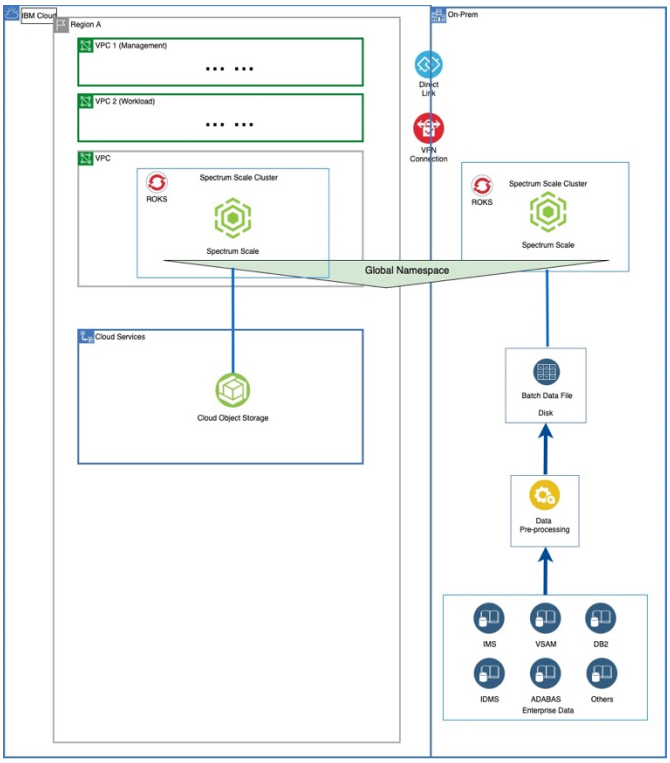
7. Tested AI models can be deployed on IBM Watson Machine Learning for z/OS*.

8. AI model inference is invoked to gain insights from the data.

9. AI models are continuously monitored via IBM Watson OpenScale for bias and drift*.

** All traffics between on-prem and IBM Cloud are through Direct Link and/or VPN.*

Rather than working with IBM Cloud Object Storage, another alternative is to use product like IBM Spectrum Scale, which is a high-performance and highly available clustered file system available on a variety of platforms (including the public cloud service providers). IBM Active File Management (AFM) is a scalable, high-performance, intelligent file system caching layer integrated into the IBM Spectrum Scale file system. This enables you to implement a single global name space across various sites including the public cloud offerings. With IBM Spectrum Scale available as software-defined storage, organizations can deploy a hybrid cloud environment. This creates an environment where data files can be moved and cached between on-premises to Public Cloud offerings. Here is a reference architecture that only focuses on the data movement.



Conclusion

As IBM Z platform is used for conducting core business transactions across the industries, unleash the power of transactional data and use it to derive better insights and support new business opportunities that can bring enterprises competitive advantages.

This article discusses different data integration strategies and data integration patterns to make the data accessible in the hybrid cloud to leverage and monetize valuable mainframe data. Protecting mainframe data with strong data security is critical to enterprises. Data and information need to be protected in motion, in use, and at rest. IBM Cloud for Financial Services can help enterprises setup a secure cloud environment and run sensitive workloads that leverage high value mainframe data.