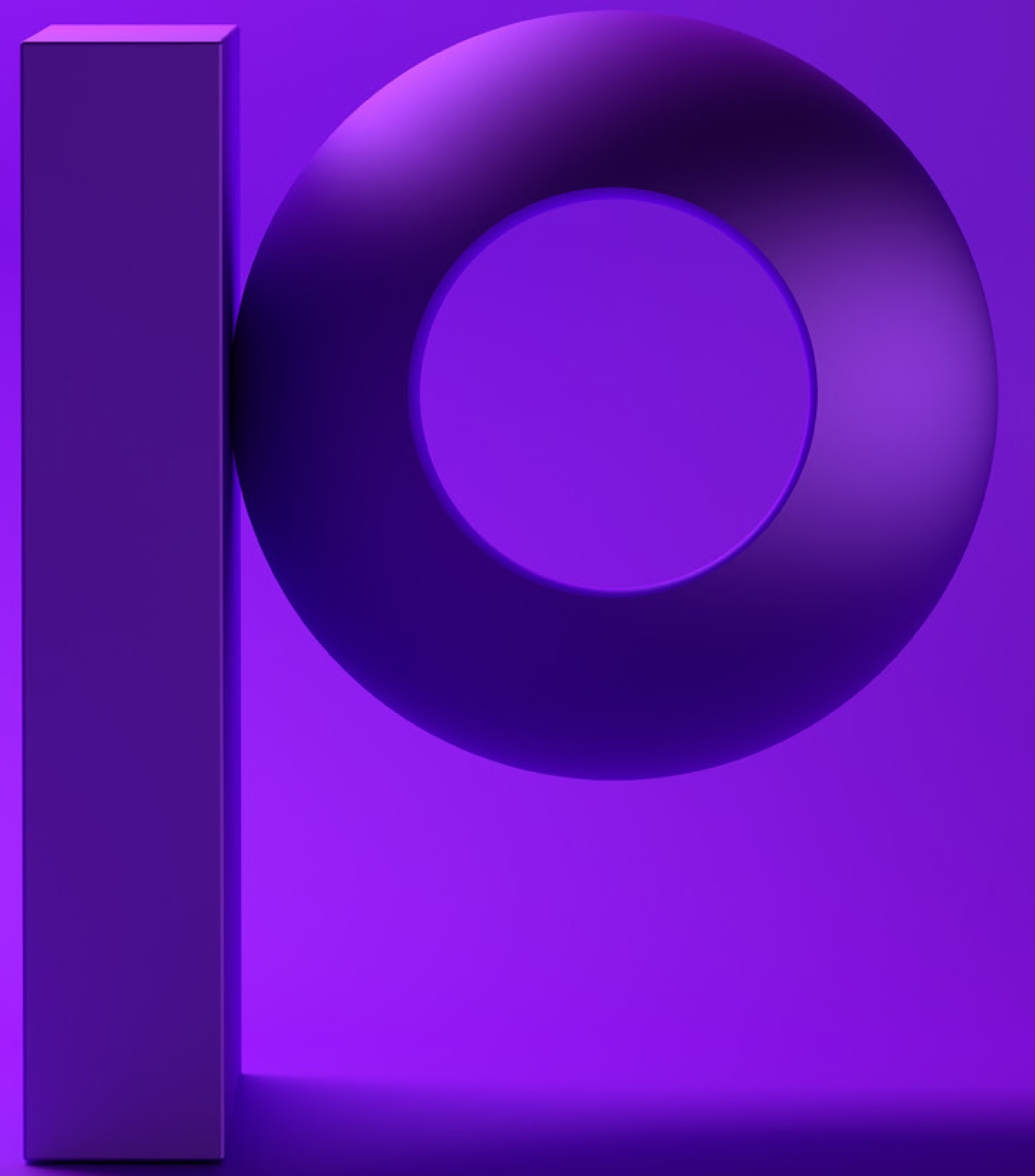
precisely

Streaming Legacy Data to Kafka

Real Industry Stories



Introduction

IT executives are constantly challenged by the need to accelerate the adoption of the most modern and powerful technologies while still having to rely upon their existing installed base of storage and compute infrastructure. The only practical path forward is to enable backwards compatibility. The new must connect and interoperate with the old. Leading-edge must leverage legacy.

In truth, given the rate of acceleration in technological advances, even the definition of "legacy" is becoming blurred. For example, would an organization transitioning from data warehouses to an enterprise data hub really consider their data warehouse infrastructure to be "legacy?" Cloud-based, distributed processing frameworks and database platforms are enabling deployment of much faster and more powerful real-time processing and analytics systems. Yet the vast majority of the data they process is still being generated by a pre-existing installed base of older servers and applications.

For many of the largest organizations, that installed base includes major installations of IBM/z mainframe servers and storage, which handle their core, high-volume processing. The costs and disruption that would result from any effort to replace these systems outright are simply incalculable. But the challenges of integrating them with modern, cloud-based distributed processing platforms are daunting.

On the following pages, you will learn how three organizations used the Precisely data integration solution Connect to fully integrate their legacy systems into their cloud platforms and analytics engines by streaming to Kafka, gaining real-time access to legacy data while eliminating the costs and delays of manual ETL processes.

Insurance

In order to support its long-established position as a market leader in their industry, this insurance and asset management company is dedicated to continuous improvement and modernization of their IT systems. The company's IT strategy centers on fully integrating its entire information base within an enterprise data hub, with the end goal of delivering rapid access to current and complete information to all areas of the business, from front-line customer services to back-office asset management, analytics and regulatory compliance.

Given the enormous scale and platform diversity of its infrastructure, one of the biggest challenges was enabling real-time ingest of transaction data to Cloudera for use as its data hub. This was especially true for the vast amount of data generated by core applications running on IBM/z mainframe systems.

Cloudera offers no native connectivity or processing capabilities for EBCDIC formatted mainframe files. Moving that data into the hub necessitated labor-intensive ETL processes that were managed by teams of exceptionally high skilled IT experts who were versed in both IBM/z mainframe and distributed cloud technologies.

In addition, using ETL laid significant additional processing loads onto the mainframe machines and also delayed delivery of data to the hub, resulting in business systems that were always using back-level data. Beyond these problems, as an insurance company, regulations require the company to preserve transactional data in its original mainframe format, creating a complicated, unsustainable "dual-books" situation for audits and reporting.

Requirements

- The ability to move IBM/z mainframe data into Cloudera for use in an enterprise datahub in real-time
- Ability for all business applications to access IBM/z data
- Retention of IBM/z data in its original format for regulatory audit and reporting

Solution

- Connect CDC for capturing and loading IBM/z data into Kafka in real-time
- Connect ETL for batching IBM/z data in Kafka for delivery to Cloudera

Key Results

- Core IBM/z data is captured, prepared and delivered in real-time to Kafka, eliminating the delays and expense caused by daily ETL processing
- IBM/z data in Kafka is then batched for delivery into Cloudera, where it is stored in its original EBCDIC format
- Real time availability of IBM/z data for use within open systems applications for analytics, customer support, etc.

International Shipping

It may seem that shipping goods via ocean-going container ships is a slow and straightforward process. But that view could not be further from the truth. The goods may move slowly while on the water, but the volume of data regarding them is massive and changes constantly, moving between customs agencies and ports and local intermodal carriers with the same urgency as any Next Day Air shipment.

A global maritime services firm needed to modernize its systems to meet increasing demand for real-time shipment tracking and information sharing with its multitude of partners, ports, and local carriers worldwide. The strategic solution was to integrate the data being generated across many different legacy platforms into a cloud-based data hub for direct access by systems managing daily operations, such as shipment tracking, as well as by analytics and regulatory compliance applications.

To achieve this, the company needed to be able to access critical container tracking and customs documentation, which is hosted and processed on IBM/z mainframe systems as well as on its Oracle SQL server platform. Within its data hub architecture, Kafka was being leveraged to gather and distribute data to and from multiple systems and applications in real-time. But the company had no way to get the IBM/z mainframe and SQL server data into Kafka. And that data also needed to be prepared for delivery to HDFS/Hive running on distributed clusters to support the analytics and compliance systems.

Requirements

- Ability to provide customers and partners with real-time updates on container location, transit time and delivery details
- Ability to move IBM/z and SQL server data into Kafka for access by multiple systems in real time
- Enable the data held in Kafka to also be delivered in batch to HDFS/Hive database clusters when required.

Solution

- Connect CDC for real-time changed-data capture and delivery to Kafka
- Connect ETL for batching data in Kafka and delivering it to distributed HDFS/Hive database clusters

Key Results

- Integrated core logistics data on IBM/z available for use by multiple lines of business across the enterprise
- Able to provide customers and partners with detailed, real-time tracking of all freight
- Critical ship, container, customs and port operations information now available to multiple logistics and reporting systems in real-time via Kafka
- IBM/z mainframe and Oracle SQL data within Kafka automatically prepared for delivery in batch mode to HDFS/Hive databases for use in analytics and regulatory reporting

Investment Management Services

A leading global investment management firm needed to fully integrate all of its data into an enterprise data platform (EDP) to modernize its business analytics and regulatory compliance applications. While the challenges of migrating large volumes of existing legacy data into its EDP were expected, capturing and integrating daily transaction data from its Oracle systems turned out to be more difficult than expected.

Initially, the IT team leveraged Oracle's Flashback capabilities, running daily SQL-based Flashback queries to collect and package new and changed data for later ingest into the EDP.

However, this proved to be a very inefficient and labor-intensive solution. Not only were the database administrators spending a significant part of their day writing and modifying SQL queries to support business case-specific requests, running all those Flashback queries each day added significant processing loads to their systems, resulting in a very noticeable degradation of Oracle application response times.

Beyond those impacts, because the data collection and delivery processes were so laborious and time-consuming, the company's analytics and governance applications were always running against data that was not fully current or complete. This was because the overall data set delivered by the Flashback queries had to be coherent within itself, even though it was being gathered by multiple queries being run sequentially across many hours each day. So, the administrators had to choose an arbitrary cut-off time and restrict all the queries in that cycle to pulling up data older than their specified point in time. The result was an unavoidable and highly variable lag between the live Oracle system data and the data delivered for analysis and reporting.

Requirements

- Eliminate costly and slow Oracle Flashback-based ETL processes
- Real-time delivery of Oracle data directly to Kafka
- No impacts on or modifications to Oracle applications

Solution

- Connect CDC implemented on commodity Windows servers

Key Results

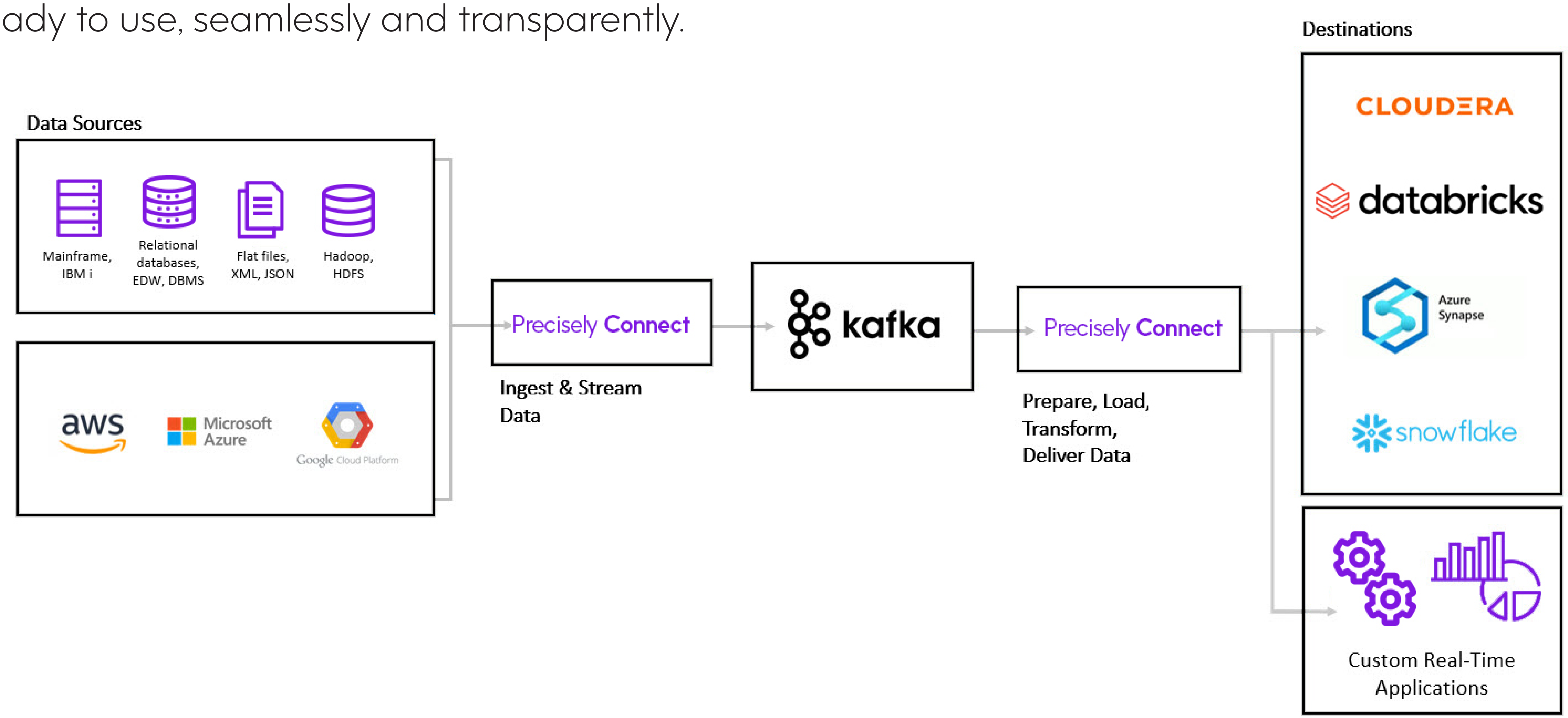
- Real-time capture and replication of Oracle transaction data to Kafka
- Elimination of Oracle system processing loads and application lag times caused by manual daily Flashback query runs
- Analytics and business governance applications using complete and up-to-the-minute data.
- Oracle DBMS Admins freed from routine, manual Flashback tasks and available to apply their higher-level skills to more important projects

Time to Retire the Word "Legacy"

The final result for each of the organizations we reviewed can be boiled down to this: The word "legacy" has become irrelevant to the discussion of how to build modern enterprise data platforms. Your data and systems are what they are. The best approach is to let your data be what it is, live where it performs best, and just virtualize your views into it. Ultimately, what you really want is a unified, source-agnostic, real-time data stream that is easily accessed by any application or analytics engine you choose to implement, now or in the future.

Connect makes it possible to do just that by enabling you to access all your data, capturing and delivering it in real-time to the people and systems that need it – ready to use, seamlessly and transparently.

With Connect's CDC and ETL capabilities, you can capture and transform data from virtually any source, including IBM/z mainframe systems, IBM i, Microsoft Azure, Snowflake, Oracle, Teradata, and many more. Then automatically deliver it directly to your key applications and databases or to Kafka for streaming and/or batch delivery to Hive, Spark, Cloudera, Databricks or any other analytics engine.



Data Sources

Mainframe, IBM i

Relational databases, EDW, DBMS

Flat files, XML, JSON

Hadoop, HDFS

aws

Microsoft Azure

Google Cloud Platform

Precisely Connect

Ingest & Stream Data

kafka

Precisely Connect

Prepare, Load, Transform, Deliver Data

Destinations

CLOUDERA

databricks

Azure Synapse

snowflake

Custom Real-Time Applications

Connect allows you to design your data transformation jobs once, focusing solely on business rules, not on the underlying platform or execution framework. Then simply deploy them anywhere – Spark, Hadoop, Linux, Unix, Windows – on-premises or in the cloud. No coding or tuning required.

Precisely combines cutting-edge technology with decades of experience on both mainframe and big data platforms to offer the best solution for accessing and integrating mainframe data. Your non-mainframe developers can work directly with virtualized views of native mainframe data on the cluster while the original data remains preserved exactly as it was on the mainframe, to meet governance and compliance mandates.

To learn more about Connect eliminate any barriers between your data and your enterprise data platform, visit www.precisely.com/connect

precisely