Informatica

White Paper

Change Data Capture for Real-Time Data Ingestion and Streaming Analytics

Build intelligent, automated, and real-time data pipelines

About Informatica

Digital transformation changes expectations: better service, faster delivery, with less cost. Businesses must transform to stay relevant and data holds the answers.

As the world's leader in Enterprise Cloud Data Management, we're prepared to help you intelligently lead—in any sector, category or niche. Informatica provides you with the foresight to become more agile, realize new growth opportunities or create new inventions. With 100% focus on everything data, we offer the versatility needed to succeed.

We invite you to explore all that Informatica has to offer—and unleash the power of data to drive your next intelligent disruption.

Table of Contents

The Evolution of Modern Data Architectures–Key Trends and Drivers

Data is at the core of how modern enterprises run their businesses and is a crucial enabler in driving digital transformation. Digital transformation has never been more critical than it is today, as the pace of disruption is only accelerating. According to a recent study from Innosight,[1] the average life of companies in the S&P 500 index was 33 years in 1964. It declined to 24 years by 2016, and it is declining at a much more rapid pace today. The prediction is that by 2027 it will decline to just 12 years, and three-quarters of current S&P 500 companies will not exist.

Organizations are trying to become data-centric, but the traditional approaches don't scale and don't provide insights that are required to drive innovation. Over time, enterprises accumulate terabytes and petabytes of data stored in on-premises databases, ERP, and CRM systems. They collect the data, run ETL jobs, and ingest data into a data warehouse such as Teradata, SQL server or an Oracle warehouse. And when the data increases, they add more data warehouse appliances. The challenge with this approach is that it creates data silos. As a result, organizations are unable to create end-to-end, 360-degree views of their customers, markets, and products.

With a modern data architecture, organizations can take advantage of exponential data growth and gain the benefits of end-to-end analytics insights. Migrating from a legacy on-premises data warehouse to a cloud data warehouse and cloud data lake provides benefits such as performance, availability, cost, manageability, and flexibility without compromising on security.

Data architecture is going through three fundamental shifts that are disrupting traditional methods of handling, analyzing, and structuring data.

1. Data Warehouse to Data Lake/Lakehouse

A data lake is a strong complement to a data warehouse. And many enterprises are now adopting a new combined architecture, the "lakehouse." A lakehouse merges data warehouses and data lakes in one data platform. A lakehouse brings the best of both worlds together by combining technologies for business analytics and decision-making with those for exploratory analytics and data science.

The data lake provides cost-effective processing and storage, which is distributable, highly available, and can store data without applying a schema to it. Instead, the schema can be applied later to read the data for analytics consumption. You can store many different data types: structured, unstructured, or semi-structured. Data lakes are critical for organizations that want to be innovative and intend to address artificial intelligence (AI) and machine learning (ML) use cases.

2. Batch Processing to Stream Processing

While there will always be a place for batch processing, there is a notable increase in the demand for streaming content; the need for capture and analysis of real-time data increases as the value of time-sensitive data increases. With the adoption of Kappa architecture and

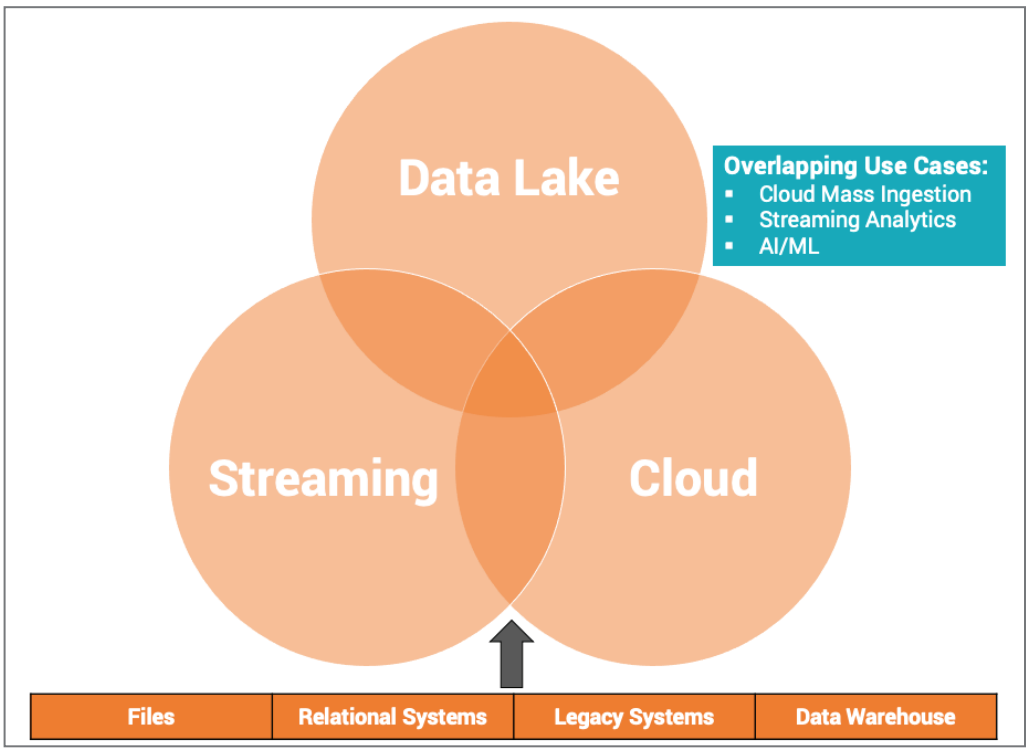
[1] Innosight Corporate Longevity Report

other streaming-first architectural patterns, stream processing has become mainstream.[2] Real-time processing of customer data can create new revenue opportunities, and tracking and analyzing IoT sensor data can improve operational efficiency. Batch processing can also be combined with stream processing to enrich the content even more. Whether it is for strategic decisions or a moment-based decision, stream processing enables organizations to make accurate and faster decisions on fresh data. For example, stream processing enables you to identify cross-sell opportunities when a customer walks into a store. Real-time stream processing helps to capture the customer's location and integrate location data in real time with historical insights from batch data to provide the correct in-moment cross-sell opportunity.

3. On-Premises to Cloud

Cloud has become mainstream as security concerns have abated in most industries. Resource elasticity and cost advantages have made cloud a significant component of multi-datacenter architectures. An industry study reports 83% of enterprise workloads are moving to the cloud.[3] The same number (83%) of organizations are moving their workloads between clouds.[4] Other studies suggest 75% of all databases will be deployed in the cloud within just a few years.

These technology trends enable enterprises to realize benefits such as agility, flexibility, and efficiency, as well as innovation. Businesses can now get better insights from their data and offer the right opportunities to the right individuals with a seamless experience. These fundamental shifts in data architecture are opening up new use cases that were not possible with traditional data management approaches. This is especially true of real-time streaming analytics use cases in the cloud. The Venn diagram below shows the overlapping use cases for data lakes, streaming, and cloud.



[2] Informatica, "Kappa Architecture – Easy Adoption with Informatica Streaming Data Management Solution"
[3] Forbes, "83% of Enterprise Workloads Will Be in the Cloud by 2020"
[4] Turbonomic, 2019 State of Multicloud

Perishable Insights and the Need for Change Data Capture

Today, every industry—healthcare, retail, telco, banking, etc.—is being transformed by data. As data continues to grow, the need for advanced data engineering architectures that combine data lakes, cloud, and streaming becomes critical. This data is also time-bound: data is created in real time and its value diminishes over time. Organizations need to take immediate action on their data when it is fresh, or else they will lose out on business opportunities. The industry term for this is "perishable insights." According Forrester Research, perishable insights can be defined as "Insights that can provide exponentially more value than traditional analytics, but the value expires and evaporates once the moment is gone."

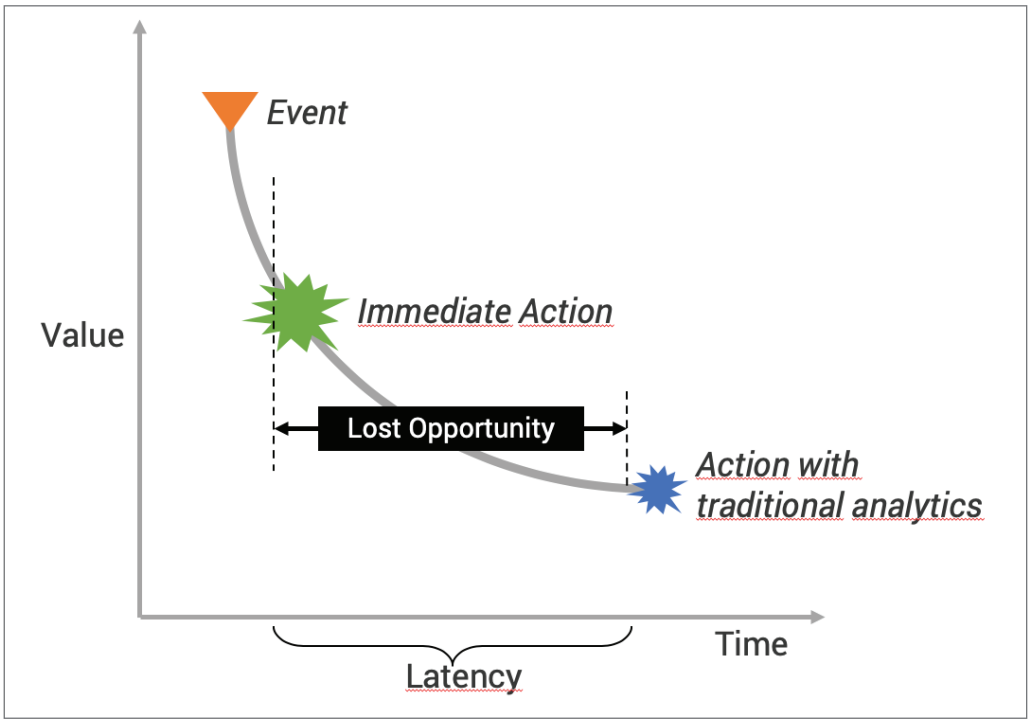


Figure 2: Gaining value from perishable insights[5]

What Is Change Data Capture?

Data with perishable value comes from various sources such as log files, machine logs, IoT devices, weblogs, social media, etc. To ensure you don't miss the opportunities in perishable insights, it's essential to have a means to rapidly capture data changes and updates from transactional data sources. Change Data Capture (CDC) is a design pattern that allows users to detect changes at the data source and then apply them throughout the enterprise.

In the case of relational transactional databases, CDC technology helps customers capture the changes in a system as they happen and propagate the changes onto analytical systems for real-time processing and analytics. For example, you have two databases (source and target), and you update a data point in the source database. Now, you would like to have the same change to be reflected in the target database. With CDC, you can collect transactional data manipulation language (DML) and data definition language (DDL) instructions (for example, insert, update, delete, create, modify, etc.) to keep target systems in sync with the source system by replication

[5] "The BI Watch: Real-Time to Real-Value", Richard Hackathorn

of these operations in near real time.

CDC is used for continuous, incremental data synchronization and loading. CDC can keep multiple systems in sync as well as monitor the source data for schema changes. CDC can dynamically modify the data stream to accommodate schema changes, which is important for different types of data coming in from live data sources. CDC continuously captures real-time changes in data in seconds or minutes. The data is then ingested into target systems such as cloud data warehouses and data lakes or cloud messaging systems, helping organizations to develop actionable data for advanced analytics and AI/ML use cases.

Change Data Capture Use Cases

CDC is used for various use cases such as synchronization of databases (a traditional use case) and real-time streaming analytics and cloud data lake ingestion (more modern use cases).

Traditional Database Synchronization

Imagine you have an online system that is continuously updating your application database. Let's say you have a customer who is registering on your web application. As part of registration, the customer will need to provide information such as name, age, and telephone number. That same record is now created in your source systems. Once the record is created in the source system, the change data is enabled. Essentially, the system will submit that change event out to some form of a listener, and that listener can then process that same change and create a record in the target system. Now, the transactions in both the source and target system will have the same customer records as the data is synchronized. This is an example of when a new record is created. With CDC, we can also capture incremental changes to the record and schema drift. So, imagine the same customer comes back and updates some information—like the telephone number. The changed telephone number gets updated in the source system. CDC will capture this event again and update the record in the target database in real time. You can also define how to treat the changes (for example, replicate or ignore them). This is an elementary example of data synchronization using CDC technology.

Reference architectures can become complex in large enterprises, where different teams are working on different sets of technologies and each has its requirements. When you apply CDC methodology, you will need to modify the application processing mechanism to create a unified solution across all systems. As your data is loaded into different types of discrete systems, you may need to apply in-memory, filtration, and transformation, and other types of operators to the data. Your data is processed and pushed down to the system that is the target for your CDC-based needs. You may want to process your data in batch (this is done mostly for the first time you load your data from source system to your target system). Once you have your initial data loaded, you may want to process the incremental changes in real time.

You can override the schema drift options when you resume a database ingestion job that is in the stopped, aborted, or failed state. The overrides affect only those tables that are currently in the error state because of the Stop Table or Stop Job Schema Drift option.

CDC also can transmit source schema/DDL updates into message streams and integrate with messaging schema registries to ensure that analytics consumers understand the metadata.

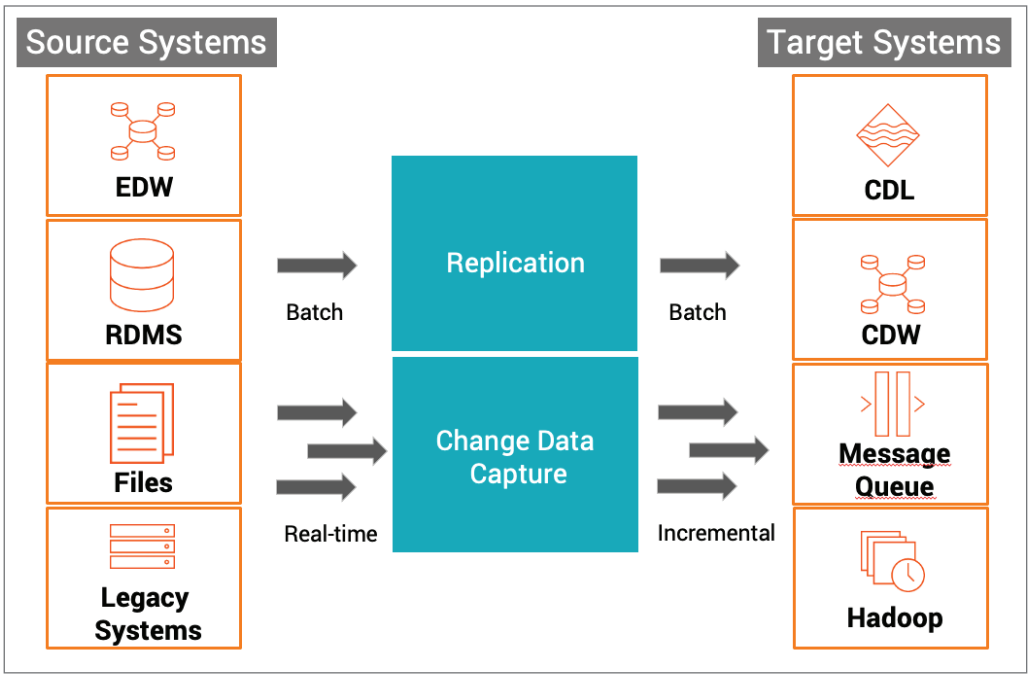


Figure 3: Synchronization of traditional database with change data capture.

Modern Real-Time Streaming Analytics and Cloud Data Lake Ingestion

In a modern data architecture, customers can continuously ingest CDC data into a data lake in the form of an automated data pipeline. CDC can propagate data to message queues to help analyze streaming log data for any kind of issue. In this case, data is pushed to a messaging queue like Apache Kafka or Amazon Kinesis as a target for reading and processing data. CDC can also be leveraged for data migration from on-premises to cloud, advanced analytics, and ML use cases by analyzing data to generate real-time insights and data modernization initiatives.

CDC helps avoid some of the bottlenecks you may encounter when provisioning large amounts of data from legacy data stores to the new data lake as the data only needs to be loaded once; thereafter, CDC just provisions the changes to the data lake. In fact, many organizations have used data lake over ETL platforms, as the data lake environment tends to be less expensive. The most difficult part of the data lake is maintaining it with current data. CDC can be extremely helpful there. It can help you save computing and network costs, especially in the case of cloud targets, as you are not moving terabytes of data unnecessarily across your network. You can focus on
the change in the data. With support for technologies like Apache Spark for real-time processing, CDC is the underlying technology for driving advanced real-time analytics, and AI/ML use cases.

Now, let's deep dive into two real-world examples of how customers are taking advantage of CDC technology to address real-time analytics use cases: real-time fraud detection at a bank and targeted, real-time marketing campaigns at a financial conglomerate.

Real-Time Fraud Detection

A large corporate bank was facing challenges with a sudden increase in fraudulent activities, which resulted in unhappy customers and loss of business. The bank wanted to build a real-time analytics platform to proactively alert customers about potential fraud–so that customers could take remedial actions. To do that, the bank needed to ingest transactional information from its database in real time and apply a fraud detection ML model to identify potentially fraudulent transactions.

Informatica® Change Data Capture captures the change data in real time from the transactional database and publishes it into Apache Kafka. Informatica Data Engineering Streaming reads this data and then transforms and enriches the data for real-time fraud analytics that enable the fraud monitoring tool to proactively send text and email alerts to customers about potential fraud detection. As a result, the bank is able to improve customer experience, thereby helping to retain and grow the customer base.

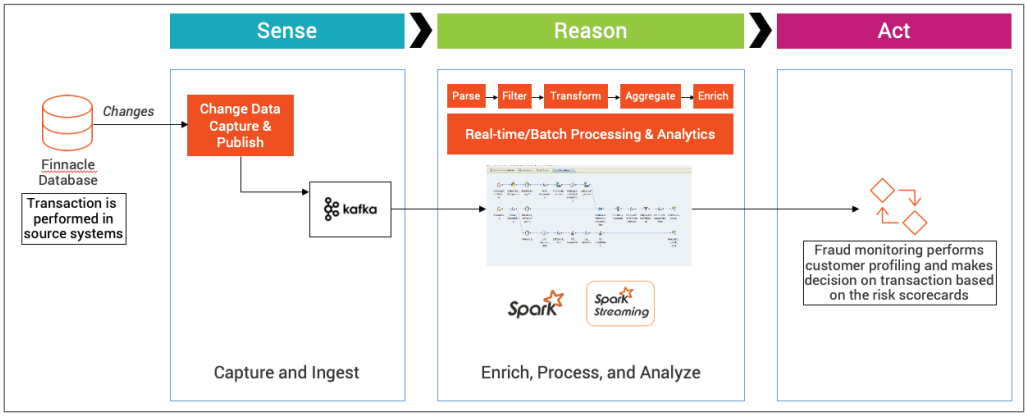


Figure 4: Reference architecture for real-time fraud detection.

Targeted, Real-Time Marketing Campaigns to Improve Customer Experience

A financial conglomerate with a mobile payment platform processes millions of transactions a day for hundreds of thousands of customers. All this data is processed through payments and purchasing transactions that take place in shopping malls where the business's merchants reside. The company tracks real-time customer activities on mobile phones (for example, transactions made while customers are topping up their cash card, making in-store payments, paying for taxi rides, etc.) All these transactions are captured by CDC and fed into Apache Kafka. Informatica Data Engineering Streaming reads this data, integrates these streams, and feeds them to a data lake for offline batch processing and into Kinetica, the company's real-time processing engine. The company correlates the real-time data with historical data, then matches the data against business rules to provide the next-best action to the customer, sending an offer on the mobile
app while the customer is at the shopping mall.

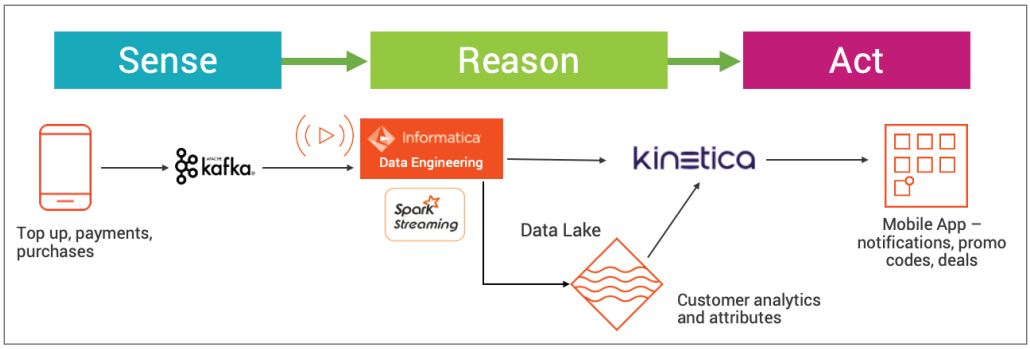


Figure 5: Reference architecture for targeted, real-time marketing campaigns.

Benefits of Change Data Capture

CDC captures changes from the database transaction log and publishes it to a destination such as a cloud data lake, cloud data warehouse, or message hub. This has several benefits for the organization.

1. Enables Faster Decision Making: The biggest advantage of CDC technology is that it fuels data for real-time analytics. This helps organizations to make faster and more accurate decisions in real time by capitalizing on fast-moving data with perishable value.

This means finding data, analyzing it, and acting on it in real time. Organizations can create hyper-personal, real-time digital experiences for their customers with real-time analytics. For example, real-time analytics can enable restaurants to create a personalized menu for individual customers based on historical data, along with data from mobile or wearable devices to provide customers with the best deals and offers.

To take another example, an online retailer wants to sell more motorcycle helmets and maximize profits. The retailer detects a pattern in real time that when a user views at least three motorcycle safety products, including at least one helmet, analytics indicate that the customer is interested. The retailer then displays the most profitable helmets. Now, add a time window parameter to this to find the real-time total sales of motorcycle helmets. The reason to do this is to move the price up and down. The first goal is to sell more motorcycle helmets, and the second goal is to maximize profitability. If sales are trending lower than usual, and this customer is price sensitive, then the retailer dynamically lowers the price.

2. Minimizes Impact on Production: When done through a tool, CDC brings many advantages compared to script-based or hand-coded implementation. Since it is time-consuming to move data from source to the production server, CDC captures incremental updates with a minimal source-to-target impact. It can read and consume incremental changes in real time to continuously feed the analytics target without disrupting production databases. This helps to scale efficiently to execute high-volume data transfers to the analytics target.

3. Improves Time to Value and Lowers TCO: CDC enables you to build your offline pipeline faster without worrying about scripting. It helps data engineers and data architects to focus on important tasks. It also helps minimize cost and total cost of ownership by removing the dependency on highly skilled users for these applications.

Methods of Change Data Capture

There are several methods and technologies to achieve CDC, and each has its merit depending on the use case. Here are the common methods, how they work, along with their advantages and disadvantages.

Timestamps: The simplest way to implement CDC is to use a timestamp column within a table. This technique depends upon a timestamp field being available in the source table(s) to identify and extract change datasets. At minimum, one timestamp field is required for implementing timestamp-based CDC. In some source systems, there are two timestamp source fields—one to store the time at which the record was created, and another field to store the time at which the record was last changed. The timestamp column should be changed every time there is a change in a row.

Timestamps are the easiest to implement and most widely used CDC technique for extracting incremental data. However, this approach only retrieves rows that have been changed since the data was last extracted. There may be issues with the integrity of the data in this method. For instance, if a row in the table has been deleted, there will be no DATE_MODIFIED column for this row, and the deletion will not be captured. This method can also slow production performance by consuming source CPU cycles.

Triggers: Another method for building CDC at the application level is defining triggers and creating your own change log in shadow tables. Shadow tables may store the entire row to keep track of every single column change, or they may store only the primary key and operation type (insert, update, or delete). Using triggers for CDC has the following drawbacks:

- Increases processing overhead
- Slows down source production operations
- Impacts application performance
- Is often not allowed by the database administrators

Log-Based CDC: Transactional databases store all changes in a transaction log that helps the database to recover in the event of a crash. With log-based CDC, new database transactions—including inserts, updates, and deletes—are read from source databases' transactions. Changes are captured without making application-level changes and without having to scan operational tables, both of which add additional workload and reduce source systems' performance. Log-based CDC is the most preferred, fastest, and least disruptive CDC method because it requires no additional modifications to existing databases or applications.

Informatica Solutions for Change Data Capture

Informatica solutions for real-time streaming analytics and mass ingestion provide end-to-end multi-latency data management with versatile connectivity.

The Informatica Cloud Mass Ingestion Service provides database ingestion capabilities that help you ingest initial and incremental loads from relational databases such as Oracle, SQL-Server, and MySQL onto cloud data lakes and cloud data warehouses as well as messaging hubs. It also offers schema drift capabilities to help customers manage changes in the schema automatically and provides real-time monitoring on ingestion jobs with lifecycle management and alerting capabilities.

Enterprises have data from a variety of sources, such as on-premises files, databases, data warehouses, streaming data, and SaaS systems like ERP and CRM. Data needs to be ingested from all these sources into a cloud data lake or stream storage like Apache Kafka to be enriched, processed, and transformed. Once the data is in the cloud data lake, data quality rules and integration are applied to the data and then land the data into the cloud data warehouse to make it ready for advanced analytics and AI/ML use cases.

The streaming pipeline feeds data for real-time analytics use cases, such as real-time dashboarding and real-time reporting. Cloud Mass Ingestion supports the following use cases:

- **Cloud Data Warehouse/Cloud Data Lake Ingestion**: Enable data ingestion from a variety of sources—such as data lakes and data warehouses, files, streaming data, IoT data, and on-premises database content—into a cloud data warehouse and cloud data lake to keep the source and target in sync.
- **Data Warehouse Modernization/Migration:** Mass ingest data from on-premise databases into a cloud data warehouse and continuously ingest CDC data to keep the source and target in sync. This applies to on-premises legacy systems, such as mainframes, and relational databases, such as Oracle, IBM DB2, Microsoft SQL, and others.

- **Accelerate Messaging Journey for Real-Time Analytics:** Ingest data from a variety of sources—such as logs, clickstream, social media, IoT, and CDC data—into Kafka or other messaging systems for real-time operationalization and reporting use cases.

Informatica Data Engineering Streaming helps customers continuously ingest and process data from a variety of streaming sources by leveraging open source technologies like Apache Spark and Apache Kafka. Data Engineering Streaming provides out-of-the-box capabilities to parse, filter, enrich, aggregate, and cleanse streaming data while also helping operationalize machine learning models on streaming data. With these capabilities, customers can perform real-time analytics on CDC data to address their streaming analytics use cases.

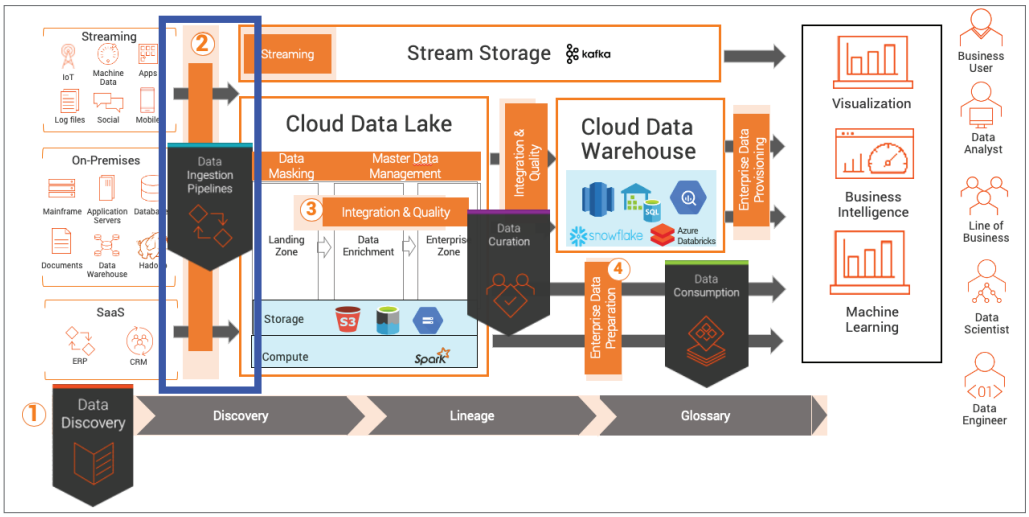


Figure 6: Cloud data warehouse/data lake reference architecture

Conclusion

Real time is the new measurement for digital success, and real-time data has increased in potential value. The need for immediate, intelligent responses is now paramount. Technologies like CDC can help companies gain digital superiority by accelerating business innovation and gaining competitive advantage. CDC technology helps businesses make better decisions, increase sales, and improve operational costs. With Informatica solutions for CDC, companies can quickly move and ingest a large volume of their enterprise data from a variety of sources onto the cloud or on-premises repositories for processing and reporting—or onto messaging hubs for real-time analytics with out-of-the-box connectivity.

Next Steps

To learn more about Informatica solutions for streaming and ingestion, visit the Cloud Mass Ingestion webpage and read the following datasheets and solution briefs:

- Informatica Cloud Mass Ingestion datasheet
- Informatica Data Engineering Streaming datasheet
- Ingest and Process Streaming and IoT Data for Real-Time Analytics solution brief

Informatica

Worldwide Headquarters 2100 Seaport Blvd., Redwood City, CA 94063, USA Phone: 650.385.5000, Toll-free in the US: 1.800.653.3871

IN09_1120_03914