Informatica | Datumize

Solution Brief

Real-Time Streaming Analytics for Dark Data Sources

Key Benefits

- Collect and leverage dark data (data not stored and therefore not used)
- Expand your data by tapping into your organization's dark data
- Manage a large volume and variety of data with on-the-edge preparation and real-time ingestion
- Discover new business and operational insights

The digitally-transformed world requires companies to be fast, adaptable, and data-driven. As a result, leading companies are adopting real-time streaming analytics solutions to unlock the potential of new data sources. Having better and faster information is a key competitive advantage.

The increased complexity of IT infrastructures and the proliferation of new systems, Internet of Things (IoT) devices, sensors, and integrations, make it difficult for companies to discover and sort the relevant data and leverage it to generate compelling insights, especially considering that the value of most of this data quickly degrades with time.

Key Features

Lightweight and High-Performance Dark Data Collection Engine

Dark data represents data not collected and therefore not used. Considering that at least 80% of data is dark,[1] it represents the great hidden resource that flows untapped through major organizations. Leveraging dark data in the data integration processes is a challenge as the most sophisticated data sources—such as network transactions, IoT, mobility, Wi-Fi, or industrial networks—require an advanced engine specially built for the purpose.

The powerful, flexible, and high-performance Datumize Data Collector engine enables organizations to collect dark data by using unobtrusive network sniffing or active polling techniques, depending on the source. The modularity of the system enables the design of real-time data gathering flows for a variety of sources, installations, architectures, integrations, and purposes. Datumize Data Collector provides the ultimate data collection and edge processing capabilities for protocols such as HTTP, SOAP, OPC-DA, SNMP, and others, so that data can be prepared and ingested into streaming platforms.

[1] Outside Insight, "What is dark data and how can it benefit your company?"

Leverage and Prepare Dark Data for Streaming Analytics

Datumize Data Collector helps enterprises capture dark data from several data sources, process that data in real time to transform and extract valuable information, and store actionable results, using a well-known ETL approach (capture, process, store). Datumize Data Collector is able to:

- Capture data in real-time from any data source based on connectors, with each connector dealing with a particular protocol (i.e., HTTP, SOAP, SNMP, OPC-DA) and capture method (i.e., network sniffing, polling). Temporary data (data flowing on a network) and closed/proprietary data (data exchanged between proprietary nodes) is preferred, as this data yields higher value.
- Process captured data in real time, at the edge, correlating and enriching individual events, applying AI algorithms (for example, A* for pathfinding), extracting valuable information, computing additional metrics using standard operations (i.e., max, avg) or custom logic (scriptable), and standardizing or transforming them into a new data format.
- Store the resulting information in any destination platform for further processing and more in-depth analytics, whether in on-premises storage or the cloud. Datumize Data Collector can be integrated with out-of-the-box analytics and raise business and technical alerts.

Streaming Analytics: Process, Enrich, and Analyze Streaming Data in Real Time

Informatica® Data Engineering Streaming uses the Sense-Reason-Act framework to help data engineers ingest, analyze, and act on the broadest range of real-time streaming data and make decisions while the events are still happening. It extracts value from fast-moving data streams and helps organizations in a wide variety of industries such as retail, healthcare, banking, telco, manufacturing, and more.

Informatica's approach to real-time streaming analytics starts with collecting the raw data from the various sources and ingesting the data into the data lake or messaging hub. Informatica also offers data transformation and data enrichment capabilities to process the streaming data and make it available for operationalization and downstream analytics.

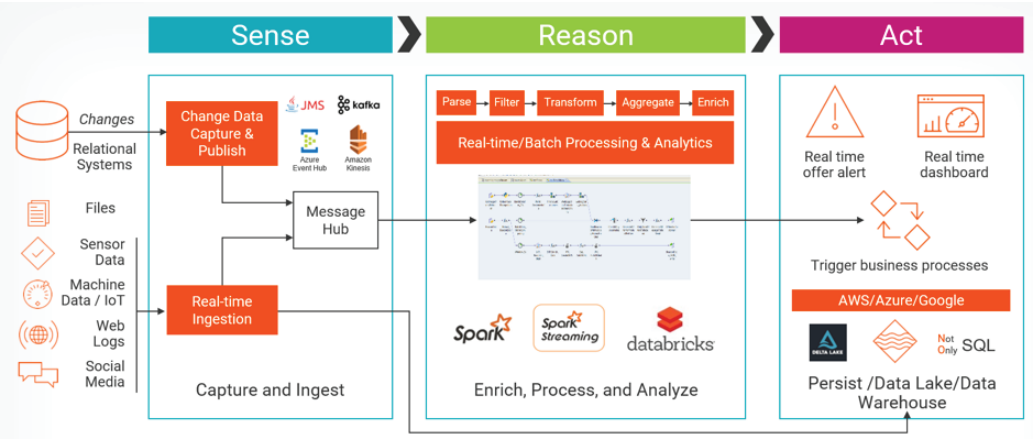

Figure 1: The Sense-Reason-Act framework.

Informatica Data Engineering Streaming is built on best-in-class open-source technologies in an easy-to-use, enterprise-grade offering. It primarily uses open-source Spark Streaming under the covers for stream processing and supports other technologies like Apache Kafka, Azure Databricks, and Databricks Delta. As new technologies inevitably evolve, Informatica Data Engineering Streaming adapts, allowing customers to reuse existing data streams. You can schedule data flows to run at any latency (real time or batch) based on available resources and business SLAs.

The Combined Datumize and Informatica Solution

For Datumize, the Informatica Data Engineering Streaming solution represents a powerful, sophisticated and enterprise-ready platform to unleash the value of data captured with Datumize Data Collector to sense, reason, and act on trusted data, all within the context of dynamically changing information, business rules, and analytic models.

For Informatica, Datumize acts as a powerful data ingestion engine that can capture data from sophisticated dark data sources, process them at the edge, and ingest this data in real time into the Informatica Data Engineering Streaming solution.

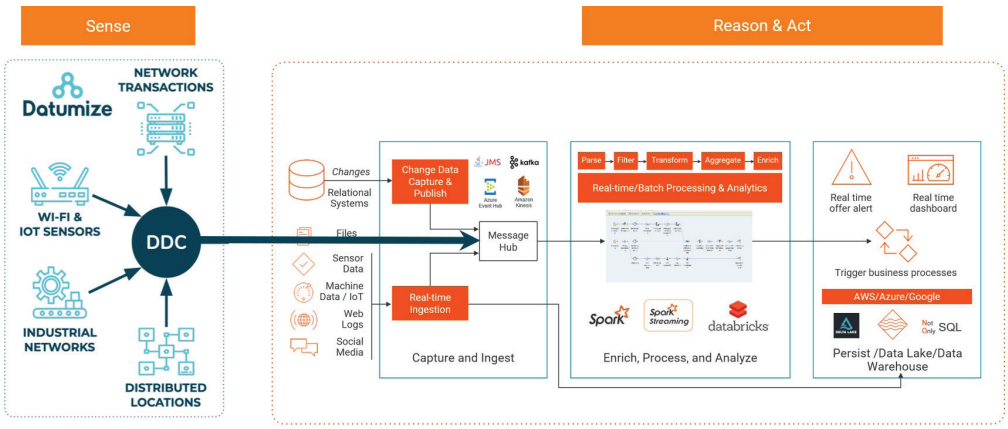

Figure 2: The combined Datumize and Informatica solution.

From Network Traffic to Real Demand

Large volumes of in-transit data remain hidden and unleveraged due to the difficulty of collecting and processing temporary transactions like API or XML integrations flowing over the network without the hassle of modifying the backend systems.

Datumize Data Collector network traffic connectors leverage network sniffing and deep packet inspection techniques to capture live network traffic flowing over physical and virtual networks and process various protocols such as HTTP, SOAP, or industrial protocols like OPC-DA. This unique approach enables the reconstruction of in-flight transactional data straight from the wire without any impediment to the existing system. It offers edge computing that will radically reduce the need for further processing down the data pipeline, and supports ingestion in real time into Informatica Data Engineering Streaming, where prepared data is transformed into valuable real demand and customer insights.

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.

About Datumize

Datumize is a software vendor established in 2014 in Barcelona, Spain, working on data integration technology.

We develop innovative products that allow companies to enjoy actionable insights based on Dark Data—data not stored and therefore not used.

Our secret sauce is a proprietary and powerful data collection engine, Datumize Data Collector (DDC), that gets data from sources that most other vendors do not consider.

From IoT to Motion Intelligence

The proliferation of mobility sensors and IoT devices represents a rich source of data, especially for motion intelligence and location-based analytics purposes. But enterprises find it challenging to efficiently access this low-level and technical data and transform it into real-time business and operational insights.

This dark data can be actively polled or unobtrusively network sniffed with Datumize Data Collector. The sniffing feature is particularly useful for IoT and industrial domains, many of which use closed and proprietary data sources. Wi-Fi mobility and sensor data are usually polled using SNMP or ad hoc APIs. This capability allows enterprises to collect very low-level metrics from an existing Wi-Fi infrastructure (for example in a logistics warehouse, a shopping center or a hotel) and transform that information into valuable motion intelligence (position, movement, distance), without the need for costly infrastructure investments.

The relevant captured data from IoT, Wi-Fi, or mobility sensors can be ingested into the Informatica Data Engineering Streaming Analytics solution. There, massive amounts of real-time data are parsed, processed, analyzed, and transformed into persisted metrics and insights, following the Informatica principles of Sense-Reason-Act.

From Industrial Data to Operational Insights

Highly valuable operational data is trapped inside industrial machines, devices, and sensors. Vendor lock-in, proprietary protocols, and a lack of interoperability have inhibited machine data from being shared and used to govern and unlock efficiencies. Industrial data integration is limited to polling techniques that represent an overhead to existing supervisory control and data acquisition (SCADA), programmable logic controller (PLC), and control systems. Network sniffing, combined with smart polling techniques, represents a competitive advantage that Datumize Data Collector exploits to capture industrial data for a number of different protocols, introducing no network overhead and secure collection from existing systems, while providing modern data ingestion capabilities into Informatica Data Engineering Streaming, to create insights and advanced metrics that go beyond the traditional operational technology (OT) capabilities.

Next Steps

Learn more by visiting the Datumize Data Collector page on the Datumize website, as well as the Informatica Data Engineering Streaming and Informatica Data Engineering Integration product pages.

Informatica

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

IN17_0220_03839