Historian Integration with Cloud Data Lakes

Challenges and Best Practices

DEEPIQ

Industrial data historians such as AVEVA's OSI PI™, AspenTech's IP21™, Honeywell's PHD™, GE's Proficy™, and Canary Labs™' are recognized as industry leaders in capturing and managing operational data. Organizations increasingly demand that this data be available directly within their data lakes, integrated with other enterprise data, instead of existing in an operational silo. This is motivated by the following:

- Need to provide operational visibility, not just at a plant or an asset level but at an enterprise level across all the operations
- Ability to apply machine learning and AI across data from multiple assets and other relevant enterprise data
- Integrate supply chain and CRM data such as orders and forecasts with operational data to manage the throughput of the manufacturing and supply chain processes
- Ability to develop digital twins and utilize agentic systems for intelligent maintenance and task automation
- Need for data residency and cost optimization when dealing with large volumes of highly sensitive data

Integrating historian data into the enterprise data lake is critical to unlocking transformative use cases such as predictive maintenance, real-time process optimization, energy management, demand driven manufacturing and AI-driven operations.

Understanding the Complexities of Historian Integration

While the need for cloud data lake integration is clear, the process is not simple. Historians' strength in ingesting large data volumes becomes their most significant challenge when extracting data. Here is a closer look at the key hurdles:

1. Data Ingest vs. Egress Speeds

Historian systems typically store years of historical data while simultaneously ingesting real-time streams.

This creates a significant challenge in data migration, as the data extraction rate often cannot keep pace with data ingestion, resulting in potential data backlogs during cloud integration. Let us examine an illustrative scenario to explore the impact on data lake integration. Although this scenario does not consider certain factors, such as data compression rates for data at rest or in motion, the primary insights remain accurate. At time *t*, your historian server has the following data attributes.

- ***Z*** units: Amount of data already stored in the historian server
- ***X* units/time**: Incoming data rate (per unit of time)
- ***Y* units/time**: Maximum data extraction throughput of the historian server

To ensure the successful migration of historian data into a data lake, such that all data is available in the data lake within a specified time frame T starting at time t without any data loss, the following relationship must be satisfied:

$$Z + X * T = Y * T$$

Rearranging for **T**:

$$T = \frac{Z}{Y - X}$$

Equation 1

In practice, a standard historian's data extraction rate (***Y***) is often significantly lower than its data ingestion rate (***X***). This discrepancy renders the time required for complete data migration (***T***) undefined, making it impractical to synchronize a new platform with the continuously growing data volume in the historian system without implementing alternative strategies.

In addition, several customers are on older versions of these systems without a desire to upgrade, further limiting their ability to achieve the current versions' performance.

Example Scenario

Consider a scenario where:

- The historian system contains **10 TB** of historical data.
- New data is ingested into the historian at **5 MB/sec**.
- Data extraction can only proceed at **2 MB/sec**.

At an extraction rate of 2 MB/sec, it would take approximately **61 days** to extract 10 TB of historical data fully. During this time, the historian continues to ingest new data at a rate of 5 MB/sec. Over 61 days, this ongoing ingestion would result in the addition of **25 TB** of new data.

By the time the historical data extraction is complete, the system will have accumulated an additional **25 TB of data** yet to be extracted. This means the data backlog would grow faster than could be cleared, making it mathematically impossible to catch up under these conditions.

2. Importance of the Asset Hierarchy & Asset Groupings

The complexity of asset hierarchies and metadata, such as those managed in OSI PI's Asset Framework (AF), is vital for understanding the data. Transferring this contextual information without loss of fidelity is a significant undertaking that requires careful handling to maintain data integrity and usefulness in the cloud environment. Companies invest considerable amounts of time and resources in building these hierarchies, making preserving contextual data a primary concern during migration. Ensuring an accurate transfer of this data to a new platform is complex, error-prone, and time-intensive, further adding to the difficulty of the process.

A Better Cloud Integration Strategy

The above section, particularly Equation 1, might seem to imply that it is impossible to have a complete and clean data replication in your cloud data lake. Here is the good news. Over the last few years, DeepIQ has developed innovative, software-driven solutions and deployment models that streamline the migration process while maintaining the integrity and usability of industrial data.

DeepIQ software is engineered to scale up to the vast data volumes typical of historian systems by integrating sophisticated technologies. Our method ensures high throughput and system efficiency, outlined through the following advanced strategies:

1. **Optimized Data Export Utilization:** DeepIQ capitalizes on the native export capabilities of historian systems, such as AF SDK for PI, SQLPlus for IP21, and OPC HDA for Honeywell PhD. This utilization is crucial for maximizing throughput by efficiently managing the export processes.
2. **Advanced Parallelization Techniques:** Our software features a configurable, multi-threaded approach to parallelizing data reads from historians. This enables significant enhancements in throughput by allowing the concurrent processing of large data volumes, thereby optimizing performance and reducing retrieval times.
3. **Comprehensive Protocol Support and Enhanced Data Handling:** DeepIQ supports various protocols to connect directly with control systems or other data sources feeding historians. We extract metadata from historians and raw telemetry data from source systems separately. These are then paired within the cloud data lake through DeepIQ workflows. This separation allows for more efficient data management and bypasses historian egress throughput limitations by delegating data pairing tasks to our powerful cloud-based workflows

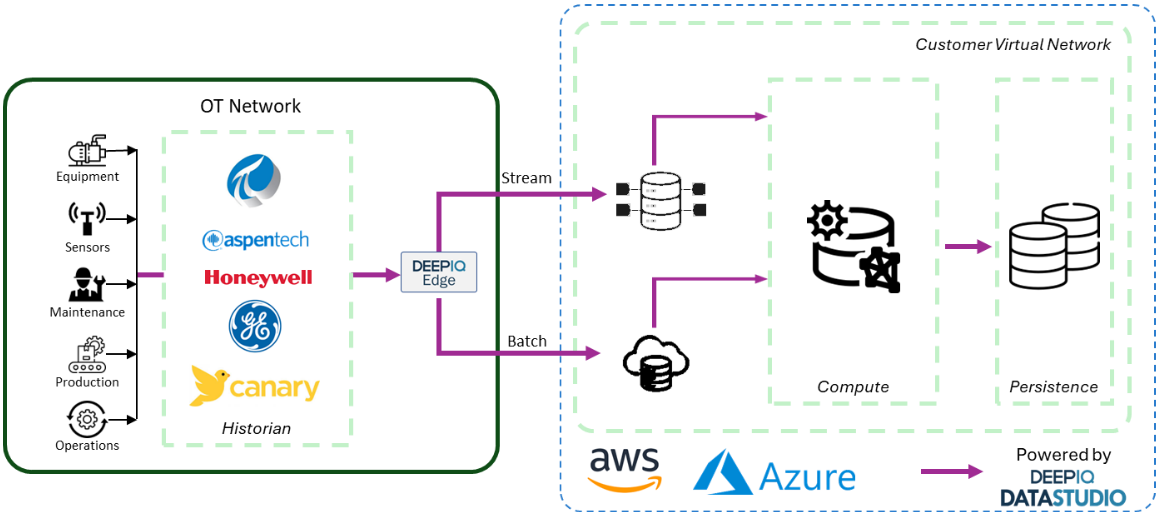

Figure 1: Architecture

Figure 1 depicts the architecture of this deployment. DeepIQ Edge software facilitates a pool of connections for data extraction from historians, dynamically adapting its requests to meet the throughput constraints of each system while capturing hierarchies and other relevant metadata. DeepIQ DataStudio serves as the central hub for orchestrating and managing these Edge requests. The Edge software is capable of pushing data to various landing zones, including Event Hub or ADLS Gen2 on Azure, or S3 or Kinesis on AWS. Meanwhile, DeepIQ DataStudio utilizes distributed computing frameworks such as Databricks, Azure Synapse Spark pools, or AWS EMR to efficiently transfer the data into your data lake.

Comprehensive Data and Context Migration

DeepIQ's software does not just transfer raw data—it ensures the migration of essential contextual elements, such as Asset Framework (AF) hierarchies, calculated tags, and the logic behind those calculations. Capturing these components is vital for maintaining the operational insights and structured organization that the historian natively provides.

Robust Versioning of Asset Hierarchies

Asset hierarchies capture important contextual data such as equipment configurations, asset groupings, and operational models. DeepIQ's software includes robust versioning capabilities, enabling organizations to capture and store changes to hierarchies over time. This functionality enables users to track modifications, compare historical and current states, and revert to previous versions as needed, ensuring data integrity and traceability throughout the migration process.

Seamless Data Mapping

Calculated tags and their associated logic are critical for maintaining operational continuity. DeepIQ's software ensures these elements are accurately mapped and migrated, enabling uninterrupted calculations and derived metrics functionality in the new environment.

Scalability and Real-Time Performance

Cloud platforms offer unparalleled scalability for batch and streaming workloads. DeepIQ's approach to time series data modeling and versioning enables cloud environments to replicate the performance of industrial historians, ensuring ultra-low latency and the ability to handle high-throughput industrial data.

Deep IQ: End-to-End Capabilities for Historian Integration

DeepIQ's platform provides a comprehensive and unified solution to address the complexities of historian migration efficiently:

- Data Reading: DeepIQ Edge seamlessly reads and normalizes data from multiple historians such as OSI PI™, IP21™, PHD™, Proficy™, and Canary Labs™. This includes capturing Asset Framework (AF) hierarchies, tag mappings, and other critical metadata, ensuring no essential structure or context is left behind.

- Edge Data Normalization: DeepIQ Edge has the capability to access source systems that supply data to historians directly. Leveraging asset hierarchies or other available metadata from these data feeds ensures that the data sent to the cloud is normalized and modeled for seamless integration.

- Protocol Versatility: The platform supports many protocols, such as MQTT, OPC UA, and others, enabling smooth and direct data extraction from source systems.

- Advanced Hierarchy and Time-Series Management: DeepIQ's tools efficiently handle evolving hierarchies and slow-changing dimensions. Additionally, the platform supports large-scale time-series data processing with distributed compute capabilities, ensuring scalability and performance for complex and data-intensive migrations while maintaining consistency and context.

DeepIQ's solutions are purpose-built to tackle large-scale IT-OT integration projects. At DeepIQ, we believe in giving our customers the freedom to adapt and evolve confidently. Visit our website (www.deepiq.com) or contact info@deepiq.com for more resources and insights into similar topics.