White Paper
May 2025

IT@Intel: Transforming Siloed Manufacturing Data into Unified Insights

Intel IT has created a resilient manufacturing data pipeline to address data explosion and security while enabling a connected data store for real-time reporting and analytics

Intel IT Authors

Subhadra Sampathkumaran
Software Engineering Manager, Industrial Process Control

Paul Schneider
Principal Engineer, Director of Automated Factory Solutions

Table of Contents

Executive Summary

Intel's manufacturing landscape is evolving rapidly, driven by digital transformation and automation. Intel's manufacturing facilities generate vast amounts of data that must be efficiently processed, analyzed, and secured to optimize operations and maintain competitiveness.

In 2013, we deployed an "ultra" data warehouse that replaced our older fragmented and siloed databases. It has stood the test of time, easily scaling as we added yet more data and data domains. As Intel transitions to a foundry model, its factories will become more complex and will generate even more data than before. Plus, security is becoming even more critical—Intel must not only protect its own manufacturing data from unauthorized access; it must also ensure that Intel Foundry customers have confidence that their data is in safe hands and guaranteed to be separate from Intel's data. We have implemented process-specific security (PSS), which helps prevent unauthorized data access while still delivering efficient query execution.

Our highly secure and high-performance data warehouse, running on modern Intel® architecture, helps unlock important insights hidden in massive manufacturing data volumes. This solution has shown that it can reliably scale to handle not only today's data challenges but also grow with Intel as it expands its factories to support its foundry model and IDM 2.0.

Intel IT Contributors

Eric Messenger, Director of Corporate Data and Analytics
Joe Sartini, Industry Engagement Manager
Robert Vaughan, Industry Engagement Manager

Acronyms

ER	entity resolution
ETL	extract, transform, load
HVM	high-volume manufacturing
IoT	Internet of Things
MES	manufacturing execution system
PSS	process-specific security
RDBMS	relational database management system
SPC	statistical process control

Business Challenge

Intel's factories have always been the heartbeat of the company, churning out silicon that powers a great portion of the world's computing environments—and generating increasingly vast amounts of manufacturing data. Intel manufacturing data has grown 2x–20x (depending on the specific area) over the last two decades (see Figure 1). This data comes from many sources, including:

- Automated systems like robotic material handling
- Real-time monitoring systems driven by Internet of Things (IoT) devices
- Increasingly complex manufacturing processes

Over the years, Intel IT has worked to efficiently manage and secure this data despite challenges presented by the continuing evolution of Intel's manufacturing data systems and the unique data characteristics of the high-volume manufacturing (HVM) environment. In the early to mid-2000s, our manufacturing data was siloed in various systems, leading to a fractured data environment that limited end-to-end data visibility, obscured insights, and hindered data-driven decisions across the factory. As one Intel factory manager stated, "If I can't see the data, I'm driving the factory blind."

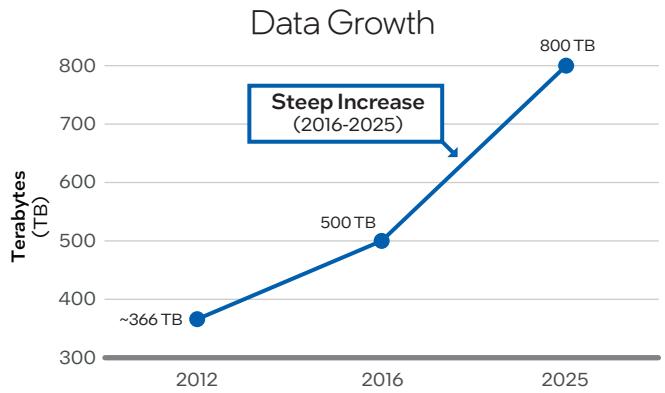

Figure 1. Intel's manufacturing data has grown steadily over the past two decades.

The Evolution of Data Systems at Intel's Factories

As Intel's HVM environment evolved, we developed a variety of in-house databases and other tools to manage factory data. One database contained quality data; another contained sort data; another contained process control data; and so on. As new data domains appeared, such as data from IoT sensors, we added new schemas. Although these systems were integral to manufacturing operations, they were not interconnected, making it difficult to achieve traceability across the wafer and lot lifecycle. The databases also had minimal analytics and reporting capabilities. Tracking wafers from inception to completion across multiple databases posed a significant challenge. The lack of a centralized data repository led to:

- Siloed data sources that made cross-system analysis difficult.
- Traceability gaps, making it challenging to track wafer movements across process steps.
- High latency in data access, with integration queries sometimes taking 30 minutes or more.
- Inconsistent data models across different databases, complicating analytics and reporting.

In addition, although the legacy databases had business continuity and disaster recovery capabilities, the entire set of databases and tools ran on legacy hardware that was becoming increasingly unreliable because of degrading performance.

HVM Data Challenges

Our manufacturing data exhibits the following complexities:

- **High dimensionality.** Each chip has multiple unique IDs based on process steps, similar to how a person can have several IDs such as a social security number, email address, and cell phone number.
- **High volume of historical data.** For engineering analytical purposes, we gather and store data about every component of every chip. In some cases, where the business logic and data transformation were intensive, we performed these tasks close to the data and stored aggregated data sets from all sites in a single table for consumption by reporting and analytical applications across Intel. In other cases, the Data Integration Server combined the data from each site and presented a unified view for consumption.
- **High velocity of incoming data.** The manufacturing execution system (MES) handles about one million transactions per minute in real time. These transactions are condensed to about 200,000 records across all tables at each site, every five minutes.
- **Varying consumption models.** Engineers require access to both historical and real-time data, leading to unpredictable query patterns.

The dynamic nature of our manufacturing data requires scalable architectures that can store, process, and retrieve insights efficiently—something the outdated, fragmented data landscape couldn't do.

Solution: A Highly Scalable Data Warehouse

In 2012, we began the transition to an engineered database system—a large, centralized repository for integrated, historical manufacturing data analysis. We chose this "ultra" commercial data warehouse solution because it could overcome the limitations of the fragmented factory data landscape that we were experiencing in the mid-2000s. We evaluated other approaches, such as NoSQL, but we chose the commercial, centralized data warehouse solution to enhance support for high-throughput analytics workloads and keep up with continued data growth.

The new data warehouse went into production in 2013. For over a decade, the system has successfully scaled with Intel's growth (our data volume is now around 800 TB per factory site) and can quickly and easily accommodate new types of data, along with new data from Intel Foundry customers as they onboard. Figure 2 depicts the system architecture for our data warehouse.

We have a centralized data warehouse that contains all factory data and operates on modern Intel® architecture. This data warehouse can serve about 8% of the applications that need cross-site data analysis and can also help run intensive queries. In addition, each factory has its own site-specific data warehouse for performance and security reasons, which also runs on Intel architecture. Each data warehouse aggregates data from multiple manufacturing systems, which helps to ensure a unified, structured, and accessible repository with a cohesive dataset for improved traceability of wafers across their lifecycle. The primary systems that send data to the data warehouse include:

- **Manufacturing execution system (MES):** Tracks work-in-progress, manages lot movement, and records process execution data.
- **Statistical process control (SPC):** Monitors and controls variations in manufacturing processes, ensuring product consistency.
- **Fault detection and classification (FDC):** Collects and processes tool sensor data to monitor equipment health and predict failures.
- **Process control and dispatching system:** Optimizes lot scheduling and reduces cycle time by prioritizing workflows.

These systems feed structured and unstructured data into the data warehouse. Data domains include the following:

- Intel wafer fabrication data
- Manufacturing indicators and line management
- Yield, integration, and quality
- Planning and new product information
- Finance and global supply chain
- Intel Foundry

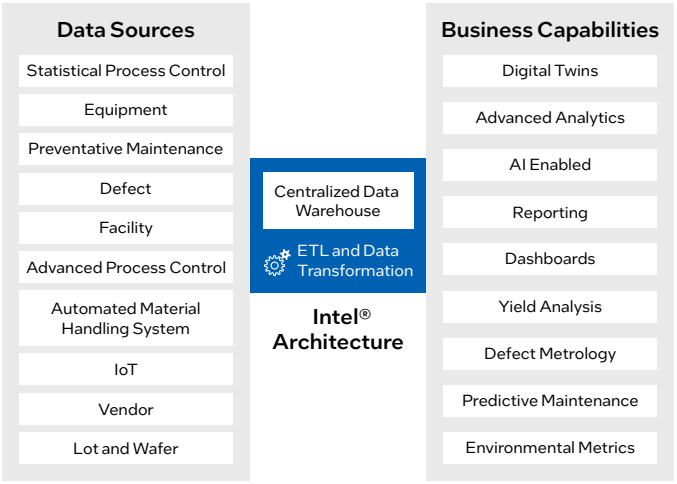

Figure 2. System architecture for our centralized manufacturing data warehouse.

Legacy Database Migration Considerations

We adopted multiple strategies for migrating data to the new data warehouse. Primarily, we used extract, transform, load (ETL) logic developed in-house that performed an incremental synchronization from a legacy database to the data warehouse. This ETL logic enabled us to transfer data from one data schema to another through data transformations. During the final cutover, we briefly shut all the systems down and performed one final synchronization. We also verified data integrity and that the migration did not result in any data loss.

All downstream applications were redirected to the new system during the final cutover. We also tested the downstream applications to verify that they were able to connect and retrieve data from the new warehouse. While performing the final cutover, after a final verification, we halted the services on the old legacy systems to prevent any applications from connecting to them. After a period of 48 hours, we began the end-of-life process for the legacy databases.

Business Continuity and Disaster Recovery Considerations

Compared to our legacy databases, the modern solution has two additional features that enable us to quickly access or restore the data when needed:

- The centralized data warehouse, which contains every factory's data, serves as the disaster recovery database for the site-specific warehouses. If issues arise at one of these warehouses, we automatically route the queries to the central warehouse, with only a slight penalty with respect to latency and performance (due to WAN bandwidth and the large size of the central database).
- We archive the text files that are generated during the archive, retrieve, and purge process. These files can be restored very quickly to the original database and tables.

Keys to a Scalable Manufacturing Data Warehouse

As we developed and tested our centralized data warehouse solution, we discovered several best practices that enabled us to meet our goals, such as fast query completion and seamless scalability. Since the solution went into production in 2013, we've grown from handling 2-3 billion sensor data points per day per factory to 6 billion data points per day per factory. When something new comes along, like a new manufacturing process technology or a new Intel Foundry customer, we don't have to build a brand-new solution—we just add new data schemas. The following sections discuss our best practices in detail.

Data Denormalization

Data normalization is the process of organizing data in a database to reduce redundancy and improve data integrity, making it easier to manage, store, and retrieve information. This is generally considered a best practice for RDBMS design—but for our unique low-latency requirements, we decided to denormalize certain data. Allowing repeated data supports better query performance and traceability.

However, we also work with manufacturing engineers to decide exactly what data needs to be denormalized, and what data can be archived in a normalized, compressed fashion (see the "Intelligent Compression and Data Lifecycle Management" section later in this paper. For example, we don't keep every single data point from the SPC system in the denormalized dataset. Instead, we keep only the value-added data, such as the time it arrives on a factory tool and the time it leaves the tool.

Keeping our denormalized database as lean and sleek as possible requires constant teamwork between Intel IT and the engineers. We jokingly refer to this as putting the engineers on a "data diet." For example, engineers may gather huge amounts of data when they design and develop a new process technology. During the design and development phases, this large dataset is necessary. However, once the process technology has been validated, the team works together to remove unnecessary data from the data warehouse. Every time we are asked to add new data, we use this methodology to verify the data will actually be used.

Flexible, Connected Data Model

Because data entering the warehouse can come from many systems—such as MES, SPC, tool sensor data, and workflow orchestrator—we have developed an extensive data dictionary and defined primary keys. We stamp data with IDs such as lot number, tool ID, time range, and process step to enable us to join tables across systems.

We've created more than 200 in-house ETL pipelines for seamless data integration. In this way, the business logic exists before the data is loaded into the warehouse. Data transformation, denormalization, and aggregation enable efficient reporting and analytics. Our ETL pipelines are capable of processing 100 GB every 15 minutes.

A query that joins 11 datasets into a single table used to take 30 minutes to complete. Now it takes only three seconds—a 100x performance improvement.[1]

Resilient Data Pipeline

In the HVM environment, data is being generated every second. Our data warehouse includes a data ingestion orchestrator with multiple threads so we can handle millions of transactions per minute. We perform query performance tuning for diverse workloads to optimize each query type response (batch or real-time). Tuning queries that are used often or are highly valued by operations can save immense time for our operations team.

We take multiple approaches to optimize query performance. Sometimes all we do is rewrite the query so that it uses a more efficient join. Other times, due to the cardinality of the data, we may need to add extra hints to guide the optimizer to choose the right path to retrieve the data. If the business logic is too complex for a single query to complete quickly, we convert the result of the query into a materialized dataset that can then be queried more efficiently.

To further streamline query performance, we use an ingestion process called entity resolution (ER) to reconcile partial ID information into comprehensive and consistent ID data for each chip.[2] The matching algorithm used for ER identifies and reports any conflicting IDs and uses approximate matching logic to reconcile differences in values. To understand how this works, think of a shipping company that must resolve minor variations in street addresses, like "2212 Pinetree Avenue" and "2212 PINE AVE." The approximate matching logic looks for clues and uses rules to reconcile the data.

Intelligent Compression and Data Lifecycle Management

We must always strive for a balance of data accessibility and storage growth. Our approach is to use different levels of compression for different ages of data. Data retention is determined by data domain, as well as if the data is likely to be needed; for example, if a wafer or lot is still active even after six months, we'll keep that data. Data about factory tools that are still in use is also retained. Here are our general compression and retention rules:

- Latest six months of data: No compression is used, because this data is frequently updated and is often queried for various reports and analytic workloads. Compressing this regularly used data would engender a high performance cost that could significantly slow response time.
- Next oldest six months of data: Low compression is used, because although this data is not frequently updated, it is still used to generate quarterly and semi-yearly trends as well as for process yield comparisons. Low compression can compress a million records to 300,000 records. The performance cost induced by low compression is acceptable at this data tier, because near-real-time response it not necessary.

[1] Based on internal Intel IT measurements and observations.

[2] Entity resolution is the process of deciphering whether multiple records are referencing the same real-world thing, such as a person, organization, address, phone number, bank account, or device. Source: https://www.quantexa.com/entity-resolution

- Data older than 12 months: If the data needs to be stored in the warehouse, it is kept under high compression for two years. For data that is not needed, we simply archive it and purge it from the warehouse, but keep the data offline as files that can be restored if necessary. For this infrequently used data tier, the performance cost of high compression is not a concern.
- The last seven years of data are stored in archive datasets, which can be quickly restored on demand for auditing purposes.

Most compression is done at the partition level, based on a work week. We generally do not compress non-partitioned tables, using common sense and archival strategies to control data growth and maintain optimal query performance.

Cost Optimization

We optimize costs through the following means:

- **Workload prioritization** for critical processes minimizes the cost of keeping important workloads idle. We create resource groups for critical, high, medium, and low criticality workloads, and assign the resource groups to user profiles. Our data warehouse includes an effective resource manager that tracks when certain parameters cross predefined thresholds. Examples of these parameters include the number of parallel executions, amount of physical I/O (data retrieved from disk), logical I/O (data retrieved from memory or cache), CPU usage, and elapsed time. When a user workload is executed and crosses a threshold for that resource group, the resource manager automatically downgrades the workload to a lower resource group.
- **Compression and encryption** hardware accelerators reduce storage overhead (see the "Intelligent Compression and Data Lifecycle Management" section earlier in this paper).
- **Dynamic scaling strategies** efficiently allocate resources like database services and memory. Our data warehouse runs on a cluster with multiple compute nodes. There is a logical split of services (sometimes called gateways) through which a workload gets assigned to one of the compute nodes. Some services run on a specific node and some across multiple nodes. Depending on the time window and need, services can be automatically shut off or expanded to run on one or more nodes. Also, we can adjust the allocated memory thresholds for the user and global cache as needed to accommodate intensive queries. We can use information from our resource manager to gain insights into workload demands so we can dynamically scale resources and avoid overprovisioning.

For cloud-based data warehouses, the focus is often on controlling cloud costs by optimizing resource utilization. Although our data warehouse is on-premises, we also strive for efficient resource allocation. We have built several dashboards and metrics to monitor resource usage, the number of query executions, the number of distinct concurrent users, CPU usage, and so on. This information can help us tune our resource manager to allocate system resources and improve overall system efficiency. We also achieved hardware cost optimization when we launched the data warehouse, because we consolidated multiple legacy databases into a single warehouse, which in turn reduced hardware and license costs.

Process-Specific Security (PSS) and Data Protection

A company's manufacturing data and intellectual property are crucial aspects of maintaining competitiveness. Whether it's Intel's own data or that of an Intel Foundry customer, we strive to provide the highest level of security possible. That means allowing only authorized users to access relevant data. However, in the HVM environment, security can interfere with query performance due to the explosion of security code combinations and performance overhead from access control enforcement.

To address these challenges, we have implemented PSS using a virtual private database feature of our data warehouse. This establishes data access control at the row level by dynamically modifying SQL queries to enforce access controls. PSS helps prevent unauthorized data access while still delivering efficient query execution.

Key components of our implementation of PSS include the following:

- Bit-masking techniques to manage security codes efficiently (also referred to as bitwise operations for optimized security filtering).
- Centralized security code management to reduce overhead.
- Integration with existing authentication frameworks for seamless user access.
- Logon triggers and session-based access control. Every user has a profile that defines a set of one or more process nodes that they are approved to view; this profile is applied when the user logs onto the system and remains in effect for the duration of the user session. Once the session expires or the user logs off, the profile is removed, recalculated, and applied the next time the user logs in. This helps to ensure that if users have been approved for additional processes or have been restricted from seeing process data that they may have had access to before, we do not violate the security constraints.

Results

As mentioned earlier, our Intel architecture-based system and data warehouse best practices combine to deliver excellent performance, such as reducing query completion time from 30 minutes to three seconds. The solution also enables our process engineers to quickly perform root-cause analysis and make fast decisions to improve product quality and factory efficacy. Real-time data can improve inline and end-of-line yield, speed fault detection, and identify real-time trends of equipment and process parameters.

Our scalable, robust data warehouse solution helps lower total cost of ownership because it requires less maintenance than the older systems, significantly reduces unplanned downtime, and accelerates factory problem-solving. Increased performance, compared to the older systems, enabled us to pursue modern smart factory solutions like factory digital twins and AI. We are currently conducting a pilot project with Intel® Factory Pathfinder (part of the Intel® Automated Factory Solutions Software Suite), which is a discreet event simulation engine that utilizes

data to perform what-if simulations to solve a multitude of issues in operations, such as identifying bottlenecks, optimizing tool utilization, and improving cost initiatives.

Next Steps

Future advancements that will further reshape the manufacturing data landscape include:

- **Edge computing.** Increasingly advanced equipment and IoT sensor capabilities are creating more and more data. Using PC-based industrial control systems, Intel's manufacturing engineers can use this new data to help make real-time, informed decisions about product and process variations to further help improve product yields and equipment availability. Edge computing can also help reduce manufacturing costs and manufacturing productivity.
- **Federated learning.** Intel's factories use AI, like computer vision models for defect detection, to improve manufacturing efficiency. The more data an AI model has, the more accurate it becomes. However, the foundry model enjoins strict data privacy and security standards on sharing data. Federated learning will allow multiple Foundry customers to train a model within their trust boundaries and private datasets. The partial models can then be centrally aggregated, without exposing data, to achieve higher model accuracy.
- **AI-powered decision systems.** Similar to edge computing, AI tools and applications will process all of the data to provide decision makers with the best information to help expand yields and predict tool downtimes.

Conclusion

The transition from fragmented data systems to a unified, centralized, and high-performance data warehouse allows Intel's factories to manage data complexity at scale. The scalable solution delivers full traceability, meaningful real-time insights, and better decision-making, along with strong security. We have discovered that deploying the solution on modern Intel architecture provides the performance manufacturing engineers need to perform their jobs and keep the factories running as efficiently as possible.

As Intel's factories evolve, we will continue to pursue improvements in security, scalability, and cost optimization. We hope that by sharing our best practices and results, we can help other manufacturers achieve similar success on their manufacturing big data journeys.

Related Content

If you liked this paper, you may also be interested in these related stories:

- IT@Intel: Minimizing Manufacturing Data Management Costs
- IT@Intel: Delivering Operational Efficiencies Using a 5G Private Network
- IT@Intel: Using Natural Language Processing to Streamline Manufacturing Failure Mode and Effects Analysis
- IT@Intel: Intel Cuts Downtime and Costs with Fault Detection Systems for Factory Equipment
- IT@Intel: Smart Manufacturing Using Computer Vision and AI for Inline Inspection
- IT@Intel: Reliability Engineering Helps Intel Cut IT Manufacturing Systems Downtime in Half
- IT@Intel: Accelerated Analytics Drives Breakthroughs in Factory Equipment Availability
- IT@Intel: Master Data – Managed!

For more information on Intel IT best practices, visit intel.com/IT.

IT@Intel

We connect IT professionals with their IT peers inside Intel. Our IT department solves some of today's most demanding and complex technology issues, and we want to share these lessons directly with our fellow IT professionals in an open peer-to-peer forum.

Our goal is simple: improve efficiency throughout the organization and enhance the business value of IT investments.

Follow us and join the conversation on **X** or **LinkedIn**.

Visit us today at **intel.com/IT** or contact your local Intel representative if you would like to learn more.