White Paper
IT Infrastructure
May 2025

IT@Intel: Modernizing Enterprise Data Analytics Using Databricks in the Cloud

Intel IT migrated data and analytics workloads to the cloud, enabling a connected data foundation; faster turnaround of business requests and data loads; and increased operational efficiency

Intel IT Authors

Pradeep Baluvaneralu
Cloud Data Warehouse Architect, Data and Analytics

Eric Messenger
IT Director, Data and Analytics

Table of Contents

Executive Summary

To address systemic data and analytics challenges across the organization, Intel IT designed and executed the migration of one of Intel's largest business unit's (BU) enterprise data analytics to the cloud. This highly successful initial project paves the way for additional migration and acceleration of the modernization of Intel's data platform.

Creating a highly efficient, accurate, and high-performing data platform is especially important as Intel expands foundry services through its IDM 2.0 strategy and builds its software business. Data—from sales, supply chain, planning, design, and other business activities—is foundational for Intel's BUs to make intelligent decisions at the speed of business.

The new cloud-based enterprise data analytics platform is based on Databricks combined with a cloud data warehouse. The new platform includes the following aspects:

- Sandbox capabilities to combine enterprise data and BU data
- Reusable data ingestion and transformation templates
- Streamlined end-to-end artificial intelligence and machine-learning workflows
- Dynamic scalability and high performance

The platform runs on Intel® Xeon® Scalable processors and benefits from Databricks' optimized data transformation capabilities. It also poses intriguing possibilities related to machine learning and generative AI.

For this single BU, the cloud migration has resulted in a 65% reduction in extract-transform-load (ETL) process time, a 58% reduction in job failures, and a 50% reduction in user issues. We will use this initial project's key learnings and best practices to accelerate Intel's transformation as we migrate additional BUs' enterprise data analytics to the cloud.

Contributors

Sachin Arora, Data Platforms Engineering Manager and Product Owner
Giri Bagusetti, Cloud Application Development Engineer
Venkadeshwaran Karunakaran, Data Platforms Cloud Software Engineer
Rajeshkumar Ramamurthy, Data and Analytics, Principal Engineer

Acronyms

BU	business unit
CIF	Cloud Ingestion Framework
ETL	extract-transform-load
GenAI	generative AI
IAO	IDM 2.0 Acceleration Office
IDM	integrated device manufacturer

Business Challenge

Intel is a data-driven organization, and Intel's business units (BUs) make critical decisions only after carefully analyzing data from different perspectives. Intel IT sought a modern data and analytics solution that could address the following key business challenges that span Intel's BUs and also eliminate functional gaps in the current solution.

- **Connected 360-degree view of the data.** To obtain maximum value from their data, BUs need a connected, 360-degree view of all their datasets. Because data requirements constantly evolve, the BUs want to frequently integrate new datasets into the data lake and query all the data via a connected foundation to obtain key insights. Currently, most BUs' datasets are dispersed across many platforms and solutions, presenting a challenge in achieving the holistic, connected data view that the business needs.
- **Faster turnaround of business data requests.** Faster delivery of data provides stronger value for decision makers. BUs can't wait a long time for new datasets to be integrated into the data lake. They need their dashboard to be quickly updated with newer datasets so they can quickly derive insights and react swiftly to changes in their business environment. However, dataset integrations consume a significant amount of development hours due to custom code and the multiple skill sets required for data pipeline development. What's more, data science use cases sprawl across platforms and siloed datasets, resulting in low productivity for data scientists, who subsequently spend more time on data preparation than data analysis.
- **Timely data availability and faster insight.** Dataset query jobs fail intermittently for various reasons and query responses are not fast enough for end users. Dashboards are not updated as quickly as the BU requires, and there are delays and synchronization issues between regions. Workloads cannot be split between read and write clusters; hence, data engineers spend significant time manually fine-tuning data platforms to achieve the necessary performance levels.
- **Ease of integration and self-service capabilities.** Business users need access to more data platform features and capabilities. Vendors' new data analytics capabilities are typically introduced in the cloud first, while on-premises adoption lags. BUs are looking for simplified data integration patterns without custom code and additional configuration. They also want extensive self-service capabilities to merge enterprise data with BUs' data without IT intervention.
- **Limited elasticity results in high capital costs.** BUs tend to overprovision on-premises systems to prepare for workload spikes. This drives up capital expenditure as well as support costs.

These challenges presented a larger IT governance and architectural problem: many data marts and containers, combined with considerable technical debt. To successfully continue our modernization efforts for enterprise data analytics, we needed to provide a unified and cost-efficient platform that could:

- Support large data volumes and handle both structured and unstructured data storage and analytics.
- Automatically and flexibly scale based on workload demand.
- Provide low-code or no-code solutions.
- Deliver disaster-recovery capabilities.
- Reduce significant downtime and effort for upgrades.

Solution Overview: Moving Enterprise Data Analytics to the Cloud

We chose Databricks running in the public cloud on Intel® architecture combined with a cloud data warehouse for more advanced analytics. In Databricks, we build the data pipeline, land the raw data on the cloud storage, transform the data, and create the necessary tables for analytics (such as fact and dimension tables). The Databricks SQL Warehouse is also used for querying raw and curated data for ad hoc analysis.

For a more detailed discussion of the solution components, see the "Solution Architecture" section.

What Is Databricks?

Databricks offers a cloud-native, unified Data Intelligence platform for Data and AI. The Data Intelligence Platform infuses Generative AI capabilities into their lakehouse platform, which combines the best of data warehouses and data lakes. Databricks includes features such as data warehousing, data engineering, data science, and business analytics.

Databricks was founded in 2013 by the original creators of Apache Spark and provides an alternative to the MapReduce system. In addition to Apache Spark, Databricks also offers a number of other open-source projects, including Delta Lake, MLflow, and Unity Catalog.

Migration by Numbers

100+	4,000+	250+	500+	100+	1,000+	3,000+
Data Sources	Tables	TB Data	Pipelines	Cubes	Dashboards	Users

Migration Timeline

As with any substantial change in the IT environment, revitalizing our enterprise data analytics platform involved a tight blend of technology, people, and processes. We chose to migrate one BU's enterprise data warehouse to the cloud for improved analytics accuracy, acceleration and operational efficiencies. This initial project would prove the value of moving to the cloud and establish a blueprint for migrating other BUs over time. As Intel ramps up its Foundry Services and continues to evolve as a software and hardware company, we can use this cloud migration blueprint to speedily multiply our success. Figure 1 illustrates our migration timeline for the initial BU.

Toward a Connected Enterprise

Since Intel's founding in 1968, it has been an integrated device manufacturer (IDM), which means that it designs and builds its own semiconductor chips. In March 2021, Intel's CEO introduced "IDM 2.0," a major evolution of that strategy that includes significant manufacturing expansions, plans for Intel to become a major provider of foundry capacity in the U.S. and Europe to serve customers globally, and expansion of Intel's use of external foundries for some of its products.

Intel's new Foundry Business and the move to IDM 2.0 introduce an inflection point that will significantly increase data volumes and the need for connected data and speedy, accurate analytics. To realize the full potential of IDM 2.0, Intel IT must continue to transform our systems, processes and tools to significantly improve scalability and transactional integrity. To that end, we have formed the IDM 2.0 Acceleration Office (IAO), which will collaborate closely with all BUs and functional teams to bring this new internal foundry model to life. The migration of enterprise data analytics to the cloud is an important aspect of the IAO's efforts because we can scale the solution to connect the entire enterprise.

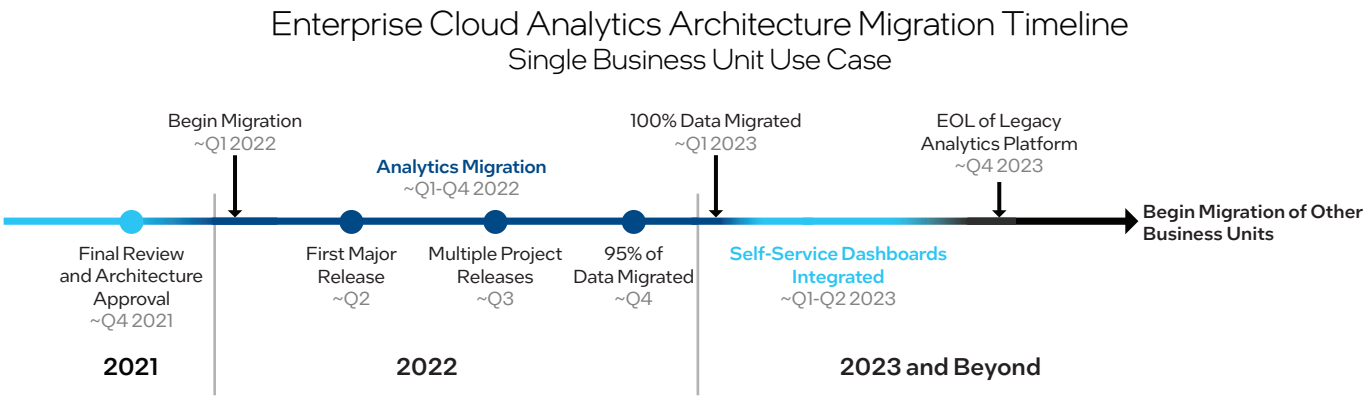


Figure 1. We used a phased approach to complete the migration of the BU's enterprise data analytics to the cloud over the course of about four quarters.

Solution Architecture

Figure 2 illustrates our cloud-based enterprise data analytics architecture. Data is pulled from a variety of sources and stored in the cloud storage system. Databricks transforms the raw data into curated data, which is then copied to the cloud data warehouse. Databricks provides essential tools for implementing and orchestrating data transformations and ETL on the raw data. In particular, the following Databricks capabilities help create reliable data that the BU can use:

- **Delta Lake.** This is the optimized storage layer that provides the foundation for storing data and tables in the Databricks lakehouse. Delta Lake is open-source software that extends Parquet data files with a file-based transaction log for ACID[2] transactions and scalable metadata handling. Delta Lake is fully compatible with Apache Spark APIs, and was developed for tight integration with Structured Streaming to enable the use of a single copy of data for both batch and streaming operations. Delta Lake also provides incremental processing at scale. Unless otherwise specified, all tables on Databricks are Delta tables.[3]
- **Improved Spark processing engine.** Databricks offers Photon, which supports multiple languages such as Python, Scala, and SQL. It is a high-performance Databricks-native vectorized query engine that runs SQL workloads and DataFrame API calls faster, helping to reduce total cost per workload. The acceleration offered by Photon was a primary reason for choosing Databricks as the foundation of our new cloud-based enterprise analytics platform.
- **Isolated clusters.** Databricks architecture allows us to establish isolated cluster environments, enabling the configuration of multiple read/write clusters that are tailored to specific requirements. This not only boosted performance but also mitigated the overhead linked to performance enhancements.

The BU's users have two options when they want to query the data. They can use Databricks' SQL endpoint to perform ad hoc queries against the raw data. Or they can use their data science platform to query the curated data in the cloud data warehouse. The results of these queries are sent to the relevant application or dashboard.

[1] ACID is an acronym for four transaction characteristics: atomicity, consistency, isolation, and durability.

[2] Source: https://docs.databricks.com/en/delta/index.html

What Are SSAS Cubes?

A simple SQL query returns one or more rows of data. More complex online analytical processing uses SQL Server Analysis Services (SSAS)[3] cubes (also referred to as in-memory cubes). This is a multidimensional reporting capability that "makes it possible to aggregate values and summary reports on multiple axes and provide a more detailed analysis by performing grouping of data along with more than one column and creating multiple grouping sets while using just a single query."[4]

While SSAS cubes are useful, they can take considerable time to complete. **Restructuring the data platform in the cloud using Databricks enabled us to reduce cube refresh time by up to 33%.**

[3] For more information on SSAS, see learn.microsoft.com/en-us/analysis-services/tabular-models/tabular-models-ssas?view=asallproducts-allversions.

[4] Source: www.educba.com/cube-in-sql/

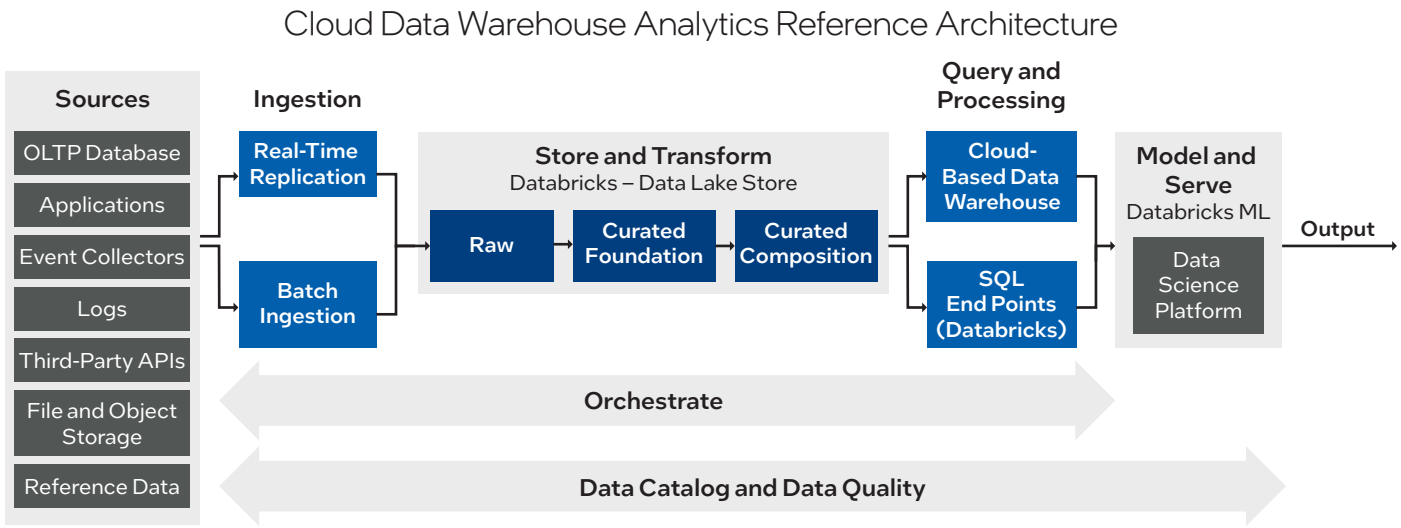


Figure 2. From data ingestion to reports and dashboards, our cloud-based enterprise data analytics platform using Databricks enables faster data ingestion and transformation as well as self-service access to the data warehouse.

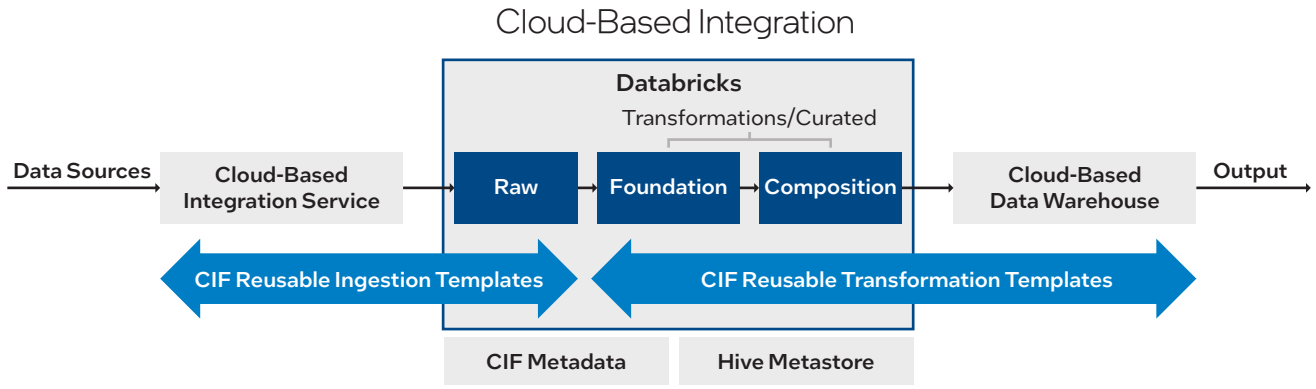


Figure 3. Our Cloud Ingestion Framework (CIF) uses reusable data ingestion and transformation templates to speed up dataset onboarding and analysis.

Cloud Ingestion Framework

We developed a metadata-driven Cloud Ingestion Framework (CIF), which includes reusable templates, including data ingestion templates for various databases, SFTP, S3, file share, and several software-as-a-service solutions. We also developed a transformation template. Figure 3 illustrates the CIF.

Databricks Cluster Configuration

We chose a cloud instance using Intel® Xeon® Scalable processors, which provide multiple accelerators for data streaming, AI, cryptography, data compression/decompression and other functions. The instance we chose delivers high memory capacity, which is important for fast analytics. The instance also delivers low latency and high-speed local storage.

We created multiple Databricks workspaces based on the data domain to carry out data engineering and ETL activities. Each workspace cluster is set up with one to five nodes in autoscaling mode. Each worker node has eight cores and 64 GB of memory. We mainly use drivers that are the same size as the executor. Each cluster is set to auto-terminate after 10 minutes of inactivity. We currently use Databricks Runtime version 11.3.x-scala2.12 and are in the process of upgrading to the latest version of Databricks Runtime.

Business Value of Cloud-Based Enterprise Analytics

Overall, our cloud migration improved our data architecture and visibility, while powerful Intel® processors contributed to performance improvements. All of the BU's data is now on a very scalable platform. Figure 4 summarizes the benefits of moving enterprise data analytics to the cloud; these benefits are discussed in more detail in the next section.

Data Engineering and Pipeline Improvements

Databricks' optimized storage, called Delta Lake, makes it easy to work with the data, helps remove boilerplate code, and provides far more efficient extract-transform-load (ETL) operations. Cloud infrastructure can scale out/in on demand based on the workload, which helps reduce the end-to-end run time of data pipelines. We can now take advantage of reusable templates for data ingestion and transformation (see the "Cloud Ingestion Framework" section for more details). The cloud data warehouse provides a single source of truth, helping to ensure consistent use and interpretation of data.

We have reduced technical debt by end-of-lifeing (EOL) several older data platforms, data marts and projects. We also achieved a 16% object reduction (nine major data sources and multiple views) across projects during the migration, based on usage analysis and system re-design.

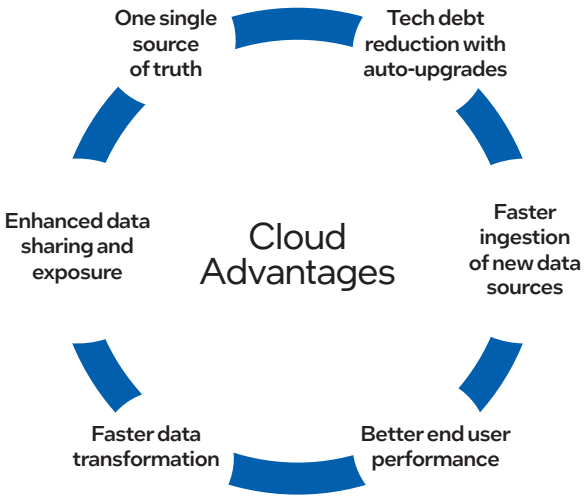


Figure 4. Moving enterprise data analytics to the cloud has several advantages.

"Transitioning to Databricks has enabled us to have direct access to the data we need. It has saved us a lot of data engineering overhead so we can focus on driving value through analytics. Porting the existing models and codebase to the platform was straightforward, and we are already taking advantage of the Azure compute clusters to significantly reduce processing time. Even so, we are just getting started, and excited to start taking advantage of the DevOps and other features as we build experience."

– Jason Cauthen, Data Science Director, Sales and Marketing Group

Connected Datasets for Better Visibility

Migrating the BU's enterprise data analytics to the cloud gave us an opportunity to organize the data lake and data warehouse layers to enable easier analytics and self-service. The data lake is designed according to the medallion architecture[5] and various connected data products are built with common dimensions that serve specific business requests. We created data model documentation that provides visibility into the relationship between datasets and assists in exploratory analysis and self-service dashboarding. The BU's users can directly query the data warehouse through business intelligence tools and connect datasets using the data model as a reference. Consistent attribute names and accurately defined data types, lengths, and relationships enhance performance and usability.

New Capabilities

Cloud service providers are often the first to make new features available for general use. Our adoption of the cloud for enterprise data analytics poses exciting new opportunities through cloud-based machine learning and integration of generative AI (GenAI) with data in the future. Intel IT is actively exploring use cases for GenAI and large language models (LLMs). Databricks is key to providing quality data to train LLMs or for retrieval-augmented generation. Our cloud-based enterprise data analytics platform positions us well for emerging demands for data.

The cloud-based solution also democratizes analytics through self-service portals and frees data scientists from the data pipelines business. They can now focus on the analytics themselves, helping the BU make better data-driven decisions.

Performance Improvements

The Databricks lakehouse platform combines the best of a data lake's openness, scalability, and flexibility with the best of a data warehouse's reliability, governance, and performance. To demonstrate the benefits of using Databricks, its native vectorized query engine (Photon), and Intel Xeon processors, we ran a test derived from the industry-standard TPC-DS power test on cloud instances.[6] The baseline for the test was 1st Gen Intel Xeon Scalable processors on the worker nodes running Databricks 10.3 without Photon enabled.

We then ran the same workload—with no code changes—on 3rd Gen Intel Xeon Scalable processors, which include Intel® Advanced Vector Extensions 512 (Intel® AVX-512), on the same version of Databricks with Photon enabled.

By enabling Databricks Photon and using Intel's 3rd Gen Xeon Scalable processors, without making any code modifications, we saved two thirds of the costs on our TPC-DS benchmark at 10 TB and the benchmark ran 6.7x quicker.[7] This translates not only to cost savings but also reduced time-to-insight. We are experiencing similar benefits to our deployed production Databricks environment.

Results

Migrating enterprise data analytics to the cloud has resulted in faster insights, reduced data processing time, less effort creating and managing data pipeline (thanks to the low code/no code platform) and improved operational efficiencies. For example, a 65% reduction in ETL processing time translates to delivering data to the end users faster, enabling them to make faster decisions.

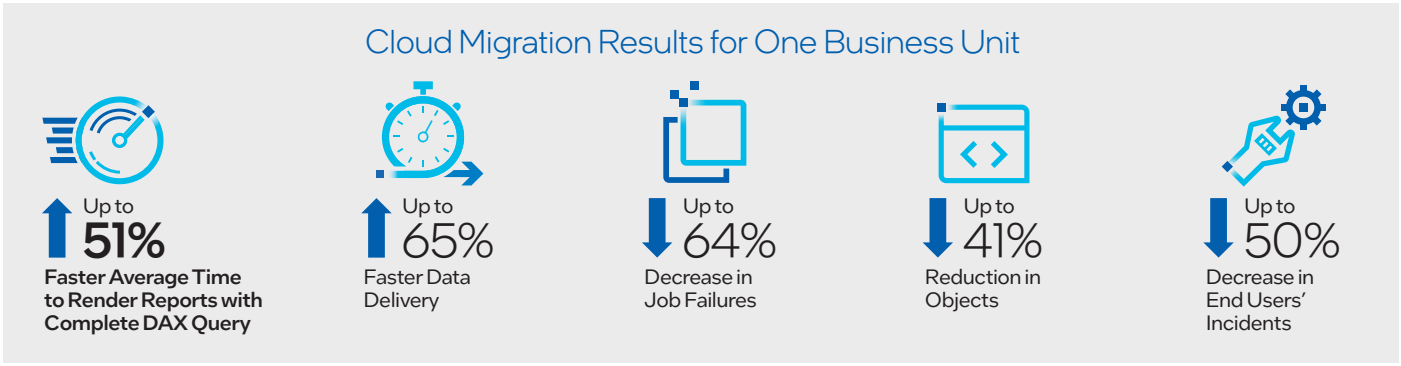
The BU now has a unified, connected data platform that provides a 360-degree view of its data, and business requests are turned around end-to-end at least 3x faster. The cloud data lake enables a "bring your own data" approach, giving the BU the flexibility to combine enterprise data with business data for prototyping and improved business agility.

The BU can also access new data platform capabilities more quickly than they could on-premises and can take advantage of the cost benefits of automatic and elastic scaling. Data engineers can spend less time on support and maintenance, and instead focus on serving the business need.

[5] "A medallion architecture is a data design pattern used to logically organize data in a lakehouse, with the goal of incrementally and progressively improving the structure and quality of data as it flows through each layer of the architecture (from Bronze ➔ Silver ➔ Gold layer tables). Medallion architectures are sometimes also referred to as "multi-hop" architectures." (source: https://www.databricks.com/glossary/medallion-architecture.)

[6] Derived from the power test consisting of all 99 TPC-DS queries that ran in sequential order within a single stream. The results are not comparable to an official, audited TPC benchmark.

[7] Test by Intel and Databricks, March 2022. See https://www.databricks.com/blog/2022/05/17/reduce-time-to-decision-with-the-databricks-lakehouse-platform-and-latest-intel-3rd-gen-xeon-scalable-processors.html for more details, including full testing configurations.



Migration Best Practices

We performed the BU's enterprise data analytics migration to the cloud using 14 data engineering teams—a total of 200 people over four quarters. Our basic recipe for success is as follows: Start small (such as one self-contained data domain), learn, generate success, create the playbook, and grow. The following sections detail several best practices that we developed to help make future similar migrations go even more smoothly.

Core Strategies for Planning and Design

- **Get buy-in.** Even in just a single BU, the data was consumed by many users, and we wanted to get them excited about the migration. To do that, we showed the BUs why the migration was needed. We brought in business leaders, collected their pain points, and then demonstrated how our cloud migration strategy would solve their concerns. The data consumers needed to understand that this wasn't just a random, IT-centric idea about shiny new technology; rather, it was a transformative effort that would benefit the BU directly and, eventually, the entire company. With business alignment in hand and a well-documented list of benefits, we also worked with IT leadership to establish the architecture as the agreed-upon enterprise architecture. With support from IT and BU leadership, we were ready to start the migration.
- **Perform design work up front.** Because we were taking a phased approach—moving only some of the data at first, followed by additional datasets—we established a bridge between certain systems so that BU users' work would not be disrupted. We also designed the data model and the CIF before any data migrated, and collaborated with the data platform teams to design and set up cloud-based resources. We also developed data products on a central, connected data foundation. Here are the steps we followed:
  1. **Analyze.** To begin the actual migration, we performed an assessment that enabled us to build an inventory of existing objects. This gave us an idea of the overall objects required for migration and the size of the migration.
  2. **Redesign and refactor.** To maximize the value of the migration, we didn't just "lift and shift" the existing data architecture and objects as-is. We also focused on solving critical business problems like lack of connected data and slow query response times as part of the migration. In this redesign/refactoring phase, we identified the overall design required for migration (such as solution and data architecture, data model, and data archival requirements). We also identified the objects and solutions that were no longer needed and EOL'ed them before starting the migration, which helped reduce scope and cost.
  3. **Rebuild.** We rebuilt the data ingestion pipeline by developing the CIF to improve data platform velocity. We didn't copy the raw data from the existing data warehouse, but instead rebuilt ingestion by pointing to the source system.
  4. **Migrate.** With all the previous steps completed, we were ready to write the code that we had designed in the redesign/refactor phase and then finalized the migration to the cloud environment.
- **Collaborate with data users.** When it was time to choose the cloud instance, we needed real-world information about how much data the BU used, data analysis patterns, and latency expectations. We also built consensus among users about data retention and data archival requirements. For example, for some data, perhaps it did not need to be kept for seven years—two would suffice. This information informed us about our choices of processors, storage, and memory. Our goal was to find the sweet spot that balanced service-level agreements versus cost versus necessary resources.

Documentation

- **Track progress and dependencies.** We built an object inventory, complete with all dependencies. This enabled us to measure migration progress weekly based on the number of migrated objects.
- **Document everything.** We kept detailed records of all data that was moved. We also formulated interim timelines and milestones based on estimates of how long we expected each aspect of the migration to take.
- **Prove value.** We also documented the positive impact of the migration, keeping metrics to demonstrate success and business value.

Expedited Migration

- **Limit change as much as possible.** Once we decided on the data model, CIF templates, and so on, we discouraged the BU's users from making changes during the migration process. When we limited change, we met our deadlines perfectly; if change crept in, the schedule was delayed.
- **Continue to collaborate with the BU after the migration is complete.** Although the pilot BU migration project is finished, we continue to work with them to plan for continuous integration/continuous delivery of new capabilities. It was key to migrate the core features first, but we understand that business analytics requirements continue to evolve, and cloud providers continue to introduce new features.

Training

- **Train the organization in parallel with migration.** A new data platform with new tools can intimidate users who are steeped in the "old way" of doing things. We found it best to offer training on the new platform during the quarter prior each migration phase. We sidestepped a fair amount of potential procrastination by setting a firm limit on design deadlines and a hard-and-fast migration date.

Conclusion

Fast and accurate data analytics will be key to Intel's BUs' success in the IDM 2.0 era. Based on the success of our initial migration of a BU's enterprise data analytics to Databricks and the cloud, we plan to move our master data to the cloud as well. Logical connectivity relies on healthy master data. For more about master data and a single source of truth, read the IT@Intel paper, "Master Data—Managed!"

Other future data platform modernization opportunities include using our learnings and best practices to accelerate IAO migration as well as migration of additional BUs' enterprise data analytics. We also plan to explore the application of GenAI to our Databricks and cloud data warehouse architecture.

Related Content

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

- Master Data – Managed!
- Enterprise Technical Debt Strategy and Framework

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** if you would like to learn more.