Microsoft tames the “wild west” of big data with modern data management

In an era of information overload, big data stories are everywhere. From battling cancer to improving traffic flow, big data paired with machine learning and AI are stealing the show. It’s true that data is now one of a company’s most strategic assets, because it’s the foundation for all the advanced analytics and deep learning that drive business transformation.

Data, once relegated to individual business groups for their internal business applications, now needs a carefully planned, modernized data infrastructure to deliver on the promise of deep, transformative insights. For Microsoft and across the industry, the transition to a fully modernized data management framework is a key factor in overall business success.

In Microsoft Core Services Engineering and Operations (CSEO), we saw an opportunity to obtain rich insights that can emerge when data crosses organization boundaries. To get there, we needed to modernize our big data strategy to provide a 360-degree view of the business and enable predictive and prescriptive analytical capabilities.

Only by taming the “wild west” of big data can we provision consistently reliable access to the advanced machine learning and AI techniques needed to power Microsoft through our digital transformation.

Challenge is people, not technology

The most difficult part of integrating data across a large organization is not technical—it’s people and organizations. We divide CSEO into functional business areas like HR, finance, and supply chain, with SAP at its core.

We reached out to each business area, bringing together experienced people to collaborate around a standardized approach to handling big data.

Soon after, we formed a cross-functional team named the Data & Analytics Working Group (DAWG), chartered with building a common set of standards, processes, patterns, and tooling for data management and governance of CSEO data assets. The DAWG team began with three overarching mandates:

- Lower data management costs and risks by reducing the duplication of data across systems.
- Improve decision making by reducing the risk of basing decisions on false or inaccurate data.
- Provide a modern data architecture that enables deeper predictive and prescriptive insights.

This team was our way of working across CSEO to build consensus around a standardized approach to big data. Knowing it wouldn’t be easy, we embraced a pragmatic “crawl, walk, run” approach to enacting these changes, as shown in Figure 1. It started by the DAWG team agreeing on a core set of draft tenets, governance, and security requirements and then piloting them with one of our business groups.

**Figure 1. Forging consensus for a new data management framework**

We took the following approach:

- **Phase 0.** Obtain sponsorship from senior engineering leaders and the executive leadership team for data and analytics tenets, standards, and recommendations from the DAWG team.
- **Phase 1.** Shift the culture to include data and analytics design principles, best practices, and standards to ensure high-quality, efficient, and highly scalable applications, services, and platforms.
- **Phase 2.** Invest in missing data management services and scalable platform design.
- **Phase 3.** Align across CSEO by tracking CSEO adoption and alignment via a measurement tracking framework (including quarterly targets).

### The DAWG mission

With an eye to curating data to provide business benefits to our customers, our mission started with the goal of implementing a fully modernized data environment. Specifically, we knew we had to enable our engineers, analysts, and data scientists to find “source-of-truth” data and apply machine learning techniques. “This requires that we have design principles for right-time data for decision making, efficiency, reduced cost, and data handling principles to ensure compliance, usability, and proper stewardship for data and issue handling,” said principal software engineer Robert Venable, a member of the DAWG team.

To build on that mission, we developed core tenets outlining our key priorities and design recommendations.

**Table 1. Data management core tenets**

<table>
<thead>
<tr>
<th>Tenet</th>
<th>Motivation</th>
<th>Actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Separate storage and compute.</td>
<td>Cost and latency</td>
<td>Keep the data in place, and move the compute to the data, minimizing data movement and duplication.</td>
</tr>
<tr>
<td>Elastic scale of storage and compute separately.</td>
<td>Cost and agility</td>
<td>Design for scale-out/up and scale-back/down solutions that provide the computation power we need when we need it.</td>
</tr>
<tr>
<td>Enable the modern data warehouse.</td>
<td>Cost, governance, flexibility, and maintenance</td>
<td>Adopt the Modern Data Warehouse Extract-Load-Transform (ELT) model and integrate metadata at load time.</td>
</tr>
<tr>
<td>Create agility by moving to API access.</td>
<td>Business cost, governance, EU data protection regulations, and latency.</td>
<td>Provide API access to data as a primary access method and fall back to batch/bulk as a secondary method.</td>
</tr>
<tr>
<td>All production data is controlled with governance.</td>
<td>Cost, risk, and governance</td>
<td>Define standards of data stewardship and responsibility. Define security, access, and audit requirements.</td>
</tr>
<tr>
<td>Discover, consume, and publish “source of truth” data.</td>
<td>Cost, risk, and governance</td>
<td>Agree on a single metadata repository with certified, published sources of truth, managed by the functional data owners.</td>
</tr>
<tr>
<td>Integrate machine learning as a core component of every new application.</td>
<td>Prescriptive analytics and business cost</td>
<td>Machine learning can help reduce confirmation bias during data analysis and optimize product feedback.</td>
</tr>
<tr>
<td>Deliver the right data at the right time.</td>
<td>Latency, cost, and flexibility</td>
<td>Ensure that new architectures can support required latency.</td>
</tr>
<tr>
<td>Abstract complexity from the consumer.</td>
<td>Flexibility and governance</td>
<td>Use semantic layers and API access to separate business needs from underlying data structures.</td>
</tr>
</tbody>
</table>

### Platform

Choosing a common storage platform was one of our first efforts, which made it easy to share data across organizations while letting each team choose the level of compute they needed.

### Infrastructure

For data storage, we chose Azure Data Lake Store (ADLS) for its flexibility and security capabilities. In addition to managing structured, partially structured, and unstructured data, we can also store files in any format as large as a petabyte (PB).
Crucially, ADLS can store any data in the system’s native format without needing teams to perform transformations on the data.

Additionally, ADLS uses the open-source Hadoop Distributed File System (HDFS) and Yarn to enable any HDFS-compute option to access data. Teams can choose from SQL offerings, Hadoop, Azure Databricks, and a growing number of HD Insight technologies such as Spark and Hive.

Data lakes provide raw data to analysts and data scientists with low latency and offer several key operational benefits. We can store data of any structure and protect downstream systems from breaking due to schema changes in the source systems. Data lakes help us provision the right storage and compute technologies for diverse types of data and version control, while reducing our overall storage costs.

As shown in Figure 2, data lakes ingest data by either traditional batch methods or streaming from an online source. The diagram shows how HR and finance groups use data management services to run analytics on raw data stored in Hadoop or another compute option (R arrow). Groups have the option of publishing their analytics results back to the data store (RW arrow), enabling users to view the results in apps like Power BI.

---

**Figure 2. Provisioning rich data analytics from ingestion to user consumption**

**Data-management services**

Our approach to building data-management services is a shared data platform, with a central team providing consistent services that other teams use as they load and extract data from the data lakes. These services are in development to meet the needs of all the data providers and lake managers, eliminating the duplicative services that inevitably would be developed if we hadn’t taken this approach. In addition, they’ll provide a consistent approach to data lake activities across all the teams and give us a way to implement and enforce standards. They are a win for all parties.

We plan to build three initial sets of services: publishing, monitoring, and then consuming services. We’ve started working on the first two sets, and here are the different services we see within each:

**Publishing services**

- **Handshaking Service.** This is the initial service used by the publisher as they load data into the data lake. It’s a job service that tracks and manages the data transfer from the source to the lake.
Control File Service. After the data is dropped, the control file service goes into action. It creates a JSON metadata file associated with that data, describing the technical information in the data (table name, field names, datatypes, primary keys, etc.). The control file enables engineers to look at the lake, read the JSON files to understand the data, and then read the data files.

Catalog Service. Another service reads the control files in the lakes and automatically populates a central catalog. Without this catalog, the data lakes become a data swamp. With this catalog, service users can find the data they need, and the management team can monitor for duplicate data sets in different lakes.

Merging Service. Data is dropped into the lake in raw form every 30 minutes, including every change made in each record. This is needed for some machine-learning/AI work, and auditing requirements for compliance reasons. However, the status of a record can change many times during a day, and analysts want to see only the latest record at the end of the day. It doesn’t make sense to have every analyst go through all those files and work out the latest record. This service does that for them. The merging process provides single files with hourly, daily, monthly, and annual data so analysts and others go to one place to see the current information.

Data Quality Service. This will initially perform data profiling and tagging, based on the data-quality rules defined later in this paper. This will be our gatekeeper service and will ensure the data that is brought into the data lake has met certain quality bars and standards.

Monitoring services

Standards Service. With a distributed data lake model, we have to ensure that each lake consistently conforms to the defined standards so that they appear as one virtual lake to consumers. This service monitors the physical implementations of each lake to ensure they are compatible with the standards.

Access Service. This monitors the telemetry and usage of data within the lakes, enabling publishers to see who is consuming their data, for security and audit purposes, and analytics on which data is key to the organization.

Data Management. The goal with a Modern Data Architecture is to bring the compute to the data, instead of copying the data everywhere to the compute engines. This requires a change in culture that takes time, and so this service enables us to track users still copying the data out of the lake. Using this information, we initiate discussions with the teams and change our approach to reduce copying as much as possible.

With these central services, we don’t have all our functional teams writing code to do the same thing. This way, we write the services according to standards we’ve developed and enforce those standards through the services. When the standards change, we update the code in one place and all the data lakes automatically pick it up.

Once we complete the publishing and monitoring services, we’ll work on a set of consumption services.

Implementation approach

As we discussed before, we took the “crawl, walk, run” approach to implementing these standards. We started with one functional area, Finance, and carried out the data lake platform standard. With this initial implementation, we were able to prove out and refine our approach. We followed up with two more implementations: Human Resources and Sales and Marketing. With these three functional areas onboarded to the data lake platform, we had a critical mass of data that not only proved out the approach, but also encouraged other teams to onboard due to the need to associate with these core systems.

Data lake benefits

Data lakes enable us to move beyond the limitations of traditional analytic methods constrained by legacy data warehouse technology. At CSEO, we’re relying on data lakes to spearhead several key initiatives:

- Integrating machine learning and AI. Enabling sophisticated predictive analytics helps data consumers across all levels of the company to make the best decisions.
- Managing data to enable collaboration. Replacing disconnected data silos with well-managed data lakes greatly boosts overall agility—teams can access data on demand and pivot quickly to adjust to changing conditions. Opening access to data repositories across groups encourages new forms of experimentation and exploration to quickly test innovative ideas and concepts.
- Redesigning storage to maximize cost efficiencies. Separating storage from compute resources, scaling compute resources up or down to meet rising or falling demand, and choosing the right storage for the data greatly lowers storage costs and promotes innovation.
• **Enabling effective data governance.** By modernizing our data management framework, we can realize greater economies of scale arising from coordinated governance, implementing data quality standards, and simplifying regulatory compliance processes.

The most value accrues from acquiring critical insights not readily possible before the advent of cloud-scale computing resources. Without expertise in data science, a broad spectrum of users can now identify, analyze, and predict previously unnoticed trends and relationships. Examples include:

  • **Incorporating precision attribution.** Augmenting traditional “small” data with other big-data attribution can produce unique insights or challenge old assumptions. One example is matching a traditional “selling” transaction with “usage telemetry” to predict consumer choices or even find upsell opportunities.

  • **Running high-speed analytics.** Discovering a data point on time series data that generates an alert in milliseconds can be as important as knowing why an issue is happening.

  • **Finding hidden patterns in data.** Using inductive analysis can uncover a new hypothesis or theory, revealing previously unseen or overlooked cause-and-effect relationships, trends, or other patterns.

Figure 3 shows how inductive analysis underpins the increasing sophistication of analytical methods from descriptive and diagnostic analytics to predictive and prescriptive analytics.

![Figure 3. The rise of predictive and prescriptive analytics](image)

**Governance**

In Microsoft CSEO, teams value their autonomy, and the culture isn’t one that values “being governed.” However, some form of governance must exist to ensure that data lakes don’t become swamps. Good governance doesn’t stop systems from accessing the data they need or make developers’ lives harder. Rather, it allows us to design a consistent framework to ensure that:

  • Data consumers can rely on transparent “source of truth” data, meaning they can trust or easily verify that data is being used for its intended purpose.

  • Users or processes avoid duplicating data without appropriate justification.

  • Administrators can efficiently manage secure role-based authentication to data resources and help ensure that change management processes maintain dependencies within downstream systems.

Read more about how we’ve set up data governance in this paper: "Driving effective data governance for improved quality and analytics."
Data catalog

Conceptually, users use data catalogs to see what data is available across the enterprise and provide data consumers with a technical and business view of the data.

Engineering teams and data scientists require a technical view of raw data sources showing:

- Data connections such as primary and foreign keys.
- Data lineage (an audit trail indicating all changes to the data and who made the changes).
- Data quality of the source.

Business analysts need metadata that shows a more generic business view, along with clear indicators of the “source of truth” to verify that the data is suited for intended business scenarios.

One of the key implementation goals of the CSEO catalog is to automate the catalog process as much as possible, without needing regular maintenance from developers. To do this, we designed the catalog to record metadata whenever source data is placed into the lakes. The catalog also automates audits to check for validity and consistency across the enterprise.

Read more about how we implemented our data catalog in this paper: “Designing a modern data catalog at Microsoft to enable business insights.”

Data quality

Having confidence in the quality of data depends on several factors, including understanding where the data has come from, the meaning of the data, the age of the data, and the accuracy of the data values.

Measuring data quality across data types can mean different things to different people. At CSEO, one of the first things we realized was the need to agree on a common vocabulary and framework.

Taking stock of the divergent ways of describing data quality, we settled on an industry-standard data-quality dimensions matrix, as shown in Table 2.

Table 2. Data quality dimensions

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>Provides a degree of certainty that data correctly reflects the fact, truth, event, or “real-life” object it’s intended to model.</td>
</tr>
<tr>
<td>Timeliness</td>
<td>Measures compliance to all expectations of data with respect to time.</td>
</tr>
<tr>
<td>Consistency</td>
<td>Describes the degree of agreement, harmony, or compatibility of the data with a point of reference.</td>
</tr>
<tr>
<td>Completeness</td>
<td>Shows the presence of all the necessary or appropriate parts.</td>
</tr>
<tr>
<td>Uniqueness</td>
<td>Indicates whether a real-life object, a concept, or an event is represented only once.</td>
</tr>
<tr>
<td>Validity</td>
<td>Measures agreement with a source of “correct” information—such as reference data, certified data, master data, or another authoritative source. Validity is also used to measure the absence of “incorrect” information, such as not on a “Do Not Contact” list, no slang words in a company name, or not a dummy email such as <a href="mailto:abc@hotmail.com">abc@hotmail.com</a>.</td>
</tr>
<tr>
<td>Conformity</td>
<td>Measures the degree to which data values comply with metadata attributes including data type, precision, format patterns, domain ranges, and underlying storage formats.</td>
</tr>
<tr>
<td>Precision</td>
<td>Provides the right level of data value granularity, such as numeric rounding or number of decimal places.</td>
</tr>
</tbody>
</table>

Our initial goal is to provide a suite of tools based on an agreed set of standards to automate assessment of the data. With these in place, our data consumers across the company can use the data quality index in the catalog for key datasets. In the longer term, we plan to automate metrics to check compliance across systems.
We’re already seeing benefits from our newly implemented data quality system for SQL-based relational data—it measures completeness, uniqueness, and timeliness. It includes a central set of data quality rules, and each business unit can add their own business rules for each dataset. Next, we’ll extend this capability to open-source solutions.

Within the data-management services we’re developing, the quality checks vary depending on the type of data in the file. The two main types of data that we load into data lakes are:

- **Raw data.** This is the data in its earliest form, as captured. For example, our main source system, SAP, has online transaction data that is dumped directly into the data lake every 30 minutes. For this raw data we check the information that is relevant for any data file and apply a basic data-quality check to make sure the data is “fit for use.”

- **Cooked data.** This is data that one of our functional business teams (e.g., Revenue Allocation) reads from multiple sources and transforms to create a new data set. This resultant data set is the cooked data and is placed back in the data lake. For this cooked data, we have more stringent quality requirements. We have a set of business rules that are specific to the data.

Automation is the key to ensuring data quality. Implementing data-quality checks as the data is loaded into the data lake, and automating the quality tagging and monitoring of data quality after the loading, keeps data healthy and ensures its hygiene continues throughout the data lifecycle.

**Benefits**

In our mission to build a fully modernized framework for managing big data, much of our success will depend on what the data shows. To help guide our priorities and measure our overall impact, we needed to set vigorous and achievable goals. As each organization adopts these strategies, we’ll start tracking progress against these targets:

- Integrate AI and machine-learning analytics into 80 percent of new projects.
- Increase data discoverability by 20 percent.
- Reduce data duplication by 20 percent.
- Reduce data latency by 25 percent.
- Lower storage costs by 25 percent

**Conclusion**

We’re continuing our journey to develop and mature our big-data management program. The DAWG team meets regularly, and each of these focus areas will continue to evolve. We’re evangelizing and expanding the data-sharing culture across our internal organizations so that our enterprise data can be joined, queried, and analyzed—all with the aim of acquiring deep new insights.

Early results from our own data lakes are helping drive new efficiencies, powered by insights previously unavailable under the traditional data warehouse model. For example, to help our Real Estate & Facilities team determine the most underutilized areas of campus buildings, data scientists collated data sourced from HR and the networking team responsible for maintaining wireless network routers. By analyzing both datasets against their own geolocation data for offices, meeting rooms, and other indoor locations, the data scientists provided the facilities team with valuable precision data. With this information, Microsoft can improve building capacity and utilization without renting or building new offices.

Today, we’ve succeeded in getting key parts of our organization aligned around a common framework for managing big data. And we’re seeing positive progress toward implementing critical components needed to yield concrete business results in the year ahead.

Taming the “wild west” of big data is an essential step toward our future. Managing data as a core strategic asset is integral to deepening our engagement with customers, employees, and partners—it’s an essential part of our business going forward. We’re well on our way toward digital transformation.

**For more information**

[Microsoft tames the “wild west” of big data with modern data management](https://microsoft.com/itshowcase)

[IT Showcase Data and AI content](https://microsoft.com/itshowcase)
Driving effective data governance for improved quality and analytics
Designing a modern data catalog at Microsoft to enable business insights
Azure Data Lake connects supply chain data for advanced analytics
Managing the data life cycle using Azure Data Factory