snowflake®

BEST PRACTICES FOR MANAGING UNSTRUCTURED DATA

CHAMPION GUIDES

TABLE OF CONTENTS

EXECUTIVE SUMMARY

Unstructured data accounts for a vast and rapidly growing amount of information. According to Computer Weekly, four-fifths of all business-relevant information—mostly text (for example, emails, reports, articles, customer reviews, client notes, and social media posts) but also audio, video, and remote system monitoring data—originates in unstructured data.[1]

However, unstructured data poses a number of challenges for organizations attempting to extract value from it using legacy data management tools. It is not easy to search, analyze, or query—especially on the fly. Its complexity creates processing problems for extracting analytical insights. Poor visibility and control create other issues with regard to governance and data security.

A modern data management platform that can effectively incorporate unstructured data (along with structured and semi-structured files) offers valuable advantages such as more complete data analysis and better insights for decision-making. An effective solution must include three core capabilities: It should eliminate data silos; provide fast and flexible data processing; and ensure easy, secure access.

THE EVOLUTION OF DATA MANAGEMENT

The ability to analyze data is why businesses, governments, and other organizations invest in computers. Extracting insights to gain tactical and strategic advantages has always been the goal. The first computers were essentially solving long, hard math problems on the small amounts of raw data available at that time.

But today, data arrives from diverse sources in massive amounts and can appear in any form—structured, semi-structured, or unstructured. Traditional data management technologies are unable to consistently support multiple data formats, causing organizations to seek out new methods for getting maximum value from all of their data.

Each form of data is important, and all must be used to form a full analytical picture.

STRUCTURED DATA

Conventional data management systems were designed decades ago, when data arrived in very predictable, structured formats. Relational data with fixed schemas was the norm because data sources were limited and didn't change very often. Table-based data warehouses offered highly controlled environments for storing and managing this kind of data. At this time, most data analysis was limited to structured data, because the data was well organized and could be easily read by analytics algorithms.

SEMI-STRUCTURED DATA

The rapid decrease in the cost of storing data and the growth in distributed systems led to an explosion of machine-generated data. Semi-structured data formats such as JSON, Avro, and others became the de facto form in which this data is sent and stored. This data was always intended to be more machine friendly—both in how it's generated and how it would later be processed programmatically.

As generally defined, semi-structured data does not obey the tabular structure of table-based data management systems developed for and by humans, but it does contain tags or other markers to separate semantic elements and enforce hierarchies.[2]

Data lakes emerged over the last decade and made it easier to manage semi-structured data. More recently, some organizations have relied on a mix of table-based and file-based management systems.

UNSTRUCTURED DATA

While data lakes expand management and analytics to more kinds of data, these architectures don't work well for the rapidly expanding quantities of unstructured data that businesses are now collecting. There has been a rapid increase in the amount of unstructured data that needs to be analyzed. According to IDC projections reported by Analytics Insight, 80% of the world's data will be unstructured by 2025—and just 0.5% of these resources are being analyzed and used today.[3]

Humans natively create unstructured data. In the same way that machines interacting with the world create huge volumes of semi-structured data, humans interacting with organizations create a huge volume of unstructured data. Unstructured data is defined by the fact that it is not organized in a pre-defined manner—which results in irregularities and ambiguities that make it difficult to manage, secure, govern, and process using traditional approaches, according to Wikipedia.[4]

Examples of unstructured include digital files that contain complex data such as images, videos, audio, and .pdf documents. It also includes many industry-specific file formats: DICOM (medical imaging); .vcf (genomics); .kdf (semiconductors); and .hdf5 (aerospace).

Unstructured data is widely regarded as an untapped resource for feeding customer analytics and marketing intelligence applications. While there's vast potential for extracting value from unstructured data, its complexity and the sheer volume of information being generated requires a new evolutionary step in how this kind of data is managed. Organizations need an easy way to access, process, and govern their stores of unstructured files.

HOW CAN UNSTRUCTURED DATA HELP YOU?

When done well, incorporating unstructured data into your analytics and decision-making can open up a new perspective for your organization—as well as new opportunities. Here are a few examples of what you can do with unstructured data:

- Analyzing customer behavior on social media to inform targeted marketing campaigns by identifying specific regions or the demographics of customers who are talking about a specific product.
- Expediting automobile insurance claims processing by automatically applying machine learning (ML) to image files for pattern recognition.
- Analyzing call center audio recordings to derive marketing insights such as sentiment analysis.
- Scanning doctors' handwritten notes for terms that could indicate good clinical trial candidates and joining that information with structured data to identify and register trial candidates faster.

CHALLENGES FOR UNSTRUCTURED DATA MANAGEMENT AND ANALYTICS

Traditional data management systems (that is, data warehouses and data lakes) aren't able to support all of the workload demands for today's data volume, velocity, and variety of formats. As a result, these systems have to add different tools to support the various types of data (structured, semi-structured, and unstructured).

According to DCIG, "Many organizations now need to manage multiple petabytes of data. At petabyte scale, storing, protecting, backing it up, and recovering it all is problematic using legacy solutions." [5]

Blob storage services provided by the public cloud providers (such as Amazon S3 and Azure Blob Containers) have become the default storage for unstructured data files. However, these have many limitations for analytics use cases. For example, listing files in blob storage can be challenging and limited to only prefix-based searches. Without the formal table-based or file-based organizational system to help guide data storage, consistently accessing, managing, controlling, searching for, and securing unstructured data with these services becomes much more difficult.

UNSTRUCTURED DATA COMPLEXITY

Unstructured data itself is complex and hard to analyze. The different file formats that make up a stored body of unstructured data are also separate, and it can be difficult to make cohesive sense of the assembled set of information.

Joining unstructured data sources with other data formats or data sets is particularly challenging—especially if the unstructured data involves audio and video media files. These issues lead to siloed and unused data. When data is stuck in silos, organizations experience limited query performance due to poor visibility, and some data is entirely inaccessible.

DATA PROCESSING ISSUES

Reliance on disparate data management tools and systems also creates complex data pipelines that degrade analytics performance. Converting unstructured data to structured data by extracting text from PDF files or using image recognition software can be cumbersome, compute-intensive, and time-consuming. Relying on legacy solutions for managing unstructured data leads to processing problems such as broken data pipelines and error-prone data movement due to frequent copying of data from one place to another. It also slows digital transformation efforts, preventing you from seeing the intended business impact of data operations and fulfilling the organization's goals.

GOVERNANCE AND SECURITY UNCERTAINTIES

When high volumes of complex unstructured data combine with the rigid architectures of traditional data systems, managing data access becomes very difficult. This is especially true when it comes to limiting access based on the specific type of data and the user's role (necessary for implementing "zero trust" security controls).

According to Security Weekly, government cybersecurity experts have clearly settled on moving to the cloud and implementing a zero-trust architecture as being the two most immediate and practical methods to improve the nation's cybersecurity posture. [6]

Data privacy laws—such as the EU's General Data Protection Regulation (GDPR)—don't distinguish between structured and unstructured data. Regardless of its form, data that contains private information must remain under the control and protection of an organization at all times. According to CPO Magazine, GDPR fines jumped by 39% in 2020 and the total fine count as of January 2021 for European Union member states totals about $332.4 million USD.[7]

Gartner predicted that, "By 2023, 65% of the world's population will have its personal data covered under modern privacy regulations, up from 10% in 2020." [8]

Specific governance and security issues surrounding unstructured data include:

- **Migrating existing permissions**. Unstructured data is often sourced from other platforms where the files already have complex permissions related to those systems. Understanding those permissions is complex, and then mapping them to a new platform is extremely challenging.
- **Data sharing**. According to Verizon, 61% of data breaches last year involved credentials, and 25% specifically used stolen credentials.[9] How can an organization give users access to data without giving them credentials?
- **Risks with data movement**. Siloed data that is copied and then resides in multiple places creates a lot of unnecessary risk exposure.
- **Right to be forgotten**. Data that's inaccessible or that has been copied across disparate management architectures can be difficult to fully expunge to maintain compliance with different regional data privacy laws. This introduces the risk of regulatory fines and potential litigation costs.

WHAT AN EFFECTIVE SOLUTION LOOKS LIKE

Storing and governing unstructured data is one of the most important tasks for data architecture administrators. An effective solution for managing unstructured data should include built-in capabilities to store, access, process, govern, secure, and share an ever-expanding volume of this data. As such, the system must specifically deliver sufficient performance, concurrency, and scale while solving the critical shortcomings of the legacy approaches in place today.

NO DATA SILOS

Modern data management needs to be based on a single, cloud-based platform that supports all data formats (structured, semi-structured, and unstructured) to easily store, access, process, share, and analyze files. Data engineers should be able to store and retrieve files in a cloud-agnostic way—so data is accessible across clouds and regions—while still enforcing unified policies.

The solution should use a simplified architecture to help reduce maintenance and management overhead. It should also offer flexibility to store unstructured data files in either an internal or external stage.

FAST, FLEXIBLE PROCESSING

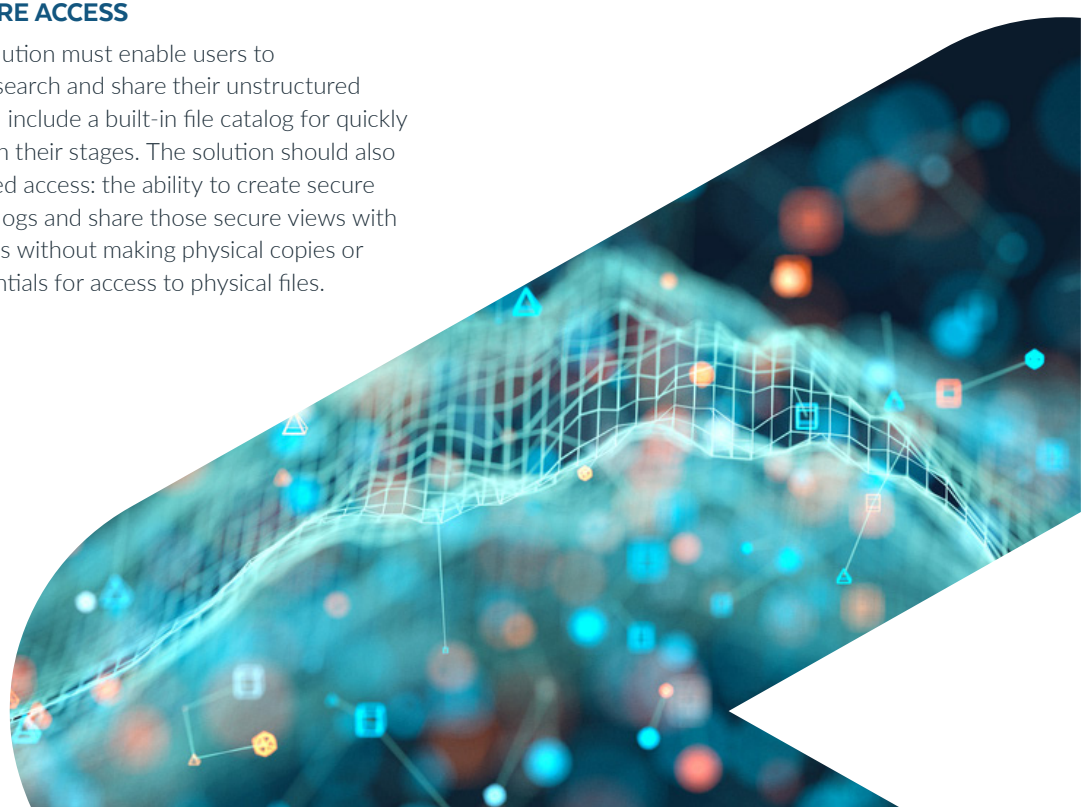
A modern management solution depends on ample processing capabilities that can transform, prepare, and enrich unstructured data to extract more-complete insights using complex analytics, data science, and interactive applications. The solution must provide fast, reliable performance without needing manual tuning or without causing workload contention. It should offer elastic workload concurrency via cost-efficient scalability across any volume of users, jobs, or data.

In addition to compute performance, data scientists need to be free to work with their tools of choice to process unstructured data, maximizing their productivity. Also, to ensure a continuous data pipeline, the solution needs to make its outputs easily available and transparent for others to use.

EASY, SECURE ACCESS

Finally, the solution must enable users to conveniently search and share their unstructured data. It should include a built-in file catalog for quickly locating files in their stages. The solution should also support scoped access: the ability to create secure views on catalogs and share those secure views with other accounts without making physical copies or sharing credentials for access to physical files.

Organizations need governance at scale with flexible policies that follow the data for consistent enforcement across users and workloads. In support of zero-trust requirements, a data management solution must help control access to sensitive data as appropriate for a user's defined role. To achieve this, governance for unstructured files should use cloud-agnostic role-based access control (RBAC) commands (such as simple GRANT and REVOKE statements). This avoids the potential complexities of security or governance policies in each cloud provider's identity and access management (IAM) system.



A GLOBAL, UNIFIED SYSTEM FOR MANAGING UNSTRUCTURED DATA

The next evolutionary step in data management should be defined by how all forms of modern data can be shared and consumed—not just by internal teams, but by customers and partners as well—to extract maximum value.

To achieve this, organizations need a global, unified system for connecting companies and data providers to the most-relevant data for their business. An effective solution must combine structured, semi-structured, and unstructured data and provide a single and seamless experience for storing, processing, and analyzing data across public clouds

The best practices in this ebook will help you start maximizing the value of all your data today. To learn more about how you can store, access, process, govern, and share unstructured data in a single data platform, watch our 7 Ways to Start Using Unstructured Data in Snowflake webinar (support for unstructured data currently in preview).

ABOUT SNOWFLAKE

Snowflake delivers the Data Cloud—a global network where thousands of organizations mobilize data with near-unlimited scale, concurrency, and performance. Inside the Data Cloud, organizations unite their siloed data, easily discover and securely share governed data, and execute diverse analytic workloads. Wherever data or users live, Snowflake delivers a single and seamless experience across multiple public clouds. Snowflake's platform is the engine that powers and provides access to the Data Cloud, creating a solution for data warehousing, data lakes, data engineering, data science, data application development, and data sharing. Join Snowflake customers, partners, and data providers already taking their businesses to new frontiers in the Data Cloud. **Snowflake.com**

CITATIONS

[1] bit.ly/3if52aK
[2] wikipedia.org/wiki/Semi-structured_data
[3] bit.ly/2XQufkz
[4] wikipedia.org/wiki/Unstructured_data
[5] bit.ly/3ClqnkS
[6] bit.ly/2WlxxvU
[7] bit.ly/3if6V7A
[8] gtnr.it/2XSRzOC
[9] vz.to/3zLjyx8