White Paper

Illuminating Insight for Unstructured Data at Scale

Sponsored by: IBM

Amita Potnis
October 2018

IDC OPINION

Data is a critical corporate asset and the value of data is tied to how it is ultimately used. Today's digital world is data-driven. Business decisions and strategy are driven by insights from analyzing data. The nature of business is transforming to deliver higher value and the key to this is to make the underlying IT infrastructure seamless. Data management systems provide specialized features for the ingestion, cataloging, transformation, rights management, retrieval, and resource optimization of data.

Large data repositories are critical for garnering insights for business development through analytics. However, the data repositories can only be tapped for the wealth of insights if the data is classified and tagged appropriately to enable search and query. In the absence of metadata – classification and tagging – datasets can be rendered unusable for business development purposes.

Organizations have come to terms with the rate of unstructured data growth and recognize they are largely unable to derive value from their vast data repositories. The problem is amplified when data sits in silos as different data types across storage deployment locations (traditional and private/public cloud). The market needs data management tools that will simplify the ETL (extract-transform-load) processes to support three main areas: governance, analytics and storage optimization. IBM Spectrum Discover is a new data management product in the market that is set to address these needs and support any organization’s goal of solidifying their business roadmap.

IN THIS WHITE PAPER

This IDC white paper evaluates the need for a data management solution in the market today and highlights IBM Spectrum Discover, a new offering designed to manage unstructured file- and object-based data that is part of the IBM Spectrum Storage portfolio.

SITUATION OVERVIEW

Today, we are in what IDC defines as the 3rd Platform that stands on four main pillars that contribute to data growth – social, mobile, Big Data and cloud. Over the past 50 years, the average life span of S&P 500 companies has shrunk from around 60 years to 18 years. To survive, companies not only have to be digital transformers but must do so while improving adaptability to stay relevant and profitable in the market. Over the last decade, data growth is driven by the influx of mobile devices, sensors, surveillance and consumer cameras, web-based social or transactional interactions. Data is created and stored at many locations: core, edge and endpoints and across several infrastructure platforms including traditional on/off-premises, private and public cloud as structured (block) or unstructured (file and object) data. As market shifts and rapidly changing technologies transform businesses, companies that don't have up-to-date, evolving skill sets and intelligent data management tools will fall behind.

IDC estimates that in 2025, the world will create and replicate 163ZB of data, tenfold what was created in 2016. Such data growth puts intense pressure on IT organizations to manage and maintain infrastructure and keep data secure and available at all times. In addition, IT organizations struggle to keep pace with business needs and in turn infrastructure requirements. The use of mobile devices and social media is a driver for growth in unstructured data such as videos, audio clips, emails, and images. This unstructured data needs to be stored for extended periods of time because of regulations or business requirements and analyzed for new growth opportunities thus increasing the size of datasets in an unprecedented manner. The file- and object-based storage forecast estimates that by 2021, 422EB of storage capacity will be deployed to support scale-up and scale-out file- and object-based storage environments which are growing at a CAGR of 30.6% between 2016-2021.

Table 1 shows raw storage capacity deployed for file- and object-based storage for both on and off-premises. Based on the high percentage of storage capacity deployed in file- and object-based storage environments, Table 2 shows the high amount of unstructured data that organizations must manage.

TABLE 1

Raw capacity deployed on- and off-premises storage in support for block, file and object environments

	1TB - 100TB	101TB - 1PB	1PB+
on-premises storage	57.6%	27.1%	15.3%
off-premises storage	58.0%	26.0%	16.0%

Source: IDC File- & Object-based Storage Survey Results, 2017 (n=450, North America only)

TABLE 2

% of raw capacity deployed by file and object-based storage

	Object-based Storage	File-based Storage
on-premises storage	35.0%	35.5%
off-premises storage	37.3%	33.5%

Source: IDC File- & Object-based Storage Survey Results, 2017 (n=450, North America only)

Systems built for analytics for machine learning and artificial intelligence, or gene sequencing, for example, support a broad number of use cases including (but not limited to) more intelligent/preemptive search, automatic metadata generation, more intelligent document classification, data extraction, and optimization of and decisioning in automated content-intensive processes. At the same time, IDC research indicates that organizations plan to incorporate hybrid/multi-cloud storage strategies in their infrastructure roadmaps thereby increasing the possibility of data silos. Cloud adoption will add to the need for robust data management systems to support greater business agility, quicker ROI and increased profitability. As unstructured datasets continue to grow within siloed storage infrastructure, it is important that organizations have the means to bring order to their data management. Without appropriate tools to manage data across silos, organizations can expect an increase in management overheads while also missing out on valuable data insights. Tools that leverage metadata – accelerated data classification, tagging and indexing – allow organizations to search, find, retrieve and analyze data at faster speeds, at greater scale, and with greater efficiency than otherwise would be possible. As organizations continue down the path of digital transformation and their datasets grow, it is important they consider data management tools that will answer several if not all the following questions in three important areas:

Governance

- Who is accessing what data and how frequently?
- What data is redundant, obsolete or trivial?
- How sensitive is the data in terms of security?
- Who within my organization owns this data?

Storage Optimization

- Where is the redundant, obsolete or trivial data stored?
- How many copies of data does my organization have today?
- What data is frequently accessed (hot), and what is not?
- What data is mission-critical and what is not?
- How much and what type of data does my organization currently have and where it is stored?

Big Data & Analytics

- Which application/s consume this data today and how?
- Where/how did the data originate? Who is using this data and how?
- Which entities, keywords, concepts, facets are in each file or object?

IMPORTANCE OF METADATA AND DATA MANAGEMENT TOOLS

Metadata can be defined as data that describes other data. Metadata offers visibility and control of data by associating relevant criteria. Creation date, author, file size, location, modification date are examples of system metadata that provides additional description of the data at hand. System metadata is typically limited in scope and use because it does not describe what is inside the data file. Custom metadata enriches the context and allows for it to be organized or structured with greater specificity. System metadata enables data to be organized or structured as well, just at the more macro level.. . Custom metadata in addition to system metadata streamlines and enhances the process of collecting, maintaining, searching, integrating and analyzing data, thereby supporting success in initiatives such as machine learning/artificial intelligence, life sciences research, and so on.

Application and custom generated metadata can help organizations run queries faster and in a simplified fashion to support governance, infrastructure optimization and analytics initiatives. The following are the ways organizations choose to deliver value for three key areas using intelligent data management tools:

- **Governance** requires strict monitoring and auditing of who has access to which data sets and controls data ownership, roles and responsibilities to keep data secure and compliant. The ideal data management solution will provide the best solution for a combination of people, policies, and processes as it relates to data governance.
- **Storage optimization** requires the ability to identify or tag data by usage and/or policy and move it to an appropriate or designated storage tier for optimal utilization of available resources. Centralized metadata management allows a common view of data residing on disparate storage including file- and object-based storage. The data management tool should have the ability to identify and report:
  - Redundant data and reduce the number of copies of it
  - Obsolete data and purge it
  - Trivial data and either purge it or tier it to an appropriate storage tier
- **Analytics** requires ELT (extract-transform-load) processes to be delivered at lightening speeds to analyze large datasets and find correlations and hidden patterns.

IBM SPECTRUM DISCOVER

As organizations grapple with managing their datasets and extracting value from them, vendors have responded to the need of the hour with appropriate data management tools. IBM has recently announced IBM Spectrum Discover, a tool that activates hidden value in data, offers visualizations of storage utilization by data categories, and automatically captures and indexes system metadata with support for custom business-oriented data tagging.

IBM has a long-standing reputation of being an established storage systems player. IBM Spectrum Storage is the company's comprehensive portfolio of software-defined storage infrastructure, storage services, and data management solutions. IBM Spectrum Discover expands and enhances the already robust portfolio with new data management capabilities designed to meet the growing demand for metadata-driven data insight for file and object storage.

Recognizing the need for a metadata-based search and governance tool, IBM has come to market with IBM Spectrum Discover. IBM expects this offering will help organizations automate cataloging of unstructured data by capturing metadata as it is created and combine custom metadata tags with system metadata for better visibility and data control.

Announced in October 2018, the product is generally available in November 2018. The offering includes the following features:

- Policy-based metadata tagging which allows rules to identify data and apply custom tags
- Ability to automatically and/or manually apply tags based on user's pre-defined schema
- Metadata tags can be open or restricted. Restricted tags apply pre-defined values vs. open tags that allow for user defined values.
- Role-based access for enhanced security

- Basic and advanced search capabilities that include key-value and range with ability to apply filters to refine search results
- Use of SDKs, custom tags and policy-based workflows to orchestrate and accelerate content identification
- SDK to integrate third party applications, open source and/or commercial
- Intuitive and easy to use dashboard and customizable reports allow users to gain granular insights
- Saved history of searches for re-running queries without having to re-write them
- Duplicate file detection capability for potentially redundant data identification and notification
- Supports efficient use of storage and helps protect against data loss

Currently, IBM Spectrum Discover indexes, classifies and manages data stored on IBM Spectrum Scale and IBM Cloud Object Storage. IBM plans to extend this offering to include competitive and established third party file and object storage products. IBM Spectrum Discover can scan 30,000 records/second. The product also leverages live event notifications to update its database when files/objects are created, updated or deleted. IBM sees this as delivering benefits across industries, but plans to target the healthcare, financial and telecom industries in this initial launch phase. End users will be charged based on the amount of data under management on a per terabyte basis.

The intent of IBM Spectrum Discover is to enable organizations to improve the following:

- **Governance**. Mitigate risk and improve data quality and lifecycle management.
- **Storage optimization**. Improve storage utilization by tiering to appropriate storage tier and eliminating redundant data therefore reducing costs.
- **Big Data & analytics**. Reduce time to achieve accurate results with deeper insights into the vast body of unstructured data by improving and enhancing analytics workflows

Overall the product is set to address these data management challenges in a simple and easy fashion such that end users can adopt IBM Spectrum Discover and immediately benefit from the functionality of the product.

CHALLENGES/OPPORTUNITIES

IBM's strength lies in its diverse and established portfolio, not just within the IBM Spectrum Storage portfolio but also in the cloud and analytics spaces. The company has expertise not only in the infrastructure space from a systems and software perspective but also within the services market. IBM keeps pace with market demands by recognizing the needs of the hour and developing solutions that address requirements for today and the future. Organizations considering IBM Spectrum Discover can take advantage of IBM's breadth of knowledge and expertise of requirements beyond data management.

In addition to the existing functionality, organizations should take into consideration support that IBM Spectrum Discover will offer for existing IBM products and platforms. For example, in the future IBM Spectrum Discover users can leverage IBM Watson Data Platform to search and analyze relevant data from IBM Spectrum Scale and IBM Cloud Object Storage.

Today, IBM Spectrum Discover has out-of-the-box support for IBM Spectrum Scale and IBM Cloud Object Storage data sources. It is important for any data management tool to be storage-agnostic for long-term success, and IBM Spectrum Discover provides an extensible platform to support other data sources via an SDK. IDC expects that IBM Spectrum Discover will penetrate the market at a rapid pace as the product expands it capabilities to support heterogenous storage products across deployment locations (traditional or private/public cloud).

CONCLUSION

Maintaining detailed and custom metadata will contribute to any organization's initiatives around governance, cost savings or analytics. Maintenance of metadata will further ensure complete and full realization of business value of any organizations' extensive data assets. Therefore, data management tools should not just be part of a data governance strategy but should also be part of a long-term business roadmap.

Organizations are not able to predict the rate of unstructured data growth and in turn are not able to astutely allocate infrastructure resources. Given the anticipated growth of data, this is only the beginning of data management challenges. Organizations are looking to derive value from data and in turn drive business — activities that demand immediate use of data management tools. Users are encouraged to conduct a proof of concept and evaluate the value IBM Spectrum Discover can bring. To the extent that IBM can address the challenges described in this paper, IDC believes the company has a significant opportunity for success in the data management marketplace.

About IDC

International Data Corporation (IDC) is the premier global provider of market intelligence, advisory services, and events for the information technology, telecommunications and consumer technology markets. IDC helps IT professionals, business executives, and the investment community make fact-based decisions on technology purchases and business strategy. More than 1,100 IDC analysts provide global, regional, and local expertise on technology and industry opportunities and trends in over 110 countries worldwide. For 50 years, IDC has provided strategic insights to help our clients achieve their key business objectives. IDC is a subsidiary of IDG, the world's leading technology media, research, and events company.

Global Headquarters

5 Speen Street
Framingham, MA 01701
USA
508.872.8200
Twitter: @IDC
idc-community.com
www.idc.com

Copyright Notice

External Publication of IDC Information and Data — Any IDC information that is to be used in advertising, press releases, or promotional materials requires prior written approval from the appropriate IDC Vice President or Country Manager. A draft of the proposed document should accompany any such request. IDC reserves the right to deny approval of external usage for any reason.

Copyright 2018 IDC. Reproduction without written permission is completely forbidden.