FOREWORD BY MANDY CHESSELL

AMNESIA OR PROGRESS?

Over the last few years, big data processing techniques and practices have matured. We no longer hear claims that the sheer volume of "big data" renders old fashioned, complex and time-consuming practices, such as data quality management, obsolete. In fact, many practices that were developed for data warehousing are being introduced into big data projects. There is a rise in the demand for data modelers, and technologies such as SQL and ETL engines are being ported and enhanced to execute on big data platforms.

So what have we gained in this movement to big data? Was it just a case of collective industry amnesia as we cast aside the years of information management experience developed, often painfully, from large-scale data warehouse projects, or has something significant happened in information management?

THE LEGACY OF DATA WAREHOUSING

During the big data frenzy, the use of data warehouses has not diminished in absolute terms. However, they no longer represent the only approach to collating and managing information outside of the operational systems that originate data.

The shift away from data warehouses was driven by cost, flexibility and time to value. A data warehouse aims to construct a coherent view of an organization's operation. This takes time because most organizations do not operate in a coherent manner. It takes thought and cross-organizational agreements to determine how the disparate operations are reconciled. This time and effort was perceived as expensive. To reduce this cost, the information governance program pushed greater conformity back into the operational systems feeding the data warehouse in a way that gave it a reputation for stifling innovation.

Thus people looked for a new approach to deliver a more agile and flexible data capability that would enable a more data-rich operating model for the organization.

The big data teams used new technology, new techniques and seemed to deliver proof of concepts, and their initial projects with lightning speed. However, as the volume and variety of data grew, the big data projects began to stall, amid a wave of concerns about the safety and trustworthiness of the data they hosted. Organizations started to question if big data had really brought them any gains over the data warehouse and began looking for another miracle technology.

LOOKING BACK AT THE IMPACT OF BIG DATA TECHNOLOGY

Despite the problems, the result of the big data movement is significant, but not as revolutionary as enthusiasts originally thought. Big data projects have been dogged with the same types of problems related to data understanding, data quality and project over-runs as data warehouse projects. There is increasing recognition that extracting, transforming, linking and collating data from heterogeneous

systems is innately hard and although big data processing makes it faster and cheaper to process data, it does not remove much of the complexity and skill needed.

For the more thoughtful organizations there is a realization that information management techniques should largely be agnostic to the data platform and the type of data. The result is that:

1. Data warehouses are seen as complementary rather than competitive to big data platforms. Their use is focused on highly optimized processing of data for standard reports and dashboards. The big data platforms offer generic capability for all types of data, which is useful for analytics development and experimentation plus production workload that are less time-critical in nature.
2. Data is processed selectively from raw format to finished data service in an agile and modular fashion. There is no longer an attempt to make all data fit into a single data model.
3. Common data models are used to create consistency between data service implementations, not to create a single coherent view of the organization's operation.
4. The data governance program is becoming targeted and selective rather than a set of standards applied to all data. It operates across all platforms.
5. Metadata is being used operationally for online data catalogs, virtualized access and active governance of data.
6. Techniques for data quality, lifecycle management and protection are being homogenized to support all types of structured and unstructured data.
7. Data is seen by the business as an asset and its use is now a discussion at the boardroom level.

Information management is getting harder due the diversity of information producers today. However, the big data movement has forced a significant step forward in information management practices beyond those developed for the data warehouse, and as a result we are better placed to manage this.

THE IMPACT OF CLOUD TECHNOLOGY

The movement to cloud is the next great playground. Business teams can select and purchase new services without involving anyone from the IT. This appears to remove a bottleneck to progress – particularly when a mature organization is trying to transform it operations. However, from a data point of view, these new cloud services are creating new silos of data distributed across different providers' data centers and the relief it offers to organizations will hit a similar wall as big data when they try to integrate their new services with their existing business.

There are also technical challenges still to be addressed. Cloud computing aims to virtualize infrastructure and IT services so that they can be seamlessly shared, to reduce cost and flexibility.

Data presents a challenge to cloud environments because it is a physical resource in an infrastructure that aims to scale by virtualizing resources.

How does an organization position data and workload across a hybrid multi-cloud ecosystem in a way that enables their business to operate coherently and efficiently?

Today this is a labor intensive, manual effort that needs significant research and investment to automate. These different infrastructure environments do not currently capture and exchange enough metadata and operational information to make this possible, even for a static environment. Given that organizations today are continuously evolving, the ecosystem must be able to dynamically evolve with it.

CONCLUSION

This is indeed an exciting time for practitioners in data-related professions. The technology is advancing, enabling the economic processing of many more types of data in many more types of systems.

To take full advantage of these advancements, information management needs to step above the technology landscape and focus on managing the movement, consumption and management of information as a coherent backbone of the organization. Supporting this backbone is a variety of technology that is selected and tuned to the workload requirements of the organization's operations.

This information backbone has to assume it is operating in a hybrid, multicloud ecosystem. Thus information management and governance capabilities need to be consistently embedded in cloud platforms from all vendors and on premises systems. They need open interfaces and well-defined behaviors to enable cloud brokerage services to correctly position both data and workloads on the most effective processing platform whilst keeping track of the organization's assets as a coherent enterprise-wide view.

Mandy Chessell
IBM's Hursley Laboratory, UK