HOW YOUR BUSINESS WORKS
information management
December 20, 2012

Data Governance for Big Data

By Krish Krishnan

The world we live in today is caught in a frenzy over the term big data. The usage of these two words in any context draws attention from executives to data scientists and business users in organizations of any size.

While there are many definitions floating around for the term, here is a simple definition for this discussion: Big data can be defined as data that cannot be processed using traditional data processing techniques due to its characteristics and complexity.

What constitutes the data that is classified as big data? Let us first explore the different categories of big data:

- Unstructured data – Text, videos, audio and images.
- Semi-structured data – Emails, earnings reports, spreadsheets, software modules.
- Structured data – Sensor data, machine data and mathematical model outputs.

If you take a step back and closely observe those categories of data, there are a few common characteristics that need to be understood by the data architecture and the business teams:

- Volume – Any of these data categories are, by default, large in volume and variable in size.
- Variety – The data can be available in a variety of formats, languages and sources.
- Ambiguity – The data can contain metadata about itself or have no metadata.
- Quality – The quality of data in the unstructured and semi-structured categories is unreliable.

These characteristics make the acquisition and processing of big data an extremely complex activity. The biggest threat in this process is to prove the associated value from integrating big data. What does a big data program look like? What are the differences in such a program from any other data management program? Why do you need to pay attention to this initiative? Today, all of these questions are arising from teams within enterprises. But beyond all of these questions lies a hidden area that needs to be a pillar for the big data initiative: data governance. Data governance for big data is a large and evolving subject and each enterprise will treat the subject according to its requirements.

Today, big data programs in organizations are tied to exploring the potential of a platform, such as Hadoop, and the associated business case includes social media data integration, some weblog parsing and, in other cases, machine learning exploration. The underlying value of these exercises is quantified in ROI of sales lift, market share and customer centricity, but the critical path is understanding the data and its relevancy to the business. Before you put a business case document together, spend some time trying to understand the data and its content, in order to comprehend the value that can be derived. This is where the data governance aspect comes into play.

Data governance concepts as applied to traditional data warehousing and business intelligence include:

- Stewardship
- Information governance
- Data definition and usage standards
- Master data management
- Metadata management
- Data lifecycle management
- Risk and cost containment

These same concepts can be applied to the world of big data, with some tweaks:

Stewardship – Big data needs strong stewardship from inception to delivery and beyond. In any enterprise, business users know the data best from a utilization perspective and they should be the stakeholders for stewardship of data that is relevant to their needs. For example, Twitter data can help marketing and competitive research. – In that case, who is the primary data steward? Sensor data and machine data, on the other hand, needs to be managed by a team of data scientists who can be aligned with a business steward, but they need to hold joint custody of the data. Bottom line is

that, without strong advocacy and support, an enterprise will not have success with big data.

Information governance – Big data, if not governed with the right approach to managing information within the enterprise, can wreak havoc. The critical stakeholders in this initiative are architects, system administrators and data center administrators. Governance in this area will include the acquisition, landing, processing requirements, infrastructure management, storage and security of big data. Each of these aspects needs to be defined clearly as a part of the initial journey into big data.

Data definition and usage standards – The biggest threat with big data is the definition of the content and how to consume it. For instance, a tweet like, "@johndoe #united #fail bad svc. long waits. faulty planes. lousy mgmt. never fly again," can be processed in multiple ways. You need to define in a well-defined standard and structure the data within the tweet and how to format and process it for each division of the enterprise that will share this data. The consumption layer will include the use of MDM and metadata programs.

Master data management – With big data, there is no net new master data to be added to the "golden record." Big data is an analytical consumer of the MDM outputs, where the keys from the MDM program can be used to parse the content and form associations of big data into the enterprise. These processes are complex and require multiple iterations of processing before you can derive a finite value.

Metadata management – With big data, there is metadata associated with the object in most cases but there is not content-related metadata. To process metadata for the content and derive the appropriate contexts for the data, we can use taxonomies, semantic libraries and ontologies. These data elements may be new to the enterprise and need to be treated as lookup data and reference data for processing, and maintained as such. The benefit of this processing technique is very straightforward: It improves the data quality of big data.

Data lifecycle management – Big data needs lifecycle management principles too. Enterprises must to define the length of time for which the data is needed online and then define the archiving and storage strategies for this data post-consumption. Even platforms like Hadoop provide a mechanism for reducing the burden of a large volume of data being in the "namenode."

Risk and cost containment – While big data can benefit enterprises, there are hidden risks in the acquisition and processing of big data, which can be mitigated by implementing a governance process for infrastructure, process and consumption of the data. The biggest risk that we see unfolding is the incorrect association of the data and its subsequent processing, which can be avoided with stewardship and standards.

As we can infer from this brief discussion, data governance plays an extremely important role in big data and its management. If implemented early on and in the right organizational process, enterprises will reap rewards with big data processing and reduce any visible risk associated with such a program from a cost perspective. This is the beginning of a long journey for many organizations, and following a known path will bring improved adoption and success in the big data world.

In his 21 plus years of professional experience, Krishnan has been solving complex solution architecture problems around data warehousing and business intelligence for Fortune 1000 clients across the world. He is the president of Sixth Sense Advisors Inc., a Chicago-based company providing Independent analyst services in big data, data warehousing and business intelligence, and serves as Senior Vice President of Innovation for CBIG Consulting, a consulting firm focusing on data warehousing and business intelligence. He co-authored Building the Unstructured Data Warehouse along with Bill Inmon. He also has written eBooks, more than 150 articles, viewpoints and case studies on big data, business intelligence, data warehousing and data warehouse appliances, and high performance architectures. Reach Krish at krish.krishnan@cbigconsulting.com.