What Is Big Data and What Does It Have to Do With IT Audit?

For many years, IT auditors have been able to rely on comparatively elementary data analysis tools to perform analyses to draw conclusions. With the recent explosion in the volume of data generated for business purposes (e.g., purchase transactions, network device logs, security appliance alerts), current tools may now not be sufficient. By necessity, big data uses data sets that are so large that it becomes difficult to process them using readily available database management tools or traditional data processing applications. The paradigm shift introduced by big data requires a transformation in the way that such information is handled and analyzed, moving away from deriving intelligence from structured data to discerning insights from large volumes of unstructured data.

There is a lot of hype and confusion regarding big data and how it can help businesses. It feels as if each new and existing technology is pushing the meme of “all your data belong to us.” It is difficult to determine the effects of this wave of innovation occurring across the big data landscape of Structured Query Language (SQL), Not Only SQL (NoSQL), NewSQL, enterprise data warehouses (EDWs), massively parallel processing (MPP) database management systems (DBMS), data marts and Apache Hadoop (to name just a few). But enterprises and the market in general can use a healthy dose of clarity on just how to use and interconnect these various technologies in ways that benefit business.

Big data not only encompasses the classic world of transactions, but also includes the new world of interactions and observations. This new world brings with it a wide range of multistructured data sources that are forcing a new way of looking at things.

Much of the work involved in conducting IT audits entails inspection of data generated from systems, devices and other applications. These data include configuration, transactional and raw data from systems or applications that are downloaded and then validated, reformatted and tested against predefined criteria.

With the sheer volume of data available for analysis, how do auditors ensure that they are drawing valid conclusions? What tools do they have available to help them? According to a report from Computer Sciences Corporation (CSC), there will be a 4,300 percent annual increase in data generation by 2020.1 Currently, a one terabyte (Tb) external drive costs around US $80. It is very common for even medium-sized enterprises to generate one Tb of data within a short period of time. Using Excel or even Access to analyze this volume of data may prove to be inadequate. More powerful enterprise tools may be cost-prohibitive for many audit firms to purchase and support. In addition, the training time and costs may also prove to be excessive.

Transactions generated as a result of common business events, such as purchases, payments, inventory changes or shipments, represent the most common types of data. Also, IT departments increasingly record events related to security, availability, modifications and approvals in order to retain accountability and for audit purposes. IT departments also record more system-related events to enable more effective support with smaller staffs. Firewalls and security appliances log thousands of events on a daily basis. Given the sheer volumes of data, these security-related events cannot be manually analyzed as they were in the past. Marketing teams may record events such as customer interactions with applications, and larger companies also record interactions between IT users and databases.

There has been significant growth in the volume of data generated by devices and by smartphones and other portable devices. End users and consumers of information generate data using multiple devices, and these devices record an increasing number of events. The landscape has evolved from an Internet of PCs to an Internet of things. These things include PCs, tablets, phones, appliances and any supporting infrastructure that underpins this entire ecosystem.

Kumar Setty, CISA, has more than 10 years of experience in the areas of data analysis, auditing and computer security. He is a manager at Grant Thornton LLP.

Rohit Bakhshi is a product manager with Hortonworks. Prior to joining Hortonworks, Bakhshi was an emerging technologies consultant at Accenture where he worked with Fortune 500 companies to incorporate big data technologies within their enterprise architecture.

Do you have something to say about this article?

Visit the Journal pages of the ISACA web site (www.isaca.org/journal), find the article, and choose the Comments tab to share your thoughts.

Go directly to the article:
Few of these new types of data were utilized or even considered in the past.

**PAST AND PRESENT**

Enterprise IT has been connecting systems via classic extract, transform, load (ETL) processing (as illustrated in step 1 of figure 1) for many years to deliver structured and repeatable analysis. In step 1, the business determines the questions to ask and IT collects and structures the data needed to answer those questions.

The big data refinery, as highlighted in step 2, is a new system capable of storing, aggregating and transforming a wide range of multistructured raw data sources into usable formats that help fuel new insights for the business. The big data refinery provides a cost-effective platform for unlocking the potential value within data and discovering the business questions worth answering with the data. A popular example of big data refining is processing blogs, clickstreams, social interactions, social feeds, and other user- or system-generated data sources into more accurate assessments of customer churn or more effective creation of personalized offers.

There are numerous ways for auditors to utilize the big data refinery. One instance is analysis of logs generated by firewalls or other security appliances. Firewalls and security appliances commonly generate thousands of alerts per day. It is unlikely that a group of individuals would be able to manually review these alerts and form meaningful associations and conclusions from this volume of data. Auditors could collaborate with IT to determine predefined thresholds to flag certain types of events and could even formulate countermeasures and actions.

---

**Figure 1—Maximizing the Value of Data**

- **Big Data Refinery**
  - Store, aggregate and transform multistructured data to unlock value.
  - Retain run-time models and historical data for ongoing refinement and analysis.

- **Business Transactions and Interactions**
  - Web, Mobile, CRM, ERP, SCM, etc.

- **Business Intelligence and Analytics**
  - Dashboards, Reports, Visualization, etc.

- **1. Classic ETL Processing**
- **2. Store, aggregate and transform multistructured data to unlock value.**
- **3. Share refined data and run-time models.**
- **4. Retain run-time models and historical data for ongoing refinement and analysis.**
- **5. Retain historical data to unlock additional value.**

to respond to such events. A centralized logging facility to
capture all security events could also be utilized to relate certain
types of events and assist in drawing conclusions to determine
appropriate follow-up actions.

Another potential use for the big data refinery is in
fraud analysis of large volumes of transactional data.
Using predefined criteria determined in collaboration with
other departments, the big data refinery could flag specific
transactions out of a large population of data to investigate for
potential instances of fraud.

The big data refinery platform provides fertile ground for
new types of tools and data processing workloads to emerge
in support of rich, multilevel data refinement solutions.

With that as a backdrop, step 3 of figure 1 takes the model
further by showing how the big data refinery interacts with
the systems powering business transactions and interactions
and business intelligence and analytics. Complex analytics
and calculations of key parameters can be performed in
the refinery and flow downstream to fuel run-time models
powering business applications, with the goal of more
accurately targeting customers with the best and most relevant
offers, for example.

Since the big data refinery is great at retaining large
volumes of data for long periods of time, the model is
completed with the feedback loops illustrated in steps 4
and 5 of figure 1. Retaining the past 10 years of historical
Black Friday\(^2\) retail data, for example, can benefit the
business, especially if it is blended with other data sources
such as 10 years of weather data accessed from a third-party
data provider. The point here is that the opportunities for
creating value from multistructured data sources available
inside and outside the enterprise are virtually endless with a
platform that can perform analysis in a cost-effective manner
and at an appropriate scale.

Enjoying this article?

- Read ISACA’s white paper Big Data: Impacts
and Benefits.
  www.isaca.org/big-data-wp
- Learn more about and discuss big data in the
Knowledge Center.
  www.isaca.org/knowledgecenter

A next-generation data architecture is emerging that
connects the classic systems powering business transactions
and interactions and business intelligence and analytics with
products such as Apache Hadoop. Hadoop or other alternate
products may be used to create a big data refinery capable of
storing, aggregating and transforming multistructured raw
data sources into usable formats that help fuel new insights
for any industry or business vertical.

One key differentiator for enterprises is the ability to quickly
yield and act promptly upon key insights gained from seemingly
disparate sources of data. Companies that are able to maximize
the value from all of their data (e.g., transactions, interactions,
observations) and external sources of data put themselves in
a position to drive more business, enhance productivity, or
discover new and lucrative business opportunities.

Emerging techniques allow auditors to draw key
conclusions from a wide range and large population of data
sources (internal and external). These conclusions or insights
may reflect changes to the overall risk profile, new risk factors
to the enterprise and specific internal risk factors such as
material misstatement to financial reporting, fraud risk and
security risk.

ENDNOTES
1 Computer Sciences Corporation (CSC), 2012
2 The day after the US Thanksgiving Day holiday, a major
shopping day in the US