Big data business intelligence analytics

Peter J. Jamack

November 20, 2012

Customers and the enterprise world are changing. It's no longer just about customer sales data. The goal of an integrated business intelligence and big data analytics platform is to dig deeper into the why, where, what, and how of a customer, product, and company. This article discusses integrating business intelligence and big data analytics.

The enterprise world is changing. The customer dynamic is changing. Customers are changing. And everyone is in a race to figure out why instead of just who or how. It's no longer okay just to understand how a company got from point A to point B. Enterprises are in a real-time competitive fight to know when customers buy something, where they buy, and what they are thinking before they ever walk into a store or visit a website. The power of big data, big data analytics, and an integrated business intelligence (BI) and big data analytics platform can help.

Big data analytics is young, and agile BI is a new concept. How do you integrate these similar but different concepts? It is not just about data or technology, it's everything—social media, customer behaviors, and customer segmentation, to name just a few. You can't plug in some big data appliance and expect to see the future. BI, master data management (MDM), big data, and analytics must be integrated into one platform, and rolled up into one visually innovative solution.

BI and big data analytics similarities and differences

BI is not a new concept. Data warehouses, data mining, and database technologies have existed in various forms for years. Big data as a term might be new, but many IT professionals have worked with large amounts of data in various industries for years.

However, now big data is not just about large amounts of data. Digging and analyzing semi-structured and unstructured data is new. Fifteen years ago, we didn't analyze email messages, PDF files, or videos. The Internet was just a fad; distributed computing was not created yesterday, but being able to distribute and scale out a system in a flash—and within smaller budgets—is new. Similarly, wanting to predict the future isn't a new concept, but being able to access and store all the data that is created is new.

Various sources claim that 90 percent of the data that exists today is only two years old. And that data is growing fast. If 90 percent of all the data in the world was created in the past two years, what does that say about the data?
Many enterprises have multiple databases and multiple database vendors, with terabytes or even petabytes of data. Some of these systems accumulated data over 30 or 40 years. Many enterprises built entire data warehouse and analytic platforms off this old data. Large retail corporations, such as Wal-Mart, became billion-dollar companies long before big data. So, it wasn't data that drove their business.

Data as a service can drive a business, though. Think of Amazon. It was an online e-commerce product company. Now, people look at Amazon as a platform as a service, software as a service, big data as a service, and cloud data center company. Amazon built an incredible recommendation engine over the years from various open source technologies. Zynga, the Facebook gaming company that is known for hits like Farmville, used Amazon's cloud services to scale its own databases and analytics.

For data to be useful to users, it must integrate customers with finance and sales data, with product data, with marketing data, with social media, with demographic data, with competitors' data, and more.

**Challenges in building a BI and big data analytics system**

Designing an integrated platform is never easy. Extract, transfer, and load (ETL) is always the longest phase in data warehouse projects. There are various ETL best practices, sometimes they work, sometimes they don't. If ETL is not done well, you suddenly have incorrect and mistrusted data. Mistrusted data becomes a mistrusted and unused system. Nobody wants that.

You might think something like a product database would be easy. But it becomes a game of versions, bugs, updates, different releases, different release cycles, different licenses, and different licenses that are based on location. And that's just in a company with a few products. It becomes more complicated for retail companies that have thousands of different products.

Integrated BI and big data platforms might have unstructured data from email messages. They might include semi-structured data from logs. Email systems might be dispersed among various databases in multiple data centers across the globe. Add a few firewalls, and suddenly moving data from one place to another becomes a logistical nightmare, a project in itself. System logs might be unformatted, semi-formatted, or a mess—another project in itself.

There's a reason why big data technologies like Apache Hadoop encourage moving the system to where the data is instead of moving the data to the system. It takes time to move data across network lines, between firewalls. You lose data, packets, files. Trust becomes a big problem.

A core concept of noSQL and Hadoop is to move the application to the data, except that it's not that simple. If you have 100 different systems, do you add 100 instances of the same application to each system? Although several people might assume that they mastered MDM, no one has. When you have one product MDM, a sales MDM, and a customer MDM that do not integrate or join easily, adding an application to each system does not suddenly integrate or join any of them. It remains a system with many silos that nobody can connect.
Even if an enterprise installed a big data application onto a perfect platform that integrated and connected all the different forms of data, there would be major issues. The truth is, you cannot suddenly run complicated algorithms on a live system that users are using. It might fail. It might slow down the performance. It might screw up the data. There might be security issues. Installing an application that requires a large amount of space, memory, and speed might cause an old system to fail. It might not even work properly on these old systems. If it did, is it any different from existing, non-connected MDM or BI systems?

A BI and big data analytics platform must be innovative. It must be next generation. It must use in-memory technologies or configure a system to use tools like Hadoop and Apache Cassandra as the staging area, the sandbox, the storage system and be a new and improved ETL system. It must integrate structured, unstructured, and semi-structured data. There are many pieces to the puzzle.

Solutions

An integrated BI and big data analytics platform is a different system. You have build versus buy options from which to choose. You must consider existing systems, use cases, and the experience levels and competence of your staff. Some companies might want to build an entire open source system using nothing more than vanilla Hadoop (the Hadoop Distributed File System [HDFS] and MapReduce), Zookeeper, Solr, Sqoop, Hive, HBase, Nagios, and Cacti, whereas someone else might be looking for more support and try and build an system using IBM® InfoSphere® BigInsights™ and IBM Netezza. Other companies might want to separate structured data and unstructured data and build a graphical user interface (GUI) layer for users, power users, and applications.

It really depends on the company. And it is not just a plug and play system. Whether going with build or buy, there are multiple pieces at every level.

ETL

ETL, data ingestion, and all the processes that are involved are always a significant first step, second step, third step, and more. You cannot dump a big data application onto a transactional system and expect things to work without degrading that original system, or expect it to integrate well with anything but the system in use. Therefore, some data ingestion into Hadoop or any other noSQL system or massively parallel processing (MPP) data warehouse is necessary. There are various tools and methodologies to follow, and much it depends on the systems, the sources, the data, the size, and staff.

You might start with something like Sqoop. It is a great tool for ingesting data from relational database management systems. Adding other open source tools like Flume or Scribe can help with logs. There are also ETL tools like Talend or IBM InfoSphere DataStage®, both of which now have big data integrators. These tools are more visual and don't require a PhD in computer science to build the infrastructure. Both tools provide technical documentation, updates, and GUI visual tools; they are always being improved; and they are in use across many industries and enterprises.
Some companies prefer open source only. Other companies might have many systems that are built on various IBM products. Obviously, integrating what is already in use with new technologies is a significant consideration.

It's time-consuming building your own ETL system, and doing so can be heartbreaking if the result doesn't do what you need it to do. Hadoop has many pieces, and you might need more than Sqoop. Integrating and adding multiple pieces can be painful, especially if you lack the experience and knowledge, or want to build your own ETL tool. The process requires time and patience. You might encounter interruptions too. You might use an open source tool that the community later dumps. Or, you might configure and develop your own ETL tool with various internal applications and open source tools, and then the open source community changes a few things or a few of your developers leave, and suddenly you have a system no one knows how to maintain or fix.

Wise enterprises look at their own staff, experiences, budgets, and potential and are realistic. For example, if an enterprise has a relatively small IT staff, looking at how a Google or Facebook built its systems is not a good idea. Do not compare your small IT shop with companies that have several servers and computer science graduates working on those particular infrastructures and systems. Sometimes, using cloud services or external staff might be the only option. Other times, big data appliances like Netezza are the best choice.

**Storage**

Data storage is a huge factor and might require that you use various technologies. In the Hadoop system, there is HBase. But some companies use Cassandra, Neo4j, Netezza, HDFS, and other technologies, depending on what's needed. HDFS is a file storage system. HBase is a columnar store similar to Cassandra. Many companies use Cassandra for more near-real-time analytics. But HBase is improving.

You might consider HBase or Cassandra when you want to use an open source database management system for big data analytics. As far as data warehouse platforms, Netezza is one of the top technologies in the analytics and BI industry. The best choice for big data integration is to use an integrated platform that consists of Hadoop and Cassandra for unstructured or semi-structured data and Netezza for structured data.

The IBM Netezza Customer Intelligence Appliance combines a few different technologies into one platform. At the top layer, that is the user layer, it relies on IBM Cognos® BI software, a business intelligence and reporting product. Cognos BI is an impressive product that many enterprises use for various BI and data warehousing needs. At the data warehouse storage layer, Netezza is great for its MPP database system. This system is geared toward structured data, but when you use Hadoop or Cassandra for unstructured and semi-structured data you create an integrated BI and big data analytics platform.

**The GUI**

At the GUI and user front-end layer, there are various other pieces to the system. Power users might use tools like IBM SPSS® Statistics, or R, for data mining, predictive modeling, machine learning, and building complex algorithms and models. Your everyday sales people might use...
something like Cognos for BI reporting, big data reporting, dashboards, and scorecards. A tool like Cognos is great for providing various kinds of users the opportunity to explore the data or view simple reports.

There are other pieces to the GUI and front-end layer, like machine learning tools (for example, Apache Mahout) or Apache Hive (for Structured Query Language), but those tools can also be a part of the infrastructure. The biggest factor is integrating the structured data and unstructured data as part of the BI and data warehousing and big data analytics infrastructure. Is it a service? Who are the users?

Users don't care about the infrastructure. They don't care if it is integrated. They only care if they are able to get the right data at the right time.

**Conclusion**

Integrating BI and big data analytics is no easy task. The goal for any data or analytical system is to make the data useful and available to as many users as possible. Big data appliances are one way to go. An open source Hadoop system is another way. Both require time, patience, and innovation.

An open source system is far quicker and less expensive to implement, but you need a staff with that experience. If you are not experienced in working with big data, a big data vendor appliance might be the better choice, though it is more expensive. Remember, not everyone wants to be a software or hardware company. Sometimes, building an integrated BI and big data platform requires a little building and buying to get where you must go.
Related topics

- Check out the blog, Big Data: The next frontier for innovation, competition, and productivity, from the McKinsey Global Institute.
- Visit the Data Warehouse Institute (TDWI), the premier resource for BI and data warehousing.
- Visit Hadoop.org for all things Hadoop.
- Visit HBase.org for more information about Apache HBase.
- The Hive project page provides the information you need to get started with Apache Hive.
- Sqoop is another Apache project that you'll want to get to know.
- Learn more about Mahout.
- Visit the Cassandra project page for all things Cassandra.
- Learn more about IBM Netezza.
- Learn more about SPSS Statistics.
- Learn more about InfoSphere BigInsights.
- Start developing with product trials, free downloads, and IBM Bluemix services.

© Copyright IBM Corporation 2012
Trademarks
(www.ibm.com/developerworks/ibm/trademarks/)