BIZCOMMUNITY

Why big data should be a priority at the moment

By Shane Moodley

11 Mar 2015

Throughout history there have been several eras of quantum leap innovations made by man. From the Stone Age, Bronze and Iron Ages, all have made their mark in our history and made their own contributions to our future. Today, with the advancement of technology and the emergence of the information highway, we are now living in what is called the 'data age'.

Shane Moodley

The reasoning for this is that as each millionth of a second passes, almost everything we see, hear and touch generates data. The use of PCs, mobile handsets, tablets, GPS devices, servers and sensors attached to vehicles, buildings and satellites leads to huge amounts of data being stored across multiple databases and data stores worldwide. The speed at which this data is generated is increasing rapidly and so is the volume and array of different structured and unstructured data types - commonly referred to as big data.

In theory, big data means having an abundance of data at your disposal. However, in practice, this data is useless if businesses cannot apply analytics to benefit and gain insight from it. Businesses must find a 'mechanism' that can perform collaborative filtering to 'net' vital information and trends and to answer key questions.

Apache Hadoop

But how easy is it to find this mechanism? Currently, frameworks have been developed to store and retrieve the data using MapReduce algorithms. Apache Hadoop is one of the frameworks of choice, and some of the large tech and social media companies have shown a keen interest in it.

Hadoop is popular because of the speed at which it can load the volume and variety of data compared with other Extract-Transform-Load (ETL) tools and frameworks. Hadoop can load data much quicker than relational databases because it has no gatekeeping rules. It stores the data on its Hadoop Distributed File System (HDFS) without needing to match the destination structure with the source structure - a relational database would require 'Table B' to be the same as 'Table A' in order to transfer data across. Instead, Hadoop stores the data across numerous data nodes using key-value pairs. Initially, querying data on HDFS was slow and involved a great deal of effort. To complete the task, a more SQL-like language was needed. However, over the years, many contributions have been made to this open-source software framework. Currently there are many 'SQL on Hadoop' languages, such as Apache Hive, Stinger, Apaches Drill and Spark SQL, which have improved data-retrieval performance considerably.

Furthermore, Hadoop's architecture is very scalable - this allows millions of servers to work together instead of running one high-specced server. Hadoop is also fault tolerant, storing multiple copies of each piece of data in different nodes. If one node goes down, the framework will automatically switch over to another in the most efficient manner.

In recent times, Hadoop has received a lot of buy-in from the big-player analytics and visualisations vendors. Cloudera is partnering NoSQL creator MongoDB, and Hortonworks has collaborated with Tableau. Other vendors, such as Qlikview, Spotfire and Microstrategy, have some Open Database Connectivity (ODBC) configuration to connect to Hadoop. Cloudera has released a universal ODBC driver that enables the connection of many applications, such as Teradata Parallel Transporter (TPT), Microsoft SSIS, IBM DataStage, Ab Initio, Informatica PowerCenter, SAP Data Services, Business Objects, OBIEE, Cognos, SAS, SPSS, Unica, Linked Server and Oracle Database Gateway.

Not taken off in SA

This all sounds very exciting and convincing; however big data has still not taken off in the South African market as initially predicted. This is evident in the fact that not many companies show demand for big data resources or have the necessary inherent skills to manage this internally. Global management consulting firm McKinsey & Company has predicted that by the year 2018, the shortfall of big data experts will be at anywhere from 140,000 to 190,000.

The good news is that tertiary institutions in many countries are now including big data in their curricula and postgraduate degrees, including South Africa, so we can expect the big data resource pool to grow over time as these students complete their studies. Until this happens, companies are investing in their current staff compliment to fill the gap. However, learning big data via frameworks, such as Apache Hadoop, is quite a steep learning curve compared with traditional, structured business intelligence involving SQL, ETL and OLAP tools, and to a certain degree, it requires a mindset shift.

There are, however, other options that could be explored. They can opt for a cloud solution like Microsoft's HDInsight hosted on their Azure platform. This option has its pros and cons. On the plus side, it collects and stores data at a reasonable price on an HDInsight cluster. There is no need to understand the background processing and architecture of Hadoop in order to do so. Microsoft has also developed a way to combine and query both non-relational data (HDInsight) with structured SQL Server Parallel Data Warehouse in the form of Polybase. This integrates perfectly with the latest version of Microsoft Excel and an assortment of PowerBI tools, such as PowerPivot and PowerQuery. One of the biggest drawbacks of this option is the cost involved in processing and retrieving the answers companies are looking for. There is also the added security risk of storing sensitive data on an off-site server, perhaps halfway across the globe. Implementing data governance can also be tricky with regard to unstructured data and defining which users are authorised to see which data.

Another option companies may choose is engaging with consulting companies that have already invested time, research and development in Hadoop and its architecture. Consultancy firms allow their staff to play around with cutting-edge technologies and often give them carte blanche to find innovative ways of solving tomorrow's business needs. This allows companies to utilise this expertise and knowledge without wasting unnecessary time and money.

Whichever option companies choose to invest in, it is important to note that big data is the essential component in every company's BI architectural framework. Due to the current skill shortage and reluctance to opt for a cloud solution, big data may not be a priority for companies at the moment, but it will definitely become one in the future. The quicker businesses get a big data solution up and running, the faster they will be able to gain a substantial leading edge over competitors. The race is on for companies to find the answers of tomorrow based on the data that was always there, but could never be analysed.

ABOUT THE AUTHOR

Shane Moodley is Entelect Team Lead of Data Solutions