TECHNOLOGY FOCUS

BIG DATA 101

DEMYSTIFYING THE ENABLING TECHNOLOGIES

Although the Big Data hype has recently invaded mainstream media and popular culture, Big Data has been on the technology radar for some time now. It's one of those terms that is hot and cool, and means different things to different people. While defining big data is in some ways minimizing its scope and scale, here are a few different ways of looking at it. One way to look at it is to examine the nature of data that are collected – its size and complexity. Another way is to look at the target population we are trying to represent. Yet another way is to look at the technologies that support it. It's not that we've not had to contend with large datasets in the past. It's also important to address the issues of validity as we extend a sample's interpretation to a population. However, all these concepts have only been made possible because we now have access to cloud computing. So, let's begin there.

Sameer Verma

In the mid-90s we would only look to a mainframe for any kind of large scale computing jobs. That changed when we got distributed computing over fast networks, whereby we could leverage several simpler computers to solve a complex job. It became the *de facto* method to employ large number of *"garden variety"* computers that would work together in a group to break down massive problems into smaller bite size chunks and solve them simultaneously. Advent of cloud computing gave this approach a boost because one did not have to own thousands of computers to crunch large data sets. Instead, one could lease the machines as and when needed.

MapReduce is one technology that features heavily in this space. MapReduce is an approach to manage and process large data sets in a distributed manner. This technology is in fact two techniques combined. There is the Map part that maps the tasks onto multiple nodes (computers) for processing, and the Reduce part that subsequently crunches the data to yield descriptive statistics. Together, MapReduce allows us to manage and process large data sets. Google is credited with bulk of the early work done on MapReduce, although in its own proprietary domain. The Google search engine employs MapReduce to process xxxbytes (a very large number) of data per minute on clusters of thousands of commodity-class Linux PCs. Nutch was one of the first projects to implement MapReduce as an open source project. A fair bit of this approach then found its way into Hadoop (also open source), a project that allows for managing the data through a distributed filesystem. Hadoop is by far the most widely used MapReduce platform. With the introduction of cloud computing, those thousands of nodes needed to run MapReduce could now be leased from Amazon, HP, Microsoft or any other cloud provider. One would simply pay for the duration of running the nodes in order to crunch large data sets.

In the mid-90s we would only look to a mainframe for any kind of large scale computing jobs. That changed when we got distributed computing over fast networks, whereby we could leverage several simpler computers to solve a complex job.

DATABASES

Now, let's switch layers to see what kind of database technologies are needed in this context. Traditional database technologies revolve around the concept of relational design with multiple tables and SQL to stitch it all together as and when needed. This works great and continues to do so for large transactional jobs such as those at banks, hospitals and universities. However, with systems such as those powering social media sites, where near-real time processing is important, we have seen a shift in the kind of data collected and analyzed. Imagine looking at a tweet from someone and clicking Like or Retweet. That kind of an activity produces a key-value pair type data where the Like or Retweet is paired with an identifier key. *So, a Retweet may look like* ***UserMeRetweet: https://twitter.com/UserMe/status/12365478912***

This is sometimes called clickstream data. This kind of data produces a single table with two columns, with several million rows. Features typically found in a relational database such as consistency, durability, etc. are usually not implemented in the database technology. Instead, these have to be managed elsewhere. Such an approach makes this kind of a database very lean, although the problems mentioned do have to be addressed elsewhere. This method is quite different from the one that employs several tables, with tens of columns and several million rows. SQL allows one to search and stitch data together. With clickstream, we are looking for aggregates, but really, there is nothing "relational" about it. So, we have different database technologies for this kind of an operation. These databases are popularly known as NoSQL databases. Several technologies exist in the field, each with different approaches to solving the problem of scale, and managing ACID compliance to some degree. MongoDB, Cassandra, CouchDB, Memcached, BigTable are to name a few.

Increasingly, due to technological advances, we now have situations where we do in fact have access to the entire population. The need to sample randomly or otherwise has diminished

Another technology that has found its way into NoSQL databases is Javascript Object Notation (JSON) serialization. This is an ▸▸

approach to address the structure of data and placement by serializing the data. Given the heavy use of Javascript in web development, both on the client side and the server side, JSON comes as a natural fit for NoSQL databases.

STATISTICS

Switching gears from the technology, let's take a look at the basics of the data crunching itself. Recall your basic probability and statistics class back in college and you would remember the importance of random sampling, normal distribution and the validities of studies based on data samples. Increasingly, due to technological advances, we now have situations where we do in fact have access to the entire population. The need to sample randomly or otherwise has diminished. We also have the technology to crunch it all in one go, and in near-real time. This leads to a marked shift in analytics where we no longer need to take a sample and draw inferences. We can use the entire population data and be descriptive.

Clickstream data, particularly in social media context tends to be for the entire population. For instance, at any given point in time, Twitter has access to all its clickstream data of Likes, Retweets, Quotes, Hashtags and Follows. They have NoSQL tables on the collective sentiment of the Twitterverse. They can slice it many different ways and tell you what's trending, what's related and so on. All we need here are the descriptive statistics and not much else.

LEVERAGE

So, what does this all mean for businesses? Big Data presents enormous opportunities for increased operational efficiency, greater customer intimacy and improved service delivery. Consider for a moment, how Amazon.com determines their customer demographics and preferences. They don't need to design a representative sample and conduct a survey, they simply analyze the behavioural, attribute and transactional data from their eCommerce site to know who their customers are, what they like, and what they are likely to purchase next.

Big Data presents enormous opportunities for increased operational efficiency, greater customer intimacy and improved service delivery

The technology is all there – the databases, the networks, the cloud – and it is very accessible and affordable even for the smallest businesses. Data can be stored and processed in more timely and meaningful ways and external data sources, such as social media are readily accessible. However, what really determines which technologies and techniques to use is still based on the nature of data and what you intend to do with it. Statistics 101 still reigns supreme. ■

Sameer Verma *is Professor of Information Systems at San Francisco State University*