Big data: what it is, its challenges and opportunities

By Dan Ourada

Whenever your credit card is swiped or inserted into a chip reader, that transaction begins its journey through the data collection process to becoming part of what is known as big data. A few 1s and 0s are transferred into the POI device, where more data are added — especially if there was a PIN entered or encryption performed. Those data move out of the POI via a cable attached to a computer, where it is further appended to other transactional data (invoice number, transaction amount, etc.), encrypted again for good measure, sent through a few more cables and blasted out of the merchant environment onto the public Internet enroute to a processor. The processor takes the data, decrypts the payload, parses the data, performs some validation, converts the data into a different format (or 2 or 3), packages it back up again, potentially routes it to another processor and finally to the card brands.

What Is Big Data
At each step in its journey, the data are annotated with more and more data. By the time the data make it to the processor, a few flakes have become a virtual data snowball. The annotation does not end at the processor. Processors link the transactional data to merchant data. Based on the merchant data, the location where the transaction originated is known. That piece of data can be annotated with census data to determine the socioeconomics of the area in which the merchant operates and, generally, the types of customers that the merchant serves. The merchant category classification (MCC) provides a way to categorize merchants into verticals. Multiple timestamps are added to the list: time of transaction, time sent from device, time of arrival at processor, etc. The geographic and temporal data can be matched to weather and other geographically pertinent news items that will further define the transaction.

The above scenario is just one of the seemingly innumerable ways that customer data can be collected by merchants. Data collection on this scale provides nearly untold opportunities for uncovering patterns, correlations, trends and customer preferences. The challenge is how to effectively analyze the data and generate actionable information — a challenge posed primarily by the three Vs of big data: volume (the amount of data), variety (the different formats in which date are stored), and velocity (the speed at which data are collected).

In the transaction example, the volume of data or the size of the "virtual snowball" is relatively small, and the number of formats the data are stored in is not significantly high. However, the number of transactions arriving every second is large (a back of the envelope estimate is ~800 transactions per second). The data over months and years and across hundreds of thousands of merchants obviously increases the volume.

The Untold Potential of Big Data
Anyone can intuitively grasp that having access to this amount of customer information is valuable, but there is another level to its potential that is less obvious. Statisticians attempt to infer truth via a specific sample size, but with big data, there is a trend toward full census. The answer *literally* is in the data. Large amounts of data have proven to shift models that were merely interesting and moderately useful into models that have significant real-world applications.

In 2001, Michele Banko and Eric Brill from Microsoft wrote a seminal work on using big data to improve natural language processing. In the beginning of their research into natural language processing, they were receiving error rates around 25 percent. With "cutting edge techniques" they were reach-

ing 19 percent error rates. Taking advantage of the enormous amount of text online, they used that massive amount of data and eventually obtained 5 percent error rates, which allowed real-world applications of the models. They also found that as data increased they were able to use traditional algorithms rather than newer algorithms that might have simply been overfitting the data.

The important takeaway from this is that as data increase (by orders of magnitude), predictive power increases, error rates drop and the subsequent products built on top of the data will lead to better decision making. Assuming the data are cleaned and the analysis is correct, more data equal a better product.

The potential applications of successful data analytics using truly big data in our industry are untold: Everything from fraud detection to detecting customer churn to external facing products that help customers analyze and/or market to their customers. The product list is large and limited only by human imagination.

The Future of Big Data
If you are a customer or partner of a processor you have likely already benefited from big data analytics. As tools, analysis and pipelines become more efficient, the time to market for big data analytic products decreases, resulting in better decision making for everyone.

There is of course a flipside to all of the opportunities that big data makes possible, namely privacy concerns. While there is tremendous value in the virtual snowball of data, there is also an obligation to handle the data with great care and consideration. As an industry, we must protect and secure all the data flowing through the system while working to deliver products that customers love.

That said, there is no denying that big data with all of its benefits and dangers will be an increasingly important factor in our industry. Your understanding of big data and, more important, your understanding of the products that sit on top of big data, will be an essential tool in building trusted-partner relationships with customers.

Dan joined Mercury (now Vantiv) in July 2004 and in his current role as technical evangelist he enjoys all the great things in life: product, technology, sales, and support.