Profisee Group, Inc.

Big Data: Business Value Driver or Detour?

Understanding when, and how, to put big data to work for your organization

Businesses are looking to augment traditional transactional and reference data with the broad-ranging intelligence potentially available from external big data sources. With big data tools and techniques, virtually unlimited volumes of raw data can be accommodated, quickly and redundantly, distributed across as many server nodes as required, with various mapping and reduction iterations that apply structure as the data is being read, i.e., "schema-on-read." For the most part, traditional data management tools and approaches that apply structure up front, i.e., "schema-on-write," are not ideal for the collection, storage, and initial mapping of the extreme high-volume, minimally- or unevenly-structured data available in the big data world. However, without master data management, big data remains largely amorphous and inscrutable, in terms of business application. MDM gives organizations the opportunity to maintain a current picture of important reference data, as well as a manageable, usable historical view to inform business analysis. With the addition of this structure added into the mapping/reducing process, big data result sets become contextualized as actionable information, turning facts into insight and intelligence.

Any discussion of "big data," and its applicability to today's business, must be founded on a shared definition of the term. You can find many definitions, implicit and inferred, in discussions of data management, analysis, and intelligence. Each will be correct, in some degree, depending on the context of the discussion. For our purposes, we will be looking at big data through the lens of operational applications in business; that is, as an information source whose value may be unlocked through the use of specific technologies, tools and approaches. We are focusing on the information aspect, as opposed to the environmental infrastructure, of big data; therefore, we need to identify what makes big data "big."

Volume is the first qualifier that comes to mind – and indeed, the challenge of managing the constant torrent of data flooding through social media channels provided the original impetus for the development of big data technologies such as Hadoop. Simply put, Hadoop is a set of algorithms that facilitate the storage of huge amounts of data, and allow for distributed processing of the data to provide efficiency and speed.

But volume alone does not serve to define big data. There is also an inherent concept of big data as being "raw," or largely unstructured, data. Where the Hadoop Distributed File System (HDFS) component distributes the data for storage across clusters of servers, the associated Map/Reduce component provides the ability to map the data across the storage nodes, and reduce redundant data from within the results. In this sense, at the time the data is being processed, the Map/Reduce algorithms serve to apply the structure that did not exist when the data was stored – what is often referred to as a "schema-on-read" approach.

Unlocking big data value

Most businesses are looking to augment their traditional transactional and reference data with the broad-ranging insights and intelligence potentially available from external big data sources. In order to do so, they need a way to further refine the structure of the mapped data so as to produce actionable information; that is, data that can be acted upon, based on context provided by the organization's operational data stores.

Looking at a typical made-to-order food delivery business will help us to see how this might work in practice. Although this is a simplified example, it will serve to illustrate key aspects of the opportunities and challenges inherent in embracing big data, and provide guidance to those evaluating whether big data should be part of their current business intelligence strategy.

We'll call our business "Pizza to Go" – PTG, for short. PTG, like any well-established pizza-delivery business, takes delivery orders for its product via phone, retail outlets,

online, and mobile apps. They employ robust point-of-sale (POS) systems and processes that capture name, phone number, delivery address and email information for orders that are placed by phone. Online orders through PTG's website capture the same type of information, on an order-by-order basis. PTG also encourages regular online customers to "make online ordering even easier and faster" by creating a profile where additional customer information will be saved; they employ these user profiles in their mobile ordering apps, as well.

The customer profile elicits all of the above information, and correlates it with optionally provided alternate phone numbers, delivery addresses and credit card information – even the customer's birthdate – all of which can serve to identify a specific returning customer down the road. Thus, if Jeri Jones has created a login, then later places a phone order while on vacation in another state, from an unknown phone number, for delivery to her grandmother's house, PTG will still be able to identify this as being an order from Jeri Jones, and add this information to her ordering history for purposes of analyzing customer buying habits.

Culling this individually identifying information from PTG's billions of order transactions and matching up each order with a unique customer record is a big data challenge in its own right, in terms of volume. Still, both the transactional data and the customer data have specific structure imposed prior to being stored – "schema-on-write", if you will – and can therefore be processed using traditional tools and approaches for mastering data. This can provide PTG with the intelligence to conduct targeted marketing and promotional activities that are based on documented customer buying patterns, without the need for a big data infrastructure. [1]

Identity stitching

For purposes of our example here, however, we will add another data source to the mix, and this data fits our big data definition perfectly: 1) There is a massive volume of data available, and 2) it is not structured in any way in which it can be readily processed with traditional data management tools, or integrated with existing customer information. This source is "social media" – specifically, for PTG's interests, dining reviews on Yelp, status updates on Facebook, and tweets on Twitter.

PTG would like to identify any of their customers who have posted on social media about a negative pizza dining experience over the past year, and send those customers a special promotional coupon for a free pizza from their local PTG locations.

[1] An actual case study of such an application, using Profisee's Master Data Maestro solution, is available here: For Leading Made-to-Order Food Chain, Maestro Delivers Customer Insight or at www.profisee.com/casestudies.

Schema-on-read

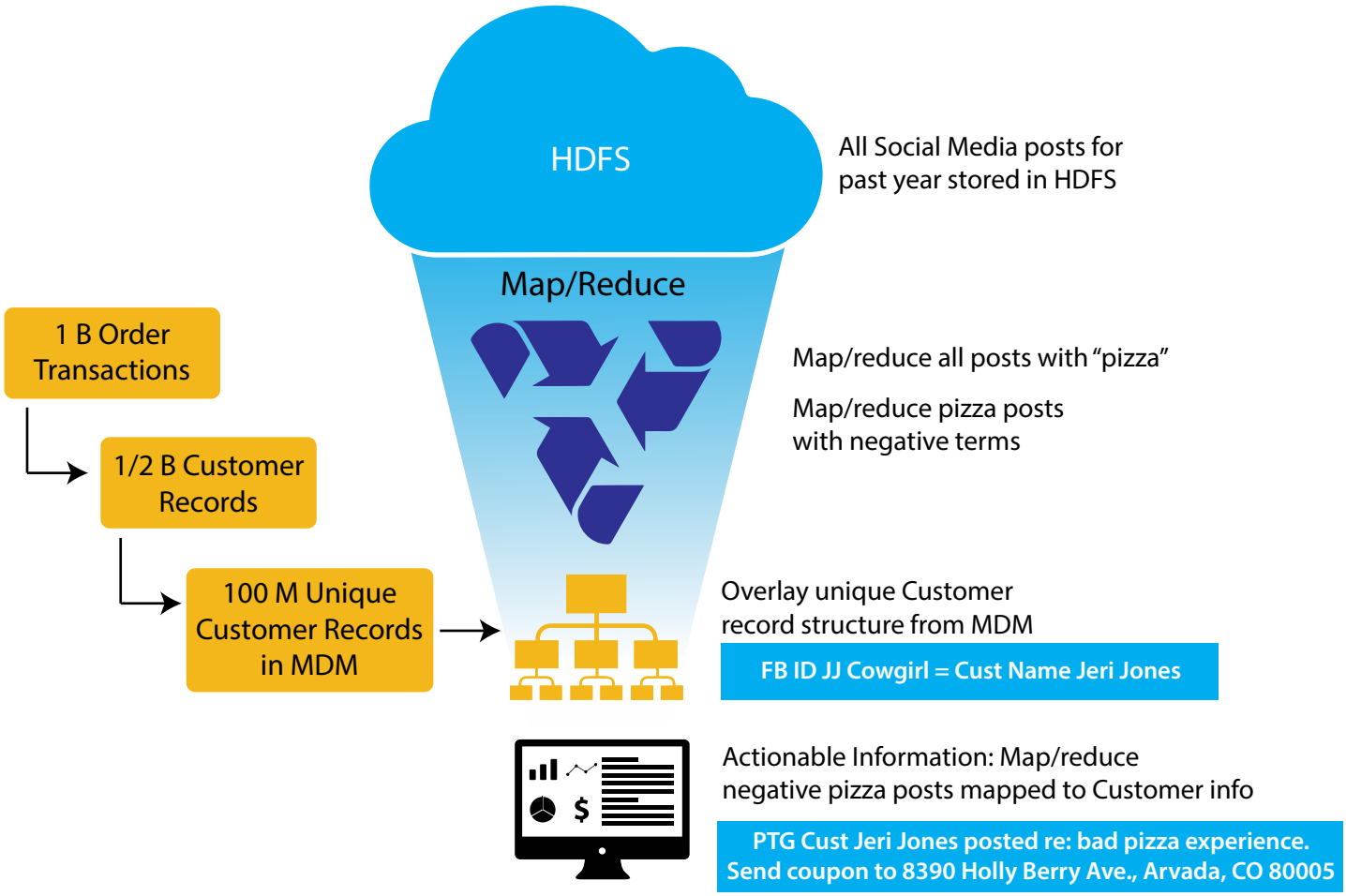
The first step in the process will be to distill from the available social media data anything that is potentially relevant to PTG's project, using the HDFS and Map/Reduce big data tools described above. Layers of structure are applied as late in the process as possible, in what is essentially a "schema-on-read" approach. This approach ensures that the big data tools and techniques are leveraged to full effect; virtually unlimited volumes of raw data can be accommodated, quickly and redundantly, distributed across as many server nodes as required, and the mapping function runs structured queries against the data as required for specific purposes. It is only in the various mapping and reduction iterations that structure is applied – that is, only as the data is being read. For example, the mapping algorithm establishes that [Name] is between character 7 and character 13 in this file; or, it is the second comma-delimited field, in this file. Using this minimal structure, name information is identified from all of the records, across all relevant data sources stored in HDFS server nodes. The reduce function then determines all the unique names from each block. It then has to further reduce the results, to eliminate the duplicates that arise from the redundant data storage across blocks within the node cluster.

In this example, the initial query had to do with posts mentioning "pizza", so that was the first round of mapping. Then, from those, the data had to be mapped to show what social media source it was from, the user name, the text of the post (tagging negative terms, like "soggy", "cold", "bland", etc.), date, and any other available identifying information, such as location.

Using big data tools, PTG collected the social media information, mapped it and reduced it to reflect posts that mention pizza, and that have the specific indicators classified as reflecting a negative experience. The results set represented two million records selected on the basis of the mapping and reduction algorithm, containing the social media ID and dates of the negative reviews. Now PTG knows how many people had a bad pizza experience, and shared it on social media. They could infer larger market trends from the big data – for example, if there were six million posts about pizza in all, it would seem to indicate that, in general, one out of every three people talking about pizza on social media have had a less-than-enjoyable experience. But they can't really act on that "insight."

The problem is the social media posts are not flagged in any way that ties them to PTG's customers or territories. It is data, but it is not actionable information. For example, in the results set is a Yelp review about soggy pizza, posted by JJ Cowgirl. It is not possible, from that information, to determine whether JJ Cowgirl is a PTG customer, much less to include her in a targeted marketing campaign. In order to make that correlation, PTG needs a way to inform their mapping and reduction algorithms with structure from their master data management (MDM) solution –

where customer Jeri Jones is linked with the social media ID JJ Cowgirl – to provide a roadmap for "stitching together" a meaningful customer identity from pieces of structured and unstructured data.



Providing the structure

PTG's mastered customer data, culled from around one billion transactions over the period of a year, reflects around 100 million customers – identifiable, unique individuals who purchased a pizza from PTG during that timeframe. These customer records have been mastered with address verification and geo-location information using best-in-class master data management tools and methodologies. As a result, all information gathered about a specific customer through any number of transactions and ordering channels is included in the master record for that customer.

For example, Jeri Jones's ordering profile and related information might look something like this:

ACTIVITY	INFORMATION GATHERED
February: Phone order for delivery to home address	Home phone number, street address
April: Jeri "Likes" and "Shares" PTG ad posted on FB to enter a contest for free pizzas	FB ID, associated phone number & email address
June: Jeri creates ordering profile online	Name, home phone, cell phone, home address, credit card info
July: Jeri uses ordering profile through PTG's mobile app to order pizzas for delivery to work address	Mobile device ID, work address, corporate credit card info

The records of Jeri's four order transactions with PTG were put in a match group through the MDM process, providing PTG with a master customer record for Jeri Jones that included the following information:

Name:	Jeri Jones
Phone:	321-555-5432
Cell:	321-555-0379
Mobile Device:	DYUBFTDST6FY
Email:	jjones@provider.com
Home:	8309 Holly Berry Ave., Arvada, CO 80005
Work:	120-9 Industrial Complex, Suite A, Westminster, CO 80021
FB ID:	JJ Cowgirl
CC1:	xxxxxxxxxxxx4284, Exp. 08/19, CSV: 893
CC2:	xxxxxxxxxxxx3905, Exp. 03/17, CSV: 3275

Address standardization and geo-location information in the customer records is further used to assign territory attributes that identify which store serves the area in which Jeri lives and orders pizza. This geo-location provides for location mapping of the company's restaurants, and reveals the location(s) of restaurant(s) the customer has shopped at before, which allows the data to be used to target marketing within specific restaurant/sales territories.

This master data structure is passed to the big data mapping/reducing process, laid onto the job framework, to identify the specific PTG customers who are among those two million who have posted about a negative pizza experience. Adding the master data identity information on top of the mapping/reducing process enables the process to determine whether a post was made by someone PTG knows about, by comparing the information about the social media post to information in the customer master

record – email, Twitter handle, FB ID, and so on; any identifying attributes that have been associated with the PTG customer master record. The structure created from PTG's master data is handed off to the mapping/reducing process to incorporate on top of its other structure-applying algorithms, essentially instructing the process, "When this ID is encountered, find its parent in the MDM table of identities, and if it has a parent, apply that to the data structure."

With this additional structure, added in as late as possible in the process to inform the map/reduce process, PTG receives a report of actionable information – the half million identifiable PTG customers who have had a bad pizza experience and shared it on social media within the past year. Among these is Jeri Jones, who Yelp'ed about soggy pizza using her FB ID, JJ Cowgirl. She, and all the others, will each receive a coupon for a free pizza from PTG.

MDM value in a big data world

For the most part, traditional data management tools and approaches are not ideal for the collection, storage, and initial mapping of the extreme high-volume, minimally- or unevenly-structured data available in the big data world. However, without master data management, big data remains largely amorphous and inscrutable, in terms of business application. MDM gives organizations the opportunity to maintain a current picture of important reference data, as well as a manageable, usable historical view to inform business analysis. In the case of PTG, their MDM solution allows them to turn the big data facts – "there were two million unhappy pizza consumers posting about it on social media last year" – into information that informs PTG's marketing initiative – "these are the half million unique, identifiable PTG customers who had a bad pizza experience and posted about it over the past year." Armed with this actionable, specific intelligence, the company can send each of these PTG customers a coupon for a free pizza.

Integrating big data into business operations

Having an integrated MDM solution with their CRM implementation, PTG can gain even greater value from big data by using their CRM system to track the resulting coupon campaign to determine its uptake, impact on consumption patterns, regional variations, etc. This would not be possible without applying the MDM structure to the greater pool of big data, thus identifying from within it that information which relates to what is known and tracked within the operations of the business, and can therefore be acted upon.

For example, PTG managers can choose to mount big data-informed campaigns or initiatives for specific geographic regions. With the integration of PTG's customer

geo-location data, the company's sales territory map, and CRM campaign-tracking functionality, a specific slice of the half million disappointed pizza consumers from the social media data can be targeted within specific geographies, and the results evaluated and tuned to greatest effect. In addition, ongoing data quality and master data maintenance processes mean that geo-location information and contact information remains up-to-date and accurate, not requiring additional big data processing to maintain actionable information.

Is it time for you to jump into the big data pool?

With all the discussion and hype about big data, many organizations are looking to gain the advertised advantages of incorporating big data into their strategic information management and business intelligence plans. The potential business insights inherent within big data offer a real temptation, if not an outright obligation, for businesses to jump into the pool. But are you in trouble if you're not yet embracing big data? The answer lies in your current data management foundation.

It may be true that, if you are not yet incorporating big data into your intelligence resources, you may be missing some analytical opportunities that you might otherwise take advantage of. But it is absolutely certain that, without a solid master data foundation for managing your traditional reference and operational data – customers, products, locations, and so on – you will miss out on key strategic intelligence and business drivers, with or without big data. And it is equally true that, without that foundation, big data will be a step too far, because there's no value big data can drive for your business unless you first have a handle on what your business is.

Here are three key questions to answer when evaluating whether you are ready for big data:

1. **Can I currently provide my business users and analysts with a single, consistent and accurate view of our customers/products/facilities/etc.?**

   If you are still struggling to reconcile different versions of data across your organization, you will not be able to readily reconcile the knowledge you have in-house with the facts contained in big data, and you will not be able to access any insight that the big data might hold for you. For example, if you do not already have a solid master data management solution in place, you may not realize that the various records you have for Jeri Jones and Jeraldine Jones and J. Jones are all really one customer; or know that only one of the three different addresses you have on file for Jeri is correct and current. Without having reconciled and consolidated – essentially, mastered – the data you have on Jeri, you will be hard put to provide that structure to your big data processes, and without it, big data will be unable to provide you with actionable information.

2. **Are the various business and infrastructure systems across my organization effectively sharing data?**

   Through business reorganizations, corporate acquisitions and mergers, even the development of one-off operational solutions, as organizations grow in size and complexity, they frequently inherit disparate operational systems, all relying on the same corporate data – which creates attendant data management issues. A robust master data management solution provides an important foundation for enterprise data integration and management, oversight and, when needed, control of data across these disparate systems. This ensures that the data structure being provided to your big data process will reflect an enterprise-wide view of your business data assets, and incorporate knowledge gained through all business interactions, regardless of which division, business unit, or transactional system is the originator of the business data.

3. **Can I be certain that all business units, systems, and initiatives within my organization are working with the latest and most accurate versions of my reference and operational data?**

   Once data has been mastered and harmonized across your enterprise systems, a master data management solution will ensure that the data is kept in sync. An MDM solution with an integrated event management system will allow organizations to accomplish this ongoing synchronization in near real-time. As new data is acquired from various source systems, through new business transactions, customer interactions, product development, and so on, the synchronization features of the MDM solution are key to ensuring that this new information is appropriately incorporated and shared among all enterprise systems, so that all data intelligence initiatives, including those that incorporate big data sources, are informed by the latest, most accurate corporate information, regardless of the initiating entity – business unit, operational department, etc. Trying to ensure this sort of enterprise-wide data synchronization on an ongoing basis without MDM is virtually impossible, and absolutely crucial to mining the real value in any big data effort.

If the answer to any of these questions is no; not really; sometimes; then you should consider building a solid MDM foundation before you jump into the big data pool. It's clear that, for many large organizations, big data has a great potential to provide business value. It's equally clear that its value will only be as great as your ability to inform your big data mining efforts with current, accurate, reliable business data – the kind of knowledge built and maintained with a strong MDM solution in place.