Big Data: Redefining Travel Business Decision Making

Written and Researched by
Bob Offutt

Sponsored by
UNIT4 BUSINESS SOFTWARE
An Introduction from UNIT4 Business Software

Turning Big Data into a Competitive Advantage

Across all industries, discussions around Big Data continue, as does the confusion around what the term really means and how organizations can take advantage of the phenomenon. While CIOs and CTOs have been talking about how to garner the power of their own data for years, it’s only now that we are starting to see organizations put that talk into practice. Concurrently, we are seeing organizations fall behind as they fail to harness that power and miss opportunities. The following paper from PhoCusWright offers insight into how this hot topic can translate to real business strategy in the travel industry.

The turbulent business environment of the modern day Travel Management Company (TMC) intensifies the need to take Big Data seriously. Many kinds of change, some anticipated and some wholly unexpected, can create challenges significant enough to compromise a TMC’s competitiveness and limit their ability to strategize into the future. While other agencies are still reeling from a new industry change, successful TMCs need to be able to focus on what the change really means and what opportunities it could present. With the right tools they can then turn the data they’ve gathered into information targeted to the strategic growth of the organization. This enables them to take advantage of industry flux to get out in front of the pack.

At UNIT4 Business Software we’ve seen our customers take advantage of our travel back-office software, Sabre® CentralCommand with Agresso™, to truly become the Intelligent Organizations mentioned in this paper. Their reporting is done in real-time; it does not require a data handoff and there is no need to scrub the data or reconcile it with other data sources. As you will read, it takes a combination of this type of reporting power, and the foresight to leverage it, to truly achieve your goals in today’s environment.
About PhoCusWright

PhoCusWright is the travel industry research authority on how travelers, suppliers and intermediaries connect. Independent, rigorous and unbiased, PhoCusWright fosters smart strategic planning, tactical decision-making and organizational effectiveness.

PhoCusWright delivers qualitative and quantitative research on the evolving dynamics that influence travel, tourism and hospitality distribution. Our marketplace intelligence is the industry standard for segmentation, sizing, forecasting, trends, analysis and consumer travel planning behavior. Every day around the world, senior executives, marketers, strategists and research professionals from all segments of the industry value chain use PhoCusWright research for competitive advantage.

To complement its primary research in North and Latin America, Europe and Asia, PhoCusWright produces several high-profile conferences in the United States and Germany, and partners with conferences in Canada, China and Singapore. Industry leaders and company analysts bring this intelligence to life by debating issues, sharing ideas and defining the ever-evolving reality of travel commerce.

The company is headquartered in the United States with Asia Pacific operations based in India and local analysts on five continents.

PhoCusWright is a wholly owned subsidiary of Northstar Travel Media, LLC.
## Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td>5</td>
</tr>
<tr>
<td>Big Data vs. the Traditional Data Warehouse</td>
<td>6</td>
</tr>
<tr>
<td>Technology</td>
<td>7</td>
</tr>
<tr>
<td>Traditional Capabilities Are Being Challenged</td>
<td>9</td>
</tr>
<tr>
<td>Business Scenarios</td>
<td>10</td>
</tr>
<tr>
<td>Business Examples (Travel)</td>
<td>11</td>
</tr>
<tr>
<td>Business Examples (Non-Travel)</td>
<td>12</td>
</tr>
<tr>
<td>Vision for the Travel Industry</td>
<td>12</td>
</tr>
<tr>
<td>What You Should Do</td>
<td>13</td>
</tr>
</tbody>
</table>

## Table of Charts

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 1</td>
<td>5</td>
<td>The Growth of the Digital Universe</td>
</tr>
<tr>
<td>Figure 2</td>
<td>6</td>
<td>Respondents by Industry Segment</td>
</tr>
<tr>
<td>Figure 3</td>
<td>7</td>
<td>Example of How MapReduce Works</td>
</tr>
<tr>
<td>Figure 4</td>
<td>9</td>
<td>Corporate Data Topology</td>
</tr>
</tbody>
</table>
Introduction

For the purpose of this Analysis, Big Data is defined as large amounts of data that cannot be processed with traditional tools and technology such as relational databases and normalized structures. The amount of data that is considered “big” ranges from tens to hundreds of terabytes to petabytes. Key to Big Data is the amount available to support business processes and the cost to store the data. Figure 1 illustrates the projected growth in global data sharing through 2015. At the same time the cost of storage is expected to drop from $18.95/GB in 2005 to $.66/GB in 2015, further accelerating the need to store and analyze Big Data.

The ability to sort and analyze Big Data couldn’t come at a better time. At one time companies thought they were doing well just by looking at purchasing trends. Now they need to consider their social reputation to develop personalized buying recommendations, detect fraud and address various customer purchasing behaviors. This thirst to consume more and more data is driven by a highly competitive business environment and an expectation that obtaining a better understanding of our customers will improve the bottom line.

There are three essential characteristics of Big Data: velocity, volume and variety (see Figure 2).

Velocity: Big Data can be received at a range of speeds, from real time (such as Walmart recording every customer transaction) to batch file updating (as is done with MapReduce, see below), with a variety of loading speed options in between.

Volume: Volume is relative. For some organizations it may be gigabytes, while others may require terabytes and beyond.

Variety: While eventually all data must be reduced to some form of structure for analysis, the source data can range from totally unstructured (such as tweets) to fully structured (fill in the form).

**Big Data vs. the Traditional Data Warehouse**

Many managers believe that with a data warehouse, they have embraced Big Data. Let's look at comparative characteristics and see how Big Data is different from the familiar data warehouses and business-oriented corporate databases:

**Big Data**
- Data can be structured, loosely structured or unstructured
- Data can come from many sources including inside and outside the business. One source could be a traditional data warehouse
- Data can be physically distributed
- Data is of unknown quality or level of usefulness
- The data model is developed iteratively, testing to see what works best
- Analytics are part of this iterative development cycle, providing a read on the quality of the data model
- The query goes to where the data resides (distributed processing)

**Traditional Corporate Database/Data Warehouse**
- The data model is carefully designed with relationships well thought out and included in the design
- All data is centralized
- Data loads are scheduled
- Much of the data is historical
- Analytics are standardized
- Characterized by production reports
Another important distinction is that Big Data analytics are often used to look for the proverbial “needle in a haystack,” seeking some form of structure in the swirl of social media and other semi/unstructured content. The analytical process requires ongoing involvement from the business to develop hypotheses and validate results. The corporate database/data warehouse, on the other hand, is designed by the IT department to handle everything the business wants to throw at it. The use of a data warehouse and Big Data analytics are complimentary strategies: each can feed the other.

**Technology**

To attack the amount of information represented by Big Data, some technology breakthroughs are necessary. The world’s most prevalent data processing structure – the relational database – can provide views of data in any way it needs to be used, for all people for all time. Relational databases were designed for an orderly world with structured data and established data sources. Unfortunately, much of today’s data is semi- or unstructured from unanticipated sources. To meet the challenge of semi- or unstructured Big Data, a new set of technologies were needed. These technologies include NoSQL, Hadoop and MapReduce. Along with the Semantic Web, these new tools form the technology foundation for processing un/semi-structured Big Data.

**NoSQL (Not only SQL):**

SQL relational databases are designed to codify the relationships between elements of data, hence the name “relational.” Conversely, NoSQL databases do not worry about data relationships and are focused on super-fast query response, often splitting large database files across multiple computers to enable very fast parallel searching. Its strength is working with unstructured or slightly structured data such as the “mapped” key value pairs discussed below (MapReduce). It is particularly useful for statistical or real-time analysis of unstructured information like Twitter posts or Internet server logs.

**Hadoop:**

Hadoop provides the capability to access and process data that resides on multiple, separate computing devices as if the data were on a single device. Hadoop is designed for batch applications and supports MapReduce jobs.

**Hadoop File System (HDFS):**

The foundation of Hadoop, HDFS is not restricted to MapReduce jobs. It can be used for other applications such as machine learning, marketing analytics, data mining, clickstream analysis, data archiving and data warehousing. We are just beginning to see solid commercial data analysis products that can leverage the distributed nature of the HDFS. While Hadoop was designed for batch processing, there are several efforts to leverage the distributed nature of the HDFS for real-time or near-real-time, data-oriented applications. One of the most aggressive approaches is by Facebook², which uses Apache’s HBase database with HDFS. To provide some perspective on the real-time performance of this combination: at startup, Facebook required a system that could handle 60,000 messages/second, 50,000 instant messages/second and 300 terabytes of data storage growth/month.

---

MapReduce:

MapReduce (see Figure 3) is a set of programming libraries that works with Hadoop distributed data to map unstructured data to key value pairs (a key might be a particular restaurant, with the value being guest opinions harvested from social media), to provide the occurrences of specific words for each key. This key value pair is then reduced through whatever analysis is desired into a single value. Key attributes of MapReduce are that it can be run on tens of thousands of commodity computers in parallel using Hadoop to organize the parallel processing, and that it brings structure to unstructured data.

FIGURE 3: Example of How MapReduce Works

Apex Amusement Park, in the face of declining visitors, is trying to determine which will be its best investment: a) Upgrade the roller coaster; b) Improve food selections; c) Add stage shows. These are the “keys.”

The management team tweeted these options and posted them on the corporate Facebook page. Apex then monitored the social media looking for mentions of the three options, along with the associated opinions. In reality, there could have been thousands of mentions in the social data streams, but in this example we will assume there were nine. The business team decided that the key word supporting the option was “excellent.” So, the monitored data is stored in a Hadoop file system, which is actually distributed across many commodity computers operating in parallel. The search through the millions of entries looking for mentions of Apex, the three options and the word “excellent” is actually chunked into smaller searches across many low-cost computers. If the option is found without the word “excellent,” then that option is assigned a “0” value. If “excellent” is found, then the option gets a value of “1”.

The social media monitoring resulted in the following mapping (three mentions did not include the word “excellent,” while six did):

<table>
<thead>
<tr>
<th>Upgrade the roller coaster: 0</th>
<th>Improve food selections: 1</th>
<th>Add stage shows: 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Upgrade the roller coaster 1</td>
<td>Improve food selections 0</td>
<td>Add stage shows: 1</td>
</tr>
<tr>
<td>Improve food selections: 0</td>
<td>Upgrade the roller coaster 1</td>
<td>Add stage shows: 1</td>
</tr>
<tr>
<td>Improve food selections: 0</td>
<td>Add stage shows: 1</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

And reduces to:

a) Upgrade the roller coaster (0,1,1)
b) Improve food selections (1,0,0)
c) Add stage shows (1,1,1)

or, more clearly:

a) Upgrade the roller coaster: 2
b) Improve food selections: 1
c) Add stage shows: 3

Imagine sifting through millions of bytes of social data looking for synonyms of “excellent” with perhaps hundreds of keys (like hotel names). This is the type of search for which MapReduce is built.

Source: PhoCusWright
© 2014 PhoCusWright Inc. All Rights Reserved.

Semantic Web:

The Semantic Web and the associated Resource Description Framework are beginning to grow in usage and importance. With this growth, traditional methods of analysis are no longer adequate, and the use of Big Data tools to analyze Semantic web data has begun.

Corporate Data Topology

Structured data comes from various operational and record keeping systems to populate the data warehouse/corporate database. This data input can also be the result of using MapReduce to convert unstructured data into a MapReduce-structured output, while the Hadoop cluster and/or the NoSQL system are populated with unstructured/semi-structured/structured data from internal and external sources, including the data warehouse. Outputs from the data warehouse are reports and fixed dashboards, while the results of Big Data analytics are more likely based on ad hoc queries and discoveries of meaningful relationships. To make this even more confusing, data warehouses are emerging that use the Hadoop File System (HDFS) (see Figure 4).

FIGURE 4: Corporate Data Topology

Source: PhoCusWright
© 2014 PhoCusWright Inc. All Rights Reserved.

Traditional Capabilities are Being Challenged

Output from un/semi-structured data processed through MapReduce can feed a traditional data warehouse/corporate database to be combined with structured enterprise data to provide integrated external/internal views of the business environment. This data integration is just part of the Big Data explosion: traditional structured data stores are being stressed by the amount of data that is processed and the speed at which the processing must occur. This stress is driving additional new technology developments. These include developments in memory databases, solid state data
storage, data warehouse appliances, complex event processing, massively parallel processing, the use of NoSQL databases and, as discussed earlier, leveraging the HDFS infrastructure.

This is a highly dynamic area, with traditional database providers such as IBM and Oracle acquiring or partnering with startups to provide a running start for Big Data analytics. Existing vendors such as IBM, Microsoft, Oracle, SAS and Informatica are making forays into Big Data processing. New vendors to watch include: Cloudera, ClickFox, Merced, Greenplum, Hadapt, Hortonworks, InfoSphere, Karmasphere and MapR.

As an example of how fluid Big Data technology is, in May 2012, Google announced the general availability of its Big Query service, which provides analytical capability for terabyte-scale datasets. Using SQL-like queries, Big Query features very fast data load times (on the order of several hundred gigabytes in 15 to 20 minutes). Even though Big Query uses a distributed query and data storage architecture, it does not use Hadoop and MapReduce, which originated in Google. Whether or not Big Query is a game changer remains to be seen.

The list of vendors and key players is by no means comprehensive. Expect a flood of new software products and startups with innovative and creative ways to support the three V’s of Big Data: Volume, Velocity and Variety. The market related to just Hadoop and MapReduce programming frameworks for Big Data is expected to grow from $77 million in 2011 to $813 million in 2016, for a compound annual growth rate of 60%, according to IDC.

Cloud computing is also becoming a resource for Big Data processing. Amazon supports Hadoop deployments in its Elastic MapReduce cloud, giving users the ability to scale up or down as needed. Microsoft has dropped plans for its own Big Data capability, and will support Hadoop in its Azure cloud. Startup Mortar Data is providing Hadoop services in a cloud and claims that if you can code (Python) you can be up and running on a Hadoop cluster cloud service in an hour. Google uses cloud computing for its Big Query product.

**Business Scenarios**

Following is a list of actual or possible scenarios for extracting business value from Big Data:

- **Recommendation engine:** Allows people matches where there are similar profiles, or to recommend products and services based on buying patterns of similar people. Linkedin uses Big Data to support its “People You May Know” capability. Amazon uses Big Data to suggest retail products to shoppers (customers who have viewed one item have also viewed other, similar items)

---

4) Perez, Juan Carlos, “Google Makes Big Query Available to All”, CIO Magazine, May 1, 2012 http://www.cio.com/article/705420/Google_Makes_Bigquery_Available_to_All
Sentiment analysis: Determines user sentiment toward products, services, destinations, cruise companies, hotels, etc. (see Figure 4 for a simplified example) http://www.utd.edu/~lkhan/papers/new/Paper-SR-Large-RDF-Graphs.pdf

Analysis of the Social Graph: Determine where some people exert influence over others. The influencers are candidates to be brand advocates

360-degree view of the customer: Analyze all of the enterprise data relating to a customer. Look at everything involved with customer interaction (and comment)

Marketing insights: Determine almost instantly how well a new marketing campaign is being received, allowing quick changes if required

Customer abandonment analysis: Determine the customers most likely to leave, and target them for retention through special considerations

R&D: Test new products (or ideas for products) and obtain instant feedback

Fraud detection: Look at multiple sources that include historical and transactional data to detect suspicious patterns.

Analyze competitors’ pricing patterns

A 2011 McKinsey report suggests suitable Big Data technologies include A/B testing, association rule learning, classification, cluster analysis, crowdsourcing, data fusion and integration, ensemble learning, genetic algorithms, machine learning, natural language processing, neural networks, pattern recognition, predictive modeling, regression, sentiment analysis, signal processing, supervised and unsupervised learning, simulation, time series analysis and visualization.

These are just a few ideas. Big Data is a new technology that continues to evolve and mature. Along with this maturation, businesses will continue to discover new and exciting uses for Big Data.

Business Examples (Travel)

Hopper (www.hoppertravel.com), a startup founded by a team of former Expedia employees, is developing a website supported by Big Data to provide consumer travel discovery, and harness a new generation of big data analysis tools. Using machine learning, NoSQL databases and Big Data processing, Hopper is transforming raw Web pages into structured and organized information, so users can conduct better, more complete searches faster than they could on traditional travel sites.

Deal Angel is a new hotel metasearch site that crunches millions of hotel rates to be able to show which rates are actually good deals.

Hotel Ratings Analysis sites like ReviewPro, TrustYou and Brand Karma crunch unstructured data from thousands of travelers to provide structured analyses of a hotel property’s reputation and service positives/deficiencies.

**Business Examples (Non-Travel)**

Walmart, one of the pioneers in the use of Big Data, used predictive analytics to observe customer preferences so that it could stock its regional distribution centers accordingly.

In overseeing response to the Deepwater Horizon oil spill disaster, the U.S. government used Big Data analytics to collect, correlate and analyze data from a variety of sources to develop actionable intelligence to help deploy cleanup forces.

A medical diagnostics company analyzed millions of lines of data (100 million gene samples) to develop a non-intrusive test for cardiac disease.

LinkedIn uses Big Data analytics to leverage content from user profiles to develop new social data products.

Medical sleuths are using Big Data to visualize the potential course of a pandemic.

Retailers are using Big Data to understand why online shoppers abandon shopping carts.

Amazon is running daily experiments: Changing prices on specific items or changing the way products are presented. Given the low cost of Big Data processing, Amazon can run hundreds of experiments daily.

Stock brokers are using Big Data-based sentiment analysis as part of their stock evaluations. This approach has become so popular that some hedge fund managers are grounding their portfolio strategy on Twitter Analytics. Other Wall Street applications include minimizing risk through better data analysis, spotting rogue traders and fraud detection.

**Vision for the Travel Industry**

- Data about customer preferences, behaviors and buying patterns will become readily available outside of the GDS, making GDS data products less valuable.
- Imagining where to go and what to do with leisure travel will become easier with structured information gleaned from unstructured content using Big Data analytics.
- Similarly, comparative shopping will be easier with Big Data enabling side-by-side comparisons.
- The travel value chain is characterized by huge amounts of information and data that overwhelm the traveler. Big Data analysis will provide the aggregation and/or subset of information that is relevant, regardless of whether the traveler is imagining, learning, shopping, booking or experiencing travel.
- Big Data collection and analysis will enable many of the changes long discussed:
  - Personalized fares and rates
  - Total trip planning
  - Many more things that haven’t yet been envisioned
What You Should Do

- Understand what Big Data is and the opportunities it offers. This can be done by reading industry articles, additional trade literature and case studies, use of consultants and strategic brainstorming.

- As appropriate, use Big Data tools such as those mentioned in this article to solve current business problems and to pursue opportunities for tactical advantage. These should be supported by business cases.

- Use cloud computing wherever possible to reduce capital investment and support tactical flexibility.

- Set aside 20% (plus or minus) of your Big Data effort as a working laboratory, exploring new ideas and new sources of and relevant uses of Big Data and associated analytics.