DataInformed
Big Data and Analytics in the Enterprise

Bridge the Gap Between Business and IT: Integrating Data into Business Workflows

Table of Contents:

Sponsored by:

Find the Needle in the Semi-Structured Haystack

by Ayush Parashar

Social media data offers a veritable goldmine of insights about users. But companies struggle to transform social data and combine it with existing data sets to enrich analysis. Ayush Parashar of UNIFi describes the steps from data discovery to insight.

Ayush Parashar, *co-founder and VP, Engineering , UNIFi Software*

The boss wants to know if the business analyst can help the merchandising department forecast the demand for the company's Halloween costumes based on social media trend data when combined with other supply chain variables.

The answer is needed by the end of the week to meet the final ERP schedule or the company may be faced with rush charges in manufacturing or higher distribution costs to ensure supplies are ready to meet customer demand.

The request seems obvious and logical. If lots of people, especially those with buying power or influence, are talking about the company's products in the second quarter, the manufacturing department would have some indication whether the company has one of the "must-haves" for this year's trick-or-treating.

Social Media Data: Lots of Promise, Lots of Headaches

Data about social network members and their usage habits is a treasure trove of insight. From detailed demographic profiles to interests and activities, there's nothing private about a person's life if they choose to be active on one or more of the growing number of social media networks.

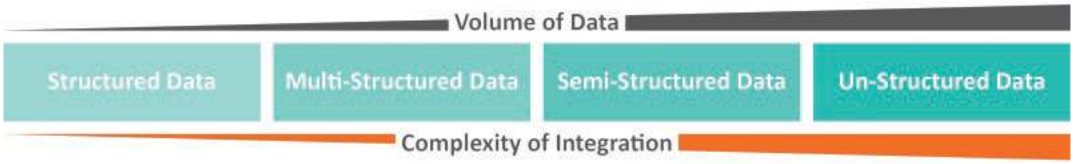
The landscape of social networks is changing and so, too, is the demographic that each attracts. Now that parents and grandparents have Facebook accounts, most of the younger generation has moved on. They have their conversations on Instagram, Yik Yak, or Snapchat. Businesses trying to get a total picture of their customer base must acquire these new data sets. Developers of these networks, looking for early revenue streams that won't alienate users with advertising, sell usage data directly or license it to third parties.

The good news is there's a ton of data. The bad news is there's a ton of data. And for most business analysts and the IT departments that support them, the real challenge is not only acquiring the data, but also transforming the data so that it can be combined with existing data and analyzed using business intelligence and visualization tools to extract business insights.

The Structure of Things

If you are reading this, you probably already know that data can be defined in a number of ways: Structured, multi-structured, semi-structured, and unstructured.

Generally, the volume of data and the complexity of integration increase as the structure decreases. Before any analysis can occur, the structure of the data must be normalized or flattened into a tabular form consisting of rows and columns so the end user can understand context and enable their visualization tool of choice to display the results.



Social media data can be classified as semi-structured or multi-structured data. That is, some structure exists, but much of the data must be parsed and normalized before any analysis can occur. The data often consist of multiple levels of nesting based on the various linear and non-linear parameters being reported. For example, time-based events can be considered linear and structured, whereas trend data based on tweets, re-tweets, follows, etc., is more unstructured, happening randomly from a time perspective and with no obvious behavior pattern.

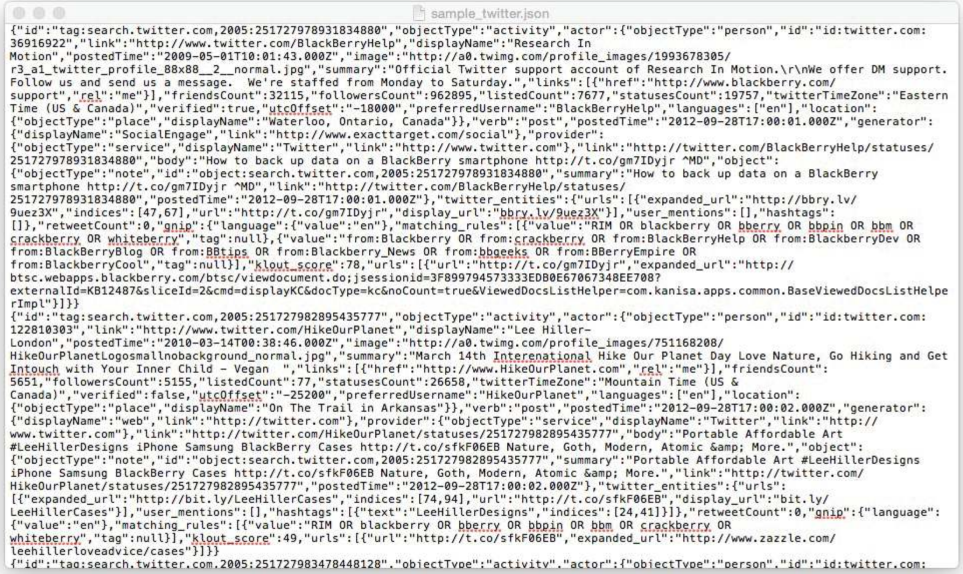
Normalizing the Data

Before any data can be analyzed, the business analyst must explain to a programmer in the IT department what question she is trying to answer or what problem she is trying to solve. This is huge a challenge – the analyst may not know exactly which data sources are available and therefore may not even know about some pivotal information she could request to have included. The IT professional is familiar with the data sources and has the technical expertise to structure the data into usable files, but may not know enough about the analyst's specific business case to include all of the relevant sources and attributes. This can mean incomplete data and frustrating, time-consuming iterations.

Before any analysis on the data can occur, the business analyst and the IT programmer must follow most or all of the following steps.

1. Data Discovery

- Data sources: Understand what data is available to the organization
- Locate data: Where in the organization does the data reside? Local storage? Remote storage? Personal desktop?



```
{"id":"tag:search.twitter.com,2005:251727978931834880","objectType":"activity","actor":{"objectType":"person","id":"id:twitter.com:
36916922","link":"http://www.twitter.com/BlackBerryHelp","displayName":"Research In
Motion","postedTime":"2009-05-01T10:01:43.000Z","image":"http://a0.twimg.com/profile_images/1993678305/
r3_a1_twitter_profile_88x88__2__normal.jpg","summary":"Official Twitter support account of Research In Motion.\r\nWe offer DM support.
Follow us and send us a message.  We're staffed from Monday to Saturday.","links":[{"href":"http://www.blackberry.com/
support","rel":"me"}],"friendsCount":32115,"followersCount":962895,"listedCount":7677,"statusesCount":19757,"twitterTimeZone":"Eastern
Time (US & Canada)","verified":true,"utcOffset":"-18000","preferredUsername":"BlackBerryHelp","languages":["en"],"location":
{"objectType":"place","displayName":"Waterloo, Ontario, Canada"}},"verb":"post","postedTime":"2012-09-28T17:00:01.000Z","generator":
{"displayName":"SocialEngage","link":"http://www.exacttarget.com/social"},"provider":
{"objectType":"service","displayName":"Twitter","link":"http://www.twitter.com"},"link":"http://twitter.com/BlackBerryHelp/statuses/
251727978931834880","body":"How to back up data on a BlackBerry smartphone http://t.co/gm7IDyjr ^MD","object":
{"objectType":"note","id":"object:search.twitter.com,2005:251727978931834880","summary":"How to back up data on a BlackBerry
smartphone http://t.co/gm7IDyjr ^MD","link":"http://twitter.com/BlackBerryHelp/statuses/
251727978931834880","postedTime":"2012-09-28T17:00:01.000Z"},"twitter_entities":{"urls":[{"expanded_url":"http://bbry.lv/
9uez3X","indices":[47,67],"url":"http://t.co/gm7IDyjr","display_url":"bbry.lv/9uez3X"}],"user_mentions":[],"hashtags":
[]},"retweetCount":0,"gnip":{"language":{"value":"en"},"matching_rules":[{"value":"RIM OR blackberry OR bberry OR bbpin OR bbm OR
crackberry OR whiteberry","tag":null},{"value":"from:Blackberry OR from:crackberry OR from:BlackBerryHelp OR from:BlackberryDev OR
from:BlackBerryBlog OR from:BBtips OR from:Blackberry_News OR from:bbgeeks OR from:BBerryEmpire OR
from:BlackberryCool","tag":null}],"klout_score":78,"urls":[{"url":"http://t.co/gm7IDyjr","expanded_url":"http://
btsc.webapps.blackberry.com/btsc/viewdocument.do;jsessionid=3F899794573333EDB0E67067348EE708?
externalId=KB12487&sliceId=2&cmd=displayKC&docType=kc&noCount=true&ViewedDocsListHelper=com.kanisa.apps.common.BaseViewedDocsListHelpe
rImpl"}]}}
{"id":"tag:search.twitter.com,2005:251727982895435777","objectType":"activity","actor":{"objectType":"person","id":"id:twitter.com:
122810303","link":"http://www.twitter.com/HikeOurPlanet","displayName":"Lee Hiller-
London","postedTime":"2010-03-14T00:38:46.000Z","image":"http://a0.twimg.com/profile_images/751168208/
HikeOurPlanetLogosmallnobackground_normal.jpg","summary":"March 14th Interenational Hike Our Planet Day Love Nature, Go Hiking and Get
Intouch with Your Inner Child - Vegan  ","links":[{"href":"http://www.HikeOurPlanet.com","rel":"me"}],"friendsCount":
5651,"followersCount":5155,"listedCount":77,"statusesCount":26658,"twitterTimeZone":"Mountain Time (US &
Canada)","verified":false,"utcOffset":"-25200","preferredUsername":"HikeOurPlanet","languages":["en"],"location":
{"objectType":"place","displayName":"On The Trail in Arkansas"}},"verb":"post","postedTime":"2012-09-28T17:00:02.000Z","generator":
{"displayName":"web","link":"http://twitter.com"},"provider":{"objectType":"service","displayName":"Twitter","link":"http://
www.twitter.com"},"link":"http://twitter.com/HikeOurPlanet/statuses/251727982895435777","body":"Portable Affordable Art
#LeeHillerDesigns iPhone Samsung BlackBerry Cases http://t.co/sfkF06EB Nature, Goth, Modern, Atomic &amp; More.","object":
{"objectType":"note","id":"object:search.twitter.com,2005:251727982895435777","summary":"Portable Affordable Art #LeeHillerDesigns
iPhone Samsung BlackBerry Cases http://t.co/sfkF06EB Nature, Goth, Modern, Atomic &amp; More.","link":"http://twitter.com/
HikeOurPlanet/statuses/251727982895435777","postedTime":"2012-09-28T17:00:02.000Z"},"twitter_entities":{"urls":
[{"expanded_url":"http://bit.ly/LeeHillerCases","indices":[74,94],"url":"http://t.co/sfkF06EB","display_url":"bit.ly/
LeeHillerCases"}],"user_mentions":[],"hashtags":[{"text":"LeeHillerDesigns","indices":[24,41]}]},"retweetCount":0,"gnip":{"language":
{"value":"en"},"matching_rules":[{"value":"RIM OR blackberry OR bberry OR bbpin OR bbm OR crackberry OR
whiteberry","tag":null}],"klout_score":49,"urls":[{"url":"http://t.co/sfkF06EB","expanded_url":"http://www.zazzle.com/
leehillerloveadvice/cases"}]}}
{"id":"tag:search.twitter.com,2005:251727983478448128","objectType":"activity","actor":{"objectType":"person","id":"id:twitter.com:
```

Semi-structured data as received from Twitter.

2. Data Acquisition

- Acquire data: License new data as needed from third-party vendors. This must factor: cost, timeline, sign-off process
- Desired outcome: Once the data sets have been identified, the analyst must explain the business requirements to the IT programmer
- Sample data: A sample of each data set must be acquired and reviewed
- Data structure: The structure of each data set must then be determined

3. Transform

- Data normalizing: A series of normalizing processes must be applied to the sample data
- Attributes of interest: Attributes of Interest must be extracted from the normalized process by writing Java Map Reduction code for each data set
- Data aggregation: Aggregation code must be written to combine multiple data sets into a single normalized view

4. Test and Fix

- Sample testing: A derived normalized, flat file must be tested against the corporation's BI visualization tools
- Analytic tests: A series of analytic tests are performed to test the functionality of the combined data and find edge cases

- Load raw data: Raw data for each data set is imported and the code is applied so the analyst can view the data holistically
- Data errors: Discrepancies in the results caused by incomplete data, data not normalized or aggregated correctly, requires code change

5. Business Insight

Even with this overly simplified description of the process that business analysts face daily, it's no surprise that many insights are delivered late or not at all. The process is challenging: Make a non-technical analyst explain in business terms, to a non-business technical person, what the end result of the analysis should be. One senior data analyst at a Fortune 500 company explained that analysts often won't even try to explain to the IT department what it is they are trying to accomplish because it is just too difficult.

The good news is that there is help in sight. Two key developments are occurring in the analytics space that will improve the lives of both the business analyst and the IT department that supports them. The first is less expensive, centralized storage and the processing of semi-structured and multi-structured data natively in Hadoop. The second is substantially more capable visualization tools that help analysts glean the nuggets of insight from the mountains of raw data.

Make a non-technical analyst explain what the end result of the analysis should be.

New emerging companies such as UNIFi Software, Tamr, and Trifacta, along with evolving industry stalwarts like Syncsort and Teradata, are helping companies integrate their data more quickly by applying complex data parsers and data logic where software developers previously had to hand code the integration with very technology-oriented tools native to the Hadoop eco-system such as Pig, Hive or Java MapReduce programs.

Ayush Parashar is co-founder and VP, Engineering at UNIFi Software, Inc. Ayush has deep software engineering expertise around big data solutions and strong domain knowledge around Hadoop, MPP database and systems, performance engineering and data integration. Before UNIFi, he was part of the founding engineering team at Greenplum.

Translate Business Talk into Predictive Analytics Projects

by James Drew

Because not all stakeholders are conversant in predictive analytics terminology, the most important aspect of the data mining process might be translating a business problem into an analytics project.

We are living in a golden age of predictive analytics and its constituent techniques of data mining and machine learning. New and ever-more clever algorithms are being invented by researchers from diverse backgrounds, and they are made available in software packages that have a variety of options, interfaces, and costs. Instruction in the methods and software is also available through formal degrees and short courses, as well as various Internet videos and other resources.

Modern Predictive Analytics (MPA – the "Modern" refers to the use of automated methods) is thus an eminently viable basis for the practice of data science, which combines techniques, software, and data acquisition with domain-specific knowledge to help real people – not just analysts – solve real-world problems.

And therein lies a major difficulty of data science practice. Not everyone in the MPA loop is up to date on the new vocabulary and capabilities of predictive analysis. A business executive with a numerical problem is unlikely to specifically ask for a regression or tree, preferring to describe the operational or marketing aspects of her problem. ("I need to find a cause for this uptick in cellular handset problems. I don't care exactly how you do this, just make sure it's useful and I can sell it to my boss.") And sometimes when such executives do incorporate MPA ideas in their narrative, they are often dated, simplistic, or irrelevant. ("Do a linear regression of Overall Satisfaction on Customer Service Satisfaction in this survey, and make sure its R2 is 90 percent!")

Therefore, perhaps the single most important aspect of the MPA/Data Mining process is not automated and perhaps not automatable: the intelligent translation of the narrative of a business problem into an analyzable project. The technical MPA requirements for this can be quite modest: a few regression, tree, and association tools go a long way in problem solving when the provisional goal is to get to, "Yes, I can help you." Of course, an ultimately good solution may well involve complicated, perhaps dirty data and an array of sophisticated algorithms, but the hard part is getting the project started so all parties are comfortable with its initial direction.

The translation part is akin to literary criticism in the sense that we are constantly asking why our business narrator uses certain sets of words, and what they evoke in her and in us. For example, in trying to calculate a customer's Tier—a desirability index—from just his monthly payment rather than the more complicated notion of profit, consider this quote from a business speaker: "Profit depends on many accounting assumptions and is hard to convincingly calculate. We think that profit is closely related to revenue alone and using only this information to calculate the threshold for a customer's Tier would save my department a lot of money and time every month." The attentive data scientist would likely downplay the first sentence, focus on the phrase "closely related," and visualize a simple scatterplot of revenue versus profit color coded by Tiers, from which several quantitative solutions (as suggested by the word "calculate") would present themselves.

Some experienced analysts, of course, perform this translation naturally, but for others I have found it helpful to develop the two sides of the needed translation. Listening for words and phrases and their careful consideration is one side. The other side is to build, or reinforce, images and other thought fragments of what a predictive solution would look like: a scatterplot, low-dimensional function, or concise rule, for example. If the translation process is having someone "ring a bell" in your mind, then the bell must be made to produce its sound, and we have to find a context in which the sound is meaningful.

The single most important aspect of the data mining process: the intelligent translation of a business problem into an analyzable project.

Both the listening and the application of MPA techniques can be fun, but is there any financial value to these complementary exercises? A few years ago, a landline phone company president asked how to identify electrical cable that was anecdotally thought to be trouble-prone, with the words, "When does this stuff go bad?" With a scatterplot of repairs as a function of cable age and a logistic regression-like model, we found that the existing cable did not, in fact, go bad with age and that most of the troubled cable had long-since been replaced. Consequently, the president got to use his replacement money on other projects more important to him and to his customers. The savings: $30 million, at a cost of a few days of analyst time and a few hundred dollars of software.

The golden age of MPA, indeed.

Dr. James Drew teaches data mining at Worcester Polytechnic Institute and has been an internal statistics and data mining consultant at Verizon Communications for 30 years. Jim holds two post-graduate degrees in mathematics from Cambridge and a Ph.D. in statistics from Iowa State.

Self-Reliant Analytics and the Implementation 'Gag Rule'

By Scott DeMers and Anthony Krinsky

Scott DeMers, *Worldwide Director of Channel Presales, Tableau*

Anthony Krinsky, *Senior Sales Consultant, Tableau Channels organization*

Organizations that talk about and adopt BI analytics implementations enjoy several business and cultural advantages over organizations that do not.

In the 1950s, Toyota empowered its least senior employees to stop the entire production line by reaching up and pulling the "Andon" cord. Andon cords connect to sign boards that illuminate and summon engineers directly to the defective unit. In Toyota's culture, this is more than a simple quality control convenience. Andon elevates and cultivates the active participation of each employee around the core organizational mission.

Self-reliant analytics can provide similar empowerment, alignment, and coordinating functions in knowledge organizations. A recent Aberdeen study showed that 70 percent of interactive visualization adopters improved collaboration and decision making. Unfortunately, adoption rates are typically too low for these systems to impact culture company-wide.

Unlike other enterprise software deployments, business intelligence rollouts tend to be unsystematic, incomplete, and short-lived. Are customers ignoring the advice and best practices of business intelligence vendors? Or is it that these vendors have been reluctant to talk about implementation at all?

The industry talks too little of implementation for three reasons. The first is that the tools have, for generations, been hard to use. They are hard to implement, and keeping up with fast-changing requirements is difficult. The learning curve is just too steep and most business users won't persevere when they are so unsure of success. Second, comprehensive post-sale implementation plans seem to tarnish "ease-of-use" bona fides and to boost total cost of ownership. Finally, implementation requires discussions between IT and business, together. By avoiding these discussions, vendors can sell to either without the consent of both.

Clever? Yes. Helpful? No.

The virtual "gag order" on implementation solves sales problems, but at what cost? There are important differences between organizations that talk a lot about implementations and those that do not.

Strong implementers did not put business intelligence in a special category. They brought over the best tools from other software deployments: special interest group mailing lists, training schedules, help desks, thoughtfully designed intranet sites, office hours, guest speakers, work that inspired, executive sponsorship, and collaboration. They also brought systematic, programmatic support.

By doing so, they discovered that the speed, power, and ease-of-use of modern analytics tools far exceeded their predecessors, and that business users could indeed summit the learning curve. They on-boarded new users quickly and maintained a continuum of excitement.

IT and business partnered together and divided roles sensibly. Back-end data sources were clean and well documented. Report development shifted to the business, and a buddy system emerged. The “water fall” report-development process disappeared. It was replaced not with Agile but with a real-time process that was even more agile. Program managers had good answers to predictable questions business users asked: “How do I get started?” “What data is available?” “Why is this important?”

Organizations taking implementation seriously reaped the full benefits of self-service BI. Operational visibility increased at every level. Good information, based on timely, factual data, made hard decisions easier. Transparency aligned people with the organization’s mission, and business users felt more independent and in control of their work lives. They worked more collaboratively with one another and felt connected to analysts around the world.

This kind of culture shift matters. Gallup finds that fewer than one third of American workers are truly engaged in their work, and Blue Ocean Leadership identifies traditional reporting as part of the problem. Performance scorecards and operational reports are “apex” instruments of the “command and control” culture that knowledge workers reject.

Self-reliant business intelligence is turning that model upside down. It’s the Andon cord of the knowledge workplace, allowing those with the most intimate understanding of operations to investigate issues themselves and to communicate their findings widely and clearly. It helps organizations move the dial on powerful intrinsic motivators: purpose, autonomy, mastery, and community. But those benefits won’t accrue to organizations that discuss implementation in hushed tones or not at all. This work is too important to be kept a secret.

Scott DeMers is the Worldwide Director of Channel Presales at Tableau. He has experience in running global and national distributed sales organizations. He has specialized experience in business intelligence including data discovery, enterprise reporting, OLAP, dashboards, workflow, and high-throughput reporting engines.

Anthony Krinsky is a Senior Sales Consultant with the Tableau Channels organization. Based in Los Angeles, he spent nine years in similar roles at IBM Cognos and SAP Business Objects, and eight years as a database application developer for Java, Cold Fusion, JavaScript, mapping, and .NET.

Accurate, Timely Analysis Requires Simplified Data Cleansing

UNIFi Software CEO Rob Carlson discusses the problem of latency due to manual data cleansing and normalization, and how tools that automate this process lead to more accurate insights and an improved bottom line.

By Rob Carlson

Rob Carlson,
co-founder and CEO,
UNIFi Software

Data analysis has become an essential part of every business. And like many office automation tools, data analytics tools have evolved from chalk marks on a blackboard to rich, dynamic, on-screen displays presenting the complete picture of a business as pie charts, scatter graphs, and three-dimensional bar graphs.

Not only has the form of data visualization changed dramatically over the past decade, but the sources of the data that are being displayed have grown dramatically at the same time. It seems that almost every day a new social media site or consumer behavior is front and center for analysis, and shifting demographics make the data analyst's job even more complex. For example, gone are the teens and tweens from Facebook, regarding it, now that their parents and – worse – grandparents are on it, as an awkward family Sunday dinner. They have moved on to more instantaneous and short-lived social media sites, like Yik Yak and Snapchat. The nature of these services allows for more controlled, more private interactions with fewer people in your inner circle. No moms or grandpas invited.

More Analysis, More Accurate Results

The availability of first- and third-party data allows the analyst to understand the habits of customers and prospects at an intimate level. At Disney/ABC, for example, the data team collects more than 1 billion pieces of data every day. These data points represent the viewing habits of its consumers based on dozens of sources, from device type and content stream to geolocation and time of day.

With this data, Disney/ABC is able to profile the viewing habits of individual users of its service and push relevant and timely content to a consumer, dramatically increasing the likelihood of engagement with sponsors' messages.

Some Data Do Not Play Nice

Due to the nature of the services that are generating data, some of the most potentially valuable data, such as that from social media networks or from news feeds like Twitter, arrive at the analyst's desktop in a totally unstructured way. This poses a problem for the analyst, as this data cannot be viewed or combined with existing data services for analysis until the structure of the data has been normalized.

Cleansing and normalizing unstructured data is a highly technical and time-consuming task. Before business analysts can dedicate themselves to the task of discovering insights from the data, they must have the data presenting to their visualization tools in such a way that it can be represented. This requires the data to be normalized to a tabular form – that is, defined in terms of rows and columns.

The problem is compounded when tables of data generated from different sources need to be combined in order to derive valuable insight. For example, to understand the influence that social media trends are having on online sales, the analyst must combine CRM data with website click stream data and social media trend data. This sounds easy, but is actually quite complex.

Before business analysts can dedicate themselves to the task of discovering insights from the data, they must have the data presenting to their visualization tools in such a way that it can be represented.

Typically, data analysts will work with a developer in the IT department to identify the data sources that are available for the specific research they require. The programmer then collects the data sources and writes software to cleanse and normalize the data. The challenge in this process is that the technologist does not fully understand the business objectives or hypotheses of the analyst. In turn, the analyst may not appreciate the technical limitations of cleansing and joining data sources together.

This operational disconnect can lead to a time-consuming and frustrating process, as each side tries to refine its requests and deliverable. The time that is required to actually deliver the insight can have a negative impact on the bottom line of the business, and opportunities to proactively react to consumer buying habits or other customer value may be lost entirely.

Improved Access to Data

For data analysis tools to achieve the same widespread adoption in the workplace that word processing and spreadsheet applications currently enjoy, the task of acquiring, cleansing, and normalizing data so that it can be viewed by analysis tools needs to improve dramatically.

Tools are emerging that remove the technical programming phase of data integration and free business analysts to explore available data sources and immediately combine sources together so they are visualized quickly. These tools, such as UNIFi Software, deliver a user-friendly interface designed for the business user and programmatically cleanse and combine data sources seamlessly "under the covers" so the analyst is completely separated from and oblivious to the complexity of the task that is required to present normalized data to the data visualization tool.

The simplification of this mundane but essential element of any data analysis frees the analyst to pursue "what if" scenarios with the data, hypothesize about their business, and employ predictive analytics in a simplified way to gain business insights about their customers at the speed of thought.

Rob Carlson is co-founder and CEO of UNIFi Software. Rob has served the enterprise technology and Big Data market for over 20 years. In that time, he has held senior leadership roles at Business Objects/SAP, EMC, PeopleSoft. Prior to co-founding UNIFi Software, Rob helped launch and grow early venture-backed big data companies Platfora and Alpine Data Labs.

Six Degrees of Data Integration

Sean Keenan of UNIFi Software breaks data integration into six distinct parts and explains how automation of this process eliminates the disconnect between business analysts and IT and leads to faster insights.

By Sean Keenan

Sean Keenan, *co-founder and VP of Products, UNIFi Software*

Data integration has become a major feature of the big data landscape. As more and more data sources become available to business analysts and visualization tools become even more powerful, the need to integrate data into the analyst environment has become paramount.

Traditionally, the task of integrating data has fallen to a dedicated team of programmers in the IT department. The task is very technical and requires a comprehensive knowledge of data structure and programming. Frequently, two or more data sets can be integrated only by writing a custom software program, often referred to as Java Map Reduction. But other tricks of the trade are used to simplify the data, identify missing attributes or delimiters, and derive a structure that can be viewed by the visualization tools.

Language Barrier

The challenge for business analysts is the total disconnect between their desired outcome and the development effort required to deliver that result. The business analyst speaks business and forms a hypothesis or wishes to pursue "what if" scenarios with the data in order to identify a business insight. The programmer, on the other hand, is looking at the problem from a purely technical perspective and sees the task as a structured data challenge.

This process of business case explanation being translated into technical lingo is the language barrier that often leads to frustration and delay. The nature of exploratory analysis is that the situation or hypothesis may change along the way and require further or different data integration. Plus, when the results are delivered by the programmer, the analyst must determine if this data is providing the right result. Obvious errors show up in the visualization program and can be spotted easily, but more subtle errors may not be detected, leading to misinterpretation of the data and possibly affecting critical business decisions. This necessitates a system of checks and balances, which means that it will take even longer to deliver the data, costing the company time and money.

The Six Degrees of Integration

Data integration can be broken down into six distinct parts. Each of these parts forms a critical piece of the end-to end-process of moving from data acquisition to visualization. The task of integration must be approached in a structured, linear way from start to finish. Jumping in halfway through undoubtedly will result in errors, delay, and further frustration. The six degrees of integration can be defined as follows:

- Acquire the data sources required for analysis.
- Search the data sources to find relevant information and objects of interest.
- Cleanse and/or enrich the data to remove delimiters, spurious fields or values, and add arguments or logic to assist in the structure of the data.
- Normalize the data for the visualization tool or combination with other data.
- Transform the data by combining two or more data sets together.
- Visualize the results by structuring the data in such a way that it can be viewed by the visualization tool.

The acquisition of data falls into two parts: The first part is sourcing the data from either first-party business operations such as salesforce.com, OSI, or ecommerce platforms or from third-party providers such as Twitter, Facebook, or data brokers such as Datasift. The second is importing and hosting the data in the analytics environment. This often requires firewall access or DMZ repositories to protect the organization's data integrity.

Search the data source to find objects of interest, such as specific customer profiles, geographies, time of day, day of week, etc.

The challenge with data cleansing is that many data sources lack structure. That is, they do not form nicely into tables of rows and columns. This semi-structured or un-structured data requires specialized knowledge, tools, and/or programming skills to normalize. For data that contain wide and varying values, the manual approach to cleansing is extremely time consuming and fraught with potential error.

Without normalizing individual data sets, it is practically impossible to join two data sets together based on individual objects of interest.

Normalizing data is the process of providing structure so that the data can be visualized. For most visualization tools, this means delivering the interface data formed as rows or columns like a spreadsheet. Without normalizing individual data sets, it is practically impossible to join two data sets together based on individual objects of interest, such as all of our Twitter followers who made a purchase on our website in the past month.

Transforming data is where real insights can be derived. This stage allows two or more data sources to be combined prior to visualization. A truly agile integration platform will allow analysts to pursue "what if" scenarios with their data quickly and easily in order to derive new insights. For example, overlaying weather data on CRM data might determine the optimum weather conditions for promoting certain types of products.

Visualization is the final stage of any data analysis. But this can occur only when the data has been presented to the visualization tool correctly. In many organizations, different analysts or departments choose different visualization tools because of personal preference or features geared toward their day-to-day tasks. For organizations that support analysts with a common set of data sources, this means structuring the output of integration in multiple ways.

Automatic for the People

If the task of data integration falls to a dedicated team of programmers supporting potentially hundreds or thousands of analysts, this task and the team to which it falls always will be a bottleneck and source of frustration. The task of data integration must be entirely automated before true data democratization can occur within the organization.

Tools that allow the data analyst at whatever level to access the available data sources and derive insights are critical to realize the full potential of the data. Many data integration tools are available to the organization, some designed to help the IT department, others geared more toward highly technical analysts or data scientists, and still others more suited to business analysts in a self-help operation. However, with the exception of UNIFi Software, the current set of tools and services available to business analysts fail to address all six degrees of integration. At any point that the analyst must contact the IT support team to assist in the integration phase, delays are introduced. The UNIFi Software suite of tools addresses the task of integration end to end. Installed on the Hadoop environment hosting the analytics data, the analyst-friendly tools allow for total self-help across any and all data types. In many cases, UNIFi automates the process of data integration, which not only simplifies the task but also delivers critical insights sooner.

Sean Keenan is co-founder and VP of Products at UNIFi Software. Sean has deep industry knowledge and experience in delivering large-scale advanced analytic solutions to clients across multiple verticals. He is a career big data practitioner and has been involved in hundreds of big data implementations across F1000 companies. Prior to co-founding UNIFi Software, where he is responsible for the products, Sean played pivotal roles at early start-ups Greenplum, Alpine Data Labs and Platfora.

Put the Power of Data in the Hands of Business

by Scott Etkin

Sean Keenan of UNIFi Software discusses the steps of data integration that enable an organization to compete in a real-time economy, automation, and the advantages of taking IT out of the equation.

As the advantages of data democracy become obvious, enterprises are committed to breaking down the barriers between their data and the people who use it to drive the business. But talking about being a data-driven enterprise, data democratization, and having a culture of information sharing is one thing. Doing it requires tools that put clean, normalized, and actionable data into the hands of those who can use it to drive business decisions without the lag that typically accompanies IT requests.

Sean Keenan of UNIFi Software spoke with *Data Informed* about the barriers to data democratization, the advantages of providing business users with greater access to data, and the tools that make it possible.

Data Informed: What do you think are the barriers to true data democratization throughout the enterprise today?

Sean Keenan: The lack of tools that the business user can relate to and use has been the biggest barrier to this movement. Data wrangling, preparation, integration, etc., have always lived in the hands of IT because the tooling focused on this problem has always been built with the engineer in mind. Most every industry analyst will tell you this work consumes 80 percent of the effort to get to insight, so it's a logical next step to get the business involved and to own this process. This isn't as easy as it sounds. It requires re-thinking the whole process from end to end and designing an application with the goal of "consumerizing" the enterprise, in my mind. This is what we have done at UNIFi.

Sean Keenan,
co-founder and VP of Products, UNIFi Software

Can you describe what is meant by data integration?

Keenan: The concept of data integration, at a high level, is the process of combining multiple, disparate datasets, regardless of location or format, to enable more informed analysis. There are six key components that make up this process: acquisition, discovery, data cleansing or enrichment, normalization, transformation, and preparation for visualization. For a solution to be considered enterprise ready, it must support all aspects of the end-to-end process for a user to be successful.

Can you briefly describe each of the stages of integration as you have just defined it?

Keenan: Data acquisition is enabled by establishing connectivity to a data source no matter location or structure, and then extracting the needed data into a centralized environment. Data discovery is the ability for an end user to quickly uncover what data exists and then understand the context of that data. Enrichment allows for an end user to build new computed attributes to augment the raw data and enhance their analytics outcome. Normalization and transformation are the process of structuring the data in a way that the analysis tools can consume this output and provide a seamless experience to the end user.

Describe the typical data cleansing and normalizing process.

Keenan: Data cleansing and normalization are key components of enabling successful analytics. Data cleansing can take many shapes and forms, but the initial step is for an end user to profile the datasets they are working with to understand any variances or anomalies in the dataset.

Something as simple as a string value being in upper case in some instances and lower case in others, or true, invalid data stored in a particular attribute, are common occurrences that an end user will face. Once discovered, a user will apply business rules to ensure that the attributes used in the analysis have conforming or cleansed values. This, typically, is an iterative journey for the analyst.

Data normalization also is a key part of any data processing pipeline. Data can arrive in many different formats and granularities. To ensure the insight they are looking to attain can be achieved, analysts must structure the data in a common format across the various datasets they are working with. This enables the last-mile analytic tools to consume the output that the analyst has generated and thus achieve analytic insight.

What challenges or issues does this process present to the business?

Keenan: Data integrity and format are key challenges for business users. From inaccurate reporting to challenges with consumption in their analytic tool of choice, these issues can be time consuming and frustrating experiences. Today's tools to address these issues are built primarily for the technologist, not the analyst, so an analyst typically must rely on others to deliver the solution. This can significantly impact time to insight. The key to enabling the right user experience is to provide automation and recommendations where appropriate and provide an application that enables an analyst to build and assign these rules in a simplistic way.

How can automation address these issues?

Keenan: Through profiling of source data and metadata enrichment, the next generation of tooling can make significant headway in uncovering the existence of these issues and ultimately suggest the potential solution. This type of tooling puts the analyst in a position of simply selecting the right pathway forward rather than focusing on how to create the rules that may or may not deliver the desired end result and again impact time to insight.

What does this mean for business users? What does this enable them to do that they could not do previously?

Keenan: The ultimate desire of the business is to have ready and simple access to data. By providing a user experience built for a business user that enables them to participate in the data acquisition, discovery, and preparation process, this single-handedly will have the biggest impact on time to insight versus any other technology advancement. This experience will open the floodgates for the business to embark on tackling use cases they have always wanted to investigate but never had the tooling or technical resources to accomplish.

What does this mean for IT? Does this change the role of technical staff and/or free them up for other tasks?

Keenan: Enterprise IT is drowning today in requests for data from the business. This is the biggest source of contention between IT and the business constituents they serve. Enabling the business to own this process from end-to-end will allow IT to focus on the maintenance and advancement of the underlying platforms that make this experience possible. IT should not be in the use-case delivery business; it should be focused on delivering the platforms that serve up data to the business in the way business wants to interact with it.

What's the bottom-line impact that automation can have on the business?

Keenan: Automation will be the key enabler to empower the business in a true self-serve manner around data. Data preparation presents challenging problems to solve, and the more that the tooling automates or provides recommendations around the solution, the more successful the business will be at leveraging and implementing these solutions.

Your company, UNIFi, offers software that automates data cleansing and normalization. What distinguishes your product from others on the market?

Keenan: UNIFi is the only vendor to deliver a solution for enterprise-class data integration and preparation that is purpose-built for a line-of-business user. The majority of the intellectual property within the UNIFi solution focuses on delivering a simple, elegant user experience that will enable a business user, for the first time, to be self-sufficient in the complete process of data integration from acquisition through visualization.

UNIFi has leveraged a number of machine-learning and advanced analytic libraries as part of our back-end server stack to deliver an experience a business analyst can engage in. This exposes itself in a number of ways throughout the application; whether as part of our recommendation engine that guides the user down the appropriate path in data preparation, doing sophisticated data format detection and parsing for multi-structured and unstructured data, or simply automating steps within our job workflow based on UNIFi's rich metadata repository. These are fundamental differentiators of our solution that make us the only true solution built specifically for the business.

Where UNIFi competitors excel at individual elements of the integration challenge, or deliver sophisticated tools to a very specialized audience such as data scientists, or deliver elegant solutions for small to medium size businesses, only UNIFi is delivering a comprehensive, end-to-end solution for today's inquisitive enterprise business analyst.

How can someone give UNIFi a test drive?

Keenan: It's really easy to get started with us. We can install anywhere – cloud or on-premise – and are super lightweight from a computing and installation standpoint. We have a journey we like to bring customers through that demonstrates the value we can bring to the organization and this can happen in as a little as a couple of days.

Scott Etkin is the managing editor of Data Informed. Email him at Scott.Etkin@wispubs.com. Follow him on Twitter: @Scott_WIS.

DataInformed

Big Data and Analytics in the Enterprise

Data Informed is the leading resource for business and IT professionals looking for expert insight and best practices to plan and implement their data analytics and management strategies. *Data Informed* is an online publication produced by Wellesley Information Services, a publishing and training organization that supports business and IT professionals worldwide.