Revolutionary BI
A Vision for Business Intelligence

Charles D. Kincaid, Engagement Director, SAS Center of Excellence, COMSYS
革命性的BI：商业智能的愿景

动机

我们使用互联网的方式在过去十年中发生了显著变化。在那个时候，特定的互联网交互模式出现了，并被证明是成功的。这些模式已逐渐渗透到公司业务智能（BI）或业务分析（BA）报告中。然而，我们知道经常需要“跳出框框思考”，特别是在公司内部有困难时。这篇论文旨在让人们以新的方式考虑业务智能和业务分析报告，以增加其成功和对公司价值的可能性。

在《连线》杂志中，作者Clive Thompson提醒我们，“马歇尔·麦克卢汉指出，每当我们掌握一种新媒介时，往往会像使用旧媒介那样使用它。早期的电视广播只是让一些人坐在那里读脚本，因为还没有意识到电视可以以不同的方式讲述故事。” —《连线》杂志

我们使用商业智能和商业分析报告的方式与旧媒介（纸张报告）的方式相似。有些人开始从这种模式中解脱出来。他们意识到，就像‘Reese’s Peanut Butter Cup’一样，我们甚至可以更好地利用这些信息和分析，以及发明新的方法来创建和分布。

考虑创建报告的例子。有些公司有正式的请求流程来协调业务和IT。该流程管理报告的实施和优先级，包括需求收集的关键步骤。这是最好的方式吗？Kay Redditt和Tom Lodahl并不这么认为。在他们的一篇文章《Business/IT Alignment Redux》中，他们显示了他们的研究，该公司在正式的请求流程方面做得越来越好。运行良好的公司自己下载数据，创建自己的报告。

互联网时代的第一批商店被称为“实体店”。当实体店拥有互联网存在时，如Blockbuster，就成为了一个“点击实体店”。20世纪的报告更像是“实体店”，使用同样的想法来传递信息，但世界正在发生变化。关于BI/BA的思考和方法可以改变，通过利用适用于其他信息流的模式。

收购信息

让我们看看报告是如何创建的。对有价值报告来说，数据是关键。公司已经花费数百万美元建立数据仓库。这种结构化数据非常重要。但是，还有很多非结构化或非正式的数据，存在于公司内部、网络上或个人的计算机上。

让我们来看看Web 2.0中数据的使用方式，看看我们可以学到什么。Mash-ups如Yahoo Pipes已经成为非常流行的方式，将原本相关的数据源整合在一起。Web开发中，Mashup是一个组合数据来自一个或多个来源的web应用程序。术语Mashup意味着快速、轻松地集成，通常通过访问开放APIs（Ed：公开定义的APIs，供他人使用）和数据源来生成结果，这些结果不是原始数据源的原始原因。

 Mash-ups如Yahoo Pipes，已经成为将不相关数据源集成在一起的一种非常流行的方式。

“在web开发中，Mashup是一个web应用程序，它汇集来自一个或多个源的数据，以创建一个整合的工具。Mashup术语意味着快速、轻松的集成，通常由访问开放APIs（Ed：应用编程接口，它们是公开定义的，供他人使用）和数据源来实现结果，这些结果不是原始数据源的原始原因。一个Mashup的例子是将Cartographic数据从Google Maps添加到Real-estate数据，从而创建一个新且独特的Web服务，该服务不是由原始数据源提供的。” —维基百科
Google Maps is a popular data source, because Google publishes an easy to use API. People have mashed it up with real estate data, gas prices, hurricane data, and multiple data sources to help the LA Fire Department fight fires more effectively.

The data sources don’t have to be limited to traditional options, such as a Database Management System or Spreadsheets. API’s (Application Programming Interfaces) are out there for many kinds of data via unusual mediums or sources. SAS can pull data from web sites or RSS feeds. Again from Wikipedia...

“RSS (an abbreviation for Really Simple Syndication) is a family of Web feed formats used to publish frequently updated works—such as blog entries, news headlines, audio, and video—in a standardized format. 2 An RSS document (which is called a “feed”, “web feed”,3 or “channel”) includes full or summarized text, plus metadata such as publishing dates and authorship.”

— Wikipedia

Enterprise Guide (EG) is a great tool for integrating data from multiple sources. Using it and the powerful F ilename statement we can create our own mashups. Weather, stocks, census data, page-rankings, number of “Digg”s for trailers of new films, or even sport scores can all be pulled into SAS data sets for integration into analyses to support decision-making. Once we’ve built an ETL project in EG we want to then be able to publish it to a collaborative server for others to use and comment on.

Once you start thinking outside the data warehouse box, then who knows what may be valuable? Have you heard about Google’s Cold and Flu Tracker? By using the number of queries related to colds and flu, Google is able to track the arrival of the flu virus in different parts of the country up to two weeks sooner than traditional surveillance systems; the sources of data are not fixed.

What if a sales forecaster for a consumer packaged goods company tied-in pictures of stores with their forecasts? The delivery person would take a picture every week and the pics displayed on the forecast graph. The forecaster could watch for visual changes whenever there is an unexpected change in sales volume.

One source for this idea comes from some YouTube videos. In 1 minute and 36 seconds you can see a daily snapshot of a person from over an 8 year period. Several people have done this. You can see changes in that person and sometimes in the background. Having this video history of sales locations (or of anything you think would be helpful) could be an undiscovered gem. Maybe no association would exist, but if a trend in sales could be correlated to changes in the pictures, the forecaster could use that information to predict future trends.

Soon, when image recognition software becomes mainstream, the forecasting software may be able to find those correlations for you. Apple has already introduced Face Recognition software into it’s iPhoto application. Once you identify a face in a photograph it finds that same face in all your photos and tags them for you.

Another example of using new data sources is Chris Hemedinger’s Mining the VP debate results according to Twitter — with SAS. He used an RSS Feed to gather informal VP votes from people all around the country after the Vice-Presidential debate. That delivery person above could be sending Twitter-like inputs to quickly capture and take advantage of useful information.

Within a company, there could be many data stores that have cropped-up over time as different business units try to solve their problems. Add to this the naturally inconsistent way that data has been delivered to the web and you have data anarchy, as coined by Todd Pearsall of the COMSYS BI Practice (formerly known as Praeos). Todd has a methodology for the data integration that is needed to support this data anarchy.

He outlines it in more detail, but it’s essentially something like this:

• A data search tool is set loose on your on-line transaction processing (OLTP) systems to scan for data relationship like a web spider building a search engine inventory (technically similar to Exceros’ approach for data mapping, but on a larger scale). The data in table XYZ of system 1 is likely related to table ABC of system 2 because there is a column where 78% of the values match if you truncate the right 3 characters.
• The server tool inventories the relationships it finds and defines a confidence level for that relationship.
• A user with an analysis or ad-hoc reporting GUI can select fields and pull back data to investigate (the UI is using the highest confidence link between the requested data to create the extract).
• If the relationship is wrong the user can flag it so it doesn’t get used again in their query.
• The more times the relationship is used (and by more people) the confidence level of the relationship is increased.
• If a critical mass in usage can be reached, incorrect relationships will become irrelevant and correct paths always selected.
dfPower from DataFlux can help here. From the web-site, dfPower can be used to:

- Profile data to understand the actual content of the data before ETL and ensure that the data brought over is consistent and useful
- Create consolidated master records from disparate customer data sources.
- Uncover relationships within your enterprise by validating:
  - Primary/foreign key relationships
  - Cross-table relationships
  - Cross-database relationships

Even Microsoft has realized the value of this approach, though their tools are a long way off.

Finally, for free form text data, SAS has just come out with a new product, SAS Content Categorization. I have not seen it, but this is exactly the right direction.

The program

“...applies natural language processing and advanced linguistic techniques to automatically categorize large volumes of multilingual content that is acquired, generated, or exists in a repository. It correctly parses and analyzes content for entities and events, which are then used to create metadata and trigger business processes.”

— SAS.com

What a great beginning for enhancing the way that unstructured data impacts your decision-making process.

Using Information

Now that they have the information, our users have to have the tools that allow them to join, sort, subset, aggregate, graph, analyze, etc. SAS has applications that do this well (Enterprise Guide, Web Report Studio and JMP, for example) and they’re only going to get better. In our new perspective, how do we want to think different, as Apple would say, about the tools and add even more value?

A valuable tool to give the people in the sandbox is the ability to generate multi-variable graphics, make them interactive, and connect multiple graphs together for greater functionality. Most good questions are complex and the information needed to answer them is complex. Packages like X-Lisp-Stat or ViSta have had this for a long time. JMP and the newer SAS application, Visual BI, also allow this capability. One can often see trends or patterns that could not be seen in other ways. Stephen Few of Perceptual Edge recently gave a seminar on BetterManagement.com about what can be done to visualize trends and patterns using JMP. These features give analysts a powerful lens through which to bring the complexity of our problems into focus. If we tie this capability to the BI/BA report, then we enhance the static, aggregate, tabular view with a dynamic, interactive, visual component. We can even take the new medium to another level by creating video reports. These short, specific videos would combine dynamic graphics with audio overlay to bring the story told by the data (historical and predictive) to a sharper relief.

Another improvement is, clearly, to convert “business intelligence” to “business analytics” through the use of Statistics (or Analytics). Due to the growth in computational power, great advances are being made by my fellow Statisticians. Some of the concern from the past about opening the Pandora's box and providing analytics to the “layman” is alleviated by these advances. For example, assumptions about normality, are not as necessary anymore because we can rely on techniques that are adaptive to the nature of the specific problem. We can make these advanced analytical techniques available in useful ways to our report builders. Doing so would still require some education and safeguards so that the techniques are used appropriately and the decisions are based on good, robust results. That’s not an easy task, but it can be done. Stephen Few gave another webinar about Risk Analysis and Prediction. He proposed that advanced analytical techniques can be understood by the broader audience “of knowledge workers” without the advanced statistical teaching if they have the right basics and the right tools. Web 2.0 is going to continue and expand upon enabling this broader audience to use and reap the benefits of advanced statistics.

This enabling works best when Statisticians and Application Developers work together. We've built applications for companies in Manufacturing, Market Research, Health & Life Sciences, and Telecommunications that bring advanced statistical techniques – State Space Analysis, Principal Components, Time Series, and Control Charts – into the hands of those making the decisions.

Another way to accelerate the creation of BI and BA Reports is to watch the sandbox and look for patterns. An analogy that is in no way out of the 21st Century is the establishment of sidewalks at a university. The story goes that when the university was first built they did not lay sidewalks. Instead they
let people walk where they wanted to. Wherever the paths were formed, that’s where they put a sidewalk. We can do the same in reporting. We will track the reports (tables, charts and other output) that people generate in their day to day work. Or we record the drill down paths taken in OLAP Server or something like Futrix. When the same, or similar, paths are generated repeatedly there may be some value there that can be taken advantage of. Identifying the patterns may be easiest with a custom-built web application, but it’s possible to do with the other applications mentioned above. Once we see the paths, we give them an alert that suggests standardizing the report. Then we create a stored process that recreates that output and put it into the mix (see below) for consideration. We can also give them a “publish” button so that if they have a report that they think is particularly useful, they can publish it to the collaborative server quickly and easily.

What else can we do, now, to bring these reports into the 21st Century Global Village? Here are some more ideas to get the creative juices flowing.

Collaboration

The sandbox should have a way for people to collaborate on what they’ve done. Neil Raden in a 2007 paper titled BI 2.0: Pervasive, Intelligent and Timely said, “Instead, new analyses will be built incrementally, using social networking with your peers, through tagging, semantics and intelligent agents.” However, we don’t want to create reports that are “junk mail,” as if they are automatically valuable to the users. If we give people ways to provide feedback, filter, and improve the reports, then we make them more targeted toward their consumer and enhance their value to the company.

In The Real World™ people post photos, blogs, and videos for others to see. Visitors comment and respond to the postings. By extending that concept to BA reports (created by any of the methods above) we tap into a company’s greater human intellectual capital. The technology is there but the culture of the group obviously has to support that kind of openness and vulnerability. For example, we’ve built systems that allow statistical modelers to publish their work from within Enterprise Guide to a collaboration site. Users from around the world can comment on a model’s effectiveness in various scenarios. Other users can then learn from the comments and the modeler can improve their reliability and robustness.

This approach would effectively be a peer review of the user generated reports. Employees are able to a) refine the “story” the reports tell, b) brainstorm new and better ways to “tell the story”, and c) cross-pollinate ideas from one domain to another. Furthermore, we avoid the cumbersome and ineffective request, review, and requirement gathering process of the past, yet still allow the “shot gun” method of generating reports that we’ve been discussing to not overwhelm the user community. This sort of collaboration can be as easy to use and as popular (within the appropriate communities) as something like Flickr.

The collaboration also gives statisticians and other analytical experts the opportunity to weigh-in on the statistical techniques being used. Tags and other metadata (see below) associated with the statistical models can be sent in alerts on the statistician’s dashboard. A review of the reports will identify places where training or tweaking needs to be done, as well as instances where novel, valid usage should get disseminated to others. Even if there aren’t statistics in a particular report, a statistician or other analyst may see an opportunity to suggest it.

The collaboration could be done in an asynchronous discussion forum, or via an online interactive environment. Consumers, statisticians and data people can use webinar tools such as iLinc or GoToMeeting to dig deeper into reports. They can look at the data that’s available, discuss the meaning, talk about any analysis the analyst has put together, and come to an agreement on a usable and efficient report format.

In the asynchronous style, we provide the wiki capability of everyone being able to edit the reports. For example, suppose I just got an e-mail saying that Dave put out a new report, so I go out to the site to review it. It looks great, but I think if Dave added variable X to the table he could see some interesting results. Or, as a statistician, I see that we could overlay some optimization. I make that change, add a new description and save it. Dave is notified via e-mail that I’ve made the changes. The version control system allows him to see the changes I made and back out of any he doesn’t care for. We just cut the development time down quite a bit.

Navigation

Another feature of the New Web is the tag. A tag is metadata, typically a keyword or phrase, that describes the report or report component. As metadata the keyword or phrase does not have to be in the body of the report, which provides a way to expand the report’s exposure without cluttering it up.
The tags could be open and unstructured like they seem to be for the blogs on sas.com, or they could be attributes analogous to the genes of an organism as in the Music Genome Project.

This project was started by

Will Glaser, Jon Kraft, and Tim Westergren to “capture the essence of music at the fundamental level” using over 400 attributes to describe songs and a complex mathematical algorithm to organize them.

A given song is represented by a vector (a list of attributes) containing approximately 150 “genes” (analogous to trait-determining genes for organisms in the field of genetics). Each gene corresponds to a characteristic of the music, for example, gender of lead vocalist, level of distortion on the electric guitar, type of background vocals, etc. Rock and pop songs have 150 genes, rap songs have 350, and jazz songs have approximately 400. Other genres of music, such as world and classical, have 300-500 genes. The system depends on a sufficient number of genes to render useful results. Each gene is assigned a number between 1 and 5, in half-integer increments.

Given the vector of one or more songs, a list of other similar songs is constructed using a distance function.

To create a song’s genome, it is analyzed by a musician in a process that takes 20 to 30 minutes per song.

— Wikipedia

Either way this metadata connects reports from a wide variety of sources. If someone is looking for a report related to price points, they search on the tag “price points” or click on “price points” in the “tag map.” Enterprise Guide, for example, could be extended so that users can add tags to reports they publish.

Now we have these reports with tags, but how do you utilize the tags to make the multitude of reports more accessible to the user? By adding the tags the reports now live in a highly multi-dimensional domain. We can use them to navigate the reports along their conceptual and business relationships.

The typical “tag map” shown with blogs or other web sites is a possible way for users to traverse these dimensions. The mind map technique for displaying these types of relationships would be even better. The Literature Map uses that technique to help you find new authors (though I don’t know what their distance measure is). You input the name of an author that you like and the site displays a graphic with that author in the center and other authors spread out around them. You can see who is near them in some “space” defined by the web site owners.

Analytics Solution has a similar capability in the tree view that represents customer site journeys, and in the Treeview applet.

In our BI/BA system, a report that you use would start as the center point on the map. Other reports that are similar to your report are shown around it. The closer they are the more they are like your report. Determining what “close” and “alike” means could be based on what they share and don’t share, for example how many tags, data sources, variables or other characteristic are in common between the two. Now click on one of those points and it will recenter with new reports around it (including yours). The report itself can pop up immediately or with another double-click. With this method you can traverse the tree to explore the reports that people have created.

Now combine that with something like Apple’s Cover Flow for iTunes. This technology “displays all the album art in your music collection in one easy-to-navigate interface that mimics a CD collection or jukebox selection. Move between covers using the forward and back arrows” or on the iPhone using just your finger. Bringing this to the BI/BA interface would allow the business user to see a thumbnail of the report (hmm, maybe that’s an “Ugh!”) as they traverse along this multi-dimensional space.

Sharing

So, awesome data has been fed into the system, brought together in unique ways, and displayed in reports that add significant value. The technique being used allows a company to generate hundreds and hundreds of BI and BA Reports. Is that a good problem to have? Now how do you find the reports that are helpful to you among all of these. The Cover Flow Mind Map is a cool way to do it, but it may get old fast. It’s designed more for an exploratory quest. Amazon and Google give us some ideas here. If you are looking at a book, say on Amazon’s web site, they will also show you a section with “Customers who bought this item also bought...”

Their purpose is to cross-sell their products and if we think of our BI and BA Reports as products, then we are doing the same thing. They are able to provide the suggestions by watching what people do. If we watch what reports people look at, then implementing their techniques can be pretty easy. Eventually, we’ll have the following scenario: You’ll be out on our company intranet looking at a report in SAS’ Information Delivery Portal (IDP), and you’ll see a sidebar that says “People who have viewed this report have also viewed the following...” We can do this for our employees as easily as we can for our customers. Now we’ve given a passive, automated way to make suggestions for which
reports to use. People don’t have to go out and create their own reports. If someone else has already done the work you don’t want to reinvent the wheel. Adding ratings to the reports based on how helpful or informative they are can further add to their value.

Now that we’re tracking what people view, let’s take that meta-view to the next level. Let’s pay attention to the patterns that show up in the reports that people view in tandem. We might see in those patterns a suggestion of new reports that are a combination of the report components or of the data within the originals reports. Of course, if someone sees a report that is particular helpful to them they may want to share it with someone else a la Digg or StumbleUpon.

We would also have an alert portlet in the IDP that would show the “hot” new reports. By looking at the “hot” reports, others might find a new way to better present their own data. Statisticians could be watching to see what reports bubble to the top. A quick discussion with the users of the reports can identify potential analytics to add. They’ve now added value to the report by taking the historical look of BI and making it forward-looking.

By tracking what people view, we can also search for leading indicators. Remember Google’s Cold and Flu Tracker above? If people start looking at a particular dataset or particular set of reports in greater number, maybe there’s something going on there. What do those reports show? Why are people looking at them? Is this an “early warning” of something that we should know about? What do our people know that hasn’t made it up the ranks yet? Again, we’re applying the standard web analytics to a new domain. SAS’ Interaction Management tool based on Event-Triggered Marketing is a good example of using these statistical methods. This solution tracks the behavior of customers and places them in one of a set of defined states. For example, a credit card company may have a state defined for the way a business traveler uses their card. If that customer changes states rapidly, that could trigger an event. Moving from a high volume state to a low volume state, could trigger a call to make sure they are not switching to another provider. Now what if we applied the idea to our reports? If people’s report or data usage changes states, an event might be triggered to find out why.

Many breakthrough ideas happen by taking something that’s standard in one domain and applying it to another. The methods above or others that increase the exposure of valuable graphs and data can contribute to cross-pollinating of ideas.

**Conclusion**

There are a lot of ideas in this paper and a lot more coming out every day as people grow more and more familiar with the new medium available to us and what it can do. Some of these ideas are implemented now. You can see them in SAS products, and others you can see in the work that COMSYS has done for our clients.

This paper is intended to get BI developers and Business management thinking about their BI reports, alerts, tools, etc in new ways in order to achieve faster, easier, more valuable use of the data and analytics that are available. Changing the way of thinking for people coming in to the business world will not be difficult at all. They live with these techniques and approaches the way we live with Excel. They use these tools every day and have done amazing things. Bringing these techniques to the way we create, deliver, disseminate, and collaborate can pull the standard, tired business reporting into the 21st century.