ABSTRACT

Business Intelligence (BI) and Analytical requirements have continued to change over the years as clients have recognized the power of analyzing and reporting on unstructured data. Unstructured data can come in a variety of different formats, customer comments and feedback, textual information, and dynamically changing source files, to name a few. This paper will focus on how organizations can leverage BI and Analytics to accurately identify and interpret trends buried deep in unstructured data. Through data consolidation and categorization it is possible to quickly pinpoint and resolve customer complaints that point to larger business issues. This enables organizations to get out in front of problems before they become crises. It also provides a continuous stream of real-time insights flowing from the front lines.

Introduction

Ever wonder how you can take data that looks like Figure 1 and turn it into an asset that will allow you improve your organization’s decision-making process?

![Figure 1 Customer Service Notes](image-url)
Free-form text data, such as that shown in the “Note” field in Figure 1, is generated in large volumes in most organizations every day. The most common sources include customer feedback from a company’s website, notes taken by the Customer Service team during phone conversations with customers and open-ended survey responses, e.g., from online surveys. While the richness of such data is nearly universally recognized, many organizations fail to turn this information into actionable insights due to the overwhelming challenge it presents to the analysts unequipped with the proper tools.

Many systems also require source data to be provided by third party vendors. These sources can be a point of contention as you attempt to develop and automate your BI processes. Reporting can be leveraged as part of the process to provide alerts and notifications as changes occur in the source systems so your warehouse does not allow garbage in and thus garbage out. As you can see from Figures 2 and 3, reporting does not mean creating a graph or table that is viewed by many users. Wikipedia states that “Reports often use persuasive elements, such as graphics, images, voice, or specialized vocabulary in order to persuade that specific audience to undertake an action.” As we are all aware these actions are not always related to an end-user being able to view sales, revenue, or some other business measure, but these actions can be related to ensuring that the business has valid data to be able to view.

---

**From:** SAS  
**Sent:** Tuesday, February 01, 2011 9:00 AM  
**To:** Bruce Bedford  
**Cc:** IT Support  
**Subject:** Exception Report

<table>
<thead>
<tr>
<th>Date</th>
<th>DailyStoreCnt</th>
</tr>
</thead>
<tbody>
<tr>
<td>25JAN2011</td>
<td>47</td>
</tr>
<tr>
<td>26JAN2011</td>
<td>47</td>
</tr>
<tr>
<td>27JAN2011</td>
<td>47</td>
</tr>
<tr>
<td>28JAN2011</td>
<td>47</td>
</tr>
<tr>
<td>29JAN2011</td>
<td>47</td>
</tr>
<tr>
<td>30JAN2011</td>
<td>47</td>
</tr>
<tr>
<td>31JAN2011</td>
<td>47</td>
</tr>
</tbody>
</table>

**Figure 2: Email Notification Report of Success in Loading the Data Mart**
Data
Unstructured data, when not handled properly, can cause more harm than good. There are many ways that data can change and the proper reporting mechanisms must be put in place to alert the responsible parties of those changes. Let’s take third party provided data for instance. Many systems rely on data provided by outside vendors. These data can be critical components of your organization’s operations, or they can be a supplemental, allowing derived business value by overlaying operational data.

For example, retailers generally offer their customers opportunities to enroll in a loyalty program, whereby certain customer behavior is rewarded with points redeemable for free or discounted products during future transactions. Loyalty programs are often administered by third parties that specialize in such programs. Customer purchases are tagged with a unique identification code that allows the third party to track retail transactions over time and present data back to the retailer that identify customer-level trends, responses to promotional offers, frequency of visits, variability of transaction values in time, etc. When a customer presents her loyalty program card, the retailer enters the unique card number along with the rest of the transaction detail. These data are collected through the retailer’s point of sale (POS) system and transmitted to the third party vendor at the completion of each transaction. The POS system, which is typically owned by the retailer, allows for the collection of extraordinary transaction-level detail, but does not record the loyalty card number. This means that transaction detail is not able to be associated with any particular customer without joining it to the loyalty program data maintained by the third party vendor. The loyalty program vendor maintains a database that must be able to accommodate the needs of multiple retailers. As needs change, the structure of the database also changes in order to best accommodate everyone. This might require the addition of new variables, the renaming of existing variables, the relocation of existing variables, and/or the elimination of variables. Thus, the loyalty card transaction history database has a dynamic structure and can be viewed as an unstructured source of data. To be able to manage the business better, it is imperative that the retailer puts mechanisms in place to ensure the arrival of data, as well as its validity and completeness.

More traditionally, unstructured textual data can provide great insight into operational issues, as observed from the perspective of a company’s customers – arguably, the most important perspective of all. For example, most contact center systems allow for free-form text to be entered by a customer service representative (CSR). This text contains nuggets of valuable information that an organization can learn from and leverage to identify patterns in product irregularity, improve operations, decrease costs, decrease fraud, increase sales, and enhance product development efforts. Insight through visualization of this data can provide an organization with the ability to identify and proactively address emerging trends, either for competitive advantage or before they become customer service problems, public relations issues or otherwise result in competitive disadvantages. Visualization is critical when analyzing such data, as the consumers of the analytical findings are frequently the busy top-level executives in an organization. These consumers must be able to make sound decisions quickly. They rarely, if ever, have time to wade through bulky reports, replete with voluminous, unstructured customer commentary regarding product or service complaints.
they require are analyzed trends and appropriate recommendations. Preferably, any recommendations will be
accompanied by predictions of outcomes. For them, converting unstructured data into easily interpreted pictures is a
business imperative. Quite simply, visualization is the difference between data being used and data being ignored.

Solution
In an effort to help Oberweis maximize their limited resources and enhance their reporting and auditing capabilities,
Zencos and Oberweis have developed processes to ensure that the data coming into the system is valid prior to
company reports being published and distributed. Audit reports are collected throughout the entire process and
distributed to the appropriate team member(s) to ensure reports can be published in a timely and effective manner.

One example of the audit and reporting capabilities uses unstructured data from the Point of Sale System (POS) from
each store. The goal of the developed solution is to provide checkpoints and audit reports in the automated ETL
process to ensure that all data coming into the system is as expected and that all reports going out of the system are
not published until all the data (unstructured and structured) is available and valid.

This begins with the incorporation of the exception reports shown in Figures 2 and 3. These reports provide
Oberweis with the ability to quickly identify changes in data, missing data, and/or incomplete data. The report is also
published when all data is available as well as a mechanism for the users to know that the data has not changed from
the stores (POS) and each store reporting the data properly. Although simple, the code in Figure 3 is the starting
point to a process that enables Oberweis to ensure the unstructured data coming in should go out.

```ods html file=out;
proc print data=storesmissing noobs;
var Date Store Name;
%if &storeCount ne 0 %then %do;
where Store not in (&storelist);
%end;
title "Stores with missing files during the last load"
run;
proc print data=staging.zerocount noobs;
where date=today()-1;
title "Stores with 0k data files during the last load"
run;
proc print data=dailystorecnt noobs;
var Date DailyStoreCnt;
title "Number of stores reporting over the past week"
run;
ods html close;
```

**Figure 4: Code for Email Exception Report.**

Once all files have been validated, i.e. the exception report only contains entries for “Number of stores reporting over
the past week”, the ETL and audit reporting can continue. This includes not running unnecessary code. The process
performs audit reporting to validate whether or not the data has already been loaded for a particular time frame and
does now run the ETL based on the response back from the report. As with Figure 4, Figure 5 is also a very simple
code example, but one that eliminates the usage of unnecessary system resources.

```%if &date eq %sysfunc(today(),date9.) and &status eq complete %then
  %put No need to run StoreLoad_Driver at this time. All data are present for
today.;
%else %StoreLoadDriver
```

**Figure 5: Code to check to see whether or not data has been loaded already.**
Once all data has been loaded into the system, data validation checks are run to ensure the daily sale totals and
individual store totals match. If a match occurs a message is published and the report is sent out to the appropriate
distribution list.

```
%macro checkReportData();
proc sql noprint;
   select
      RevContrib into :RegionTotal
   from
      RPT.rpt_sss_sales
   where
      region='Total Region'
   and
      Transdate = &transdt;
quit;
proc sql noprint;
   select
      sum(RevContrib) as RevContrib format=dollar20.2 into :StoreTotal
   from
      RPT.rpt_dailystorereport
   where
      StoreType = 'Corporate Stores'
   and
      Transdate = &transdt;
quit;
%put RegionTotal=&RegionTotal;
%put StoreTotal=&StoreTotal;
%if &RegionTotal = &StoreTotal %then %do;
   %put Totals match;
   %SendReport
   %let errorMessage=Daily report data has been updated and total match.;
%end;
%else %do;
   %let errorMessage=Region Total and Store Totals do not match for ;
   %sendErrorEmail
%end;
%mend checkReportData;
```

**Figure 6: Code to validate data and publish report.**

This logic is the final step in ensuring that the reports that are published from the unstructured data will be able to be
used to make true business decisions and the numbers will not be in question from a ETL validation perspective.
Assuming all processes pass the audit and validation reporting, the process will control the publishing of the daily
sales report shown in Figure 6.
Figure 7: Published Daily Dairy Store Sales Report.

Part of the process of publishing data and reports is selecting the proper mechanism for distribution. The reports can utilize standard SAS® programming like the email engine, or take advantage of other SAS report distribution mechanisms like batch SAS Web Report Studio® submission. Figure 6 distributes the data through ODS output, however the same report can be distributed from WRS too. This provides the consumer of the report with the ability to select how they would like to receive or view the report. The option is there to view the report online, or to register to be part of a published distribution of the report.
SAS WRS allows reports to be distributed to a list of users through a batch command process using the built-in `ReportBatch` command (example shown in Figure 9) or through the “File” → “Distribution” menu.

C:\SAS\EBIserver\Lev1\ReportBatch\rptbatch.bat --spring-xml-file C:\SAS\EBIserver\Lev1\Applications\SASBIReportServices4.2\spring.xml --repository Foundation --source "A5QPN478.C4000054" --burst --channel-age 90 --metadata-key "BurstDefinition+omi://Foundation/reposname=Foundation/Transformation;id=A5QPN478.B10001HO"

Output
At the beginning of this paper we asked how the data in Figure 1 could be analyzed to provide actionable results. Oberweis Dairy is doing just that through the powerful tools of SAS text analytics. The data shown in Figure 1 represent a sample of actual text notes taken by the company’s CSRs and entered into its enterprise resource planning (ERP) system. The ERP system is a bit quirky, in that it only allows 84 characters to populate the Note field before opening up a new record. If a CSR needs to take a note that exceeds this limit, the new record is time-stamped and entered into the database on an entirely separate line. Moreover, the timestamp is non-traditional, in that it counts minutes in ascending fashion, but seconds are entered in descending order. That is, it counts minutes down from 60 to 0 seconds, rather than the other way around, as one would expect. As a result, it is quite easy to inadvertently disaggregate the various records that comprise a complete customer note, rendering the post-hoc analysis of such notes challenging at best, when simply read by the human eye.

The Oberweis data integration strategy, developed and implemented by Zencos Consulting, addresses all of the complexities represented by the ERP data. Each day a SAS table is updated, bringing to the surface a wide array of information extracted from text notes on customers records. The table feeds a reporting macro that surfaces the primary issues in trend charts and automatically emails the results out to the executive team for decision-making. Figure 9 shows a sample of the analyzed data contained in the SAS table `Product_Complaints`.

Figure 8: WRS version of the Daily Dairy Store Sales Report.

Figure 9: Using the batch report distribution.
The variable CustNote is the result of the SAS data integration code reassembling the disaggregated note elements from the source data. The four rightmost variables illustrated in Figure 9, i.e., ProductComplaint, Qty_1010, Qty_1060 and Qty_1090 are the result of text analytics applied to the CustNote character variable. ProductComplaint is a variable that categorizes the contents of CustNote and the three Qty_ variables provide counts of key products that may be referenced in CustNote. Utilizing this table, it is a simple matter to generate a visual representation of the customer notes for use in decision-making. Figures 11-12 illustrate the power of visualization. To highlight the steps taken, the visualization process is illustrated in two parts. Figure 11 shows how the text analytics procedures extracted a spike of customer complaints related to a reduced fat milk product during the week ending Sept 26, 2010 of the rolling 8-week period from Sept 12 – Nov 7, 2010. For ease in comparison, data from the same period of the prior year are also visualized, with the prior year dates shifted to make the direct comparison as easily as possible by those who need to quickly understand the trend. Without wading through a mountain of statistical data, it is apparent that something has changed that customers feel strongly enough about to phone in and report.

<table>
<thead>
<tr>
<th>DX_CUSTNO</th>
<th>NoteDate</th>
<th>NoteTime</th>
<th>CustNote</th>
<th>ProdComplaint</th>
<th>Qty_1010</th>
<th>Qty_1060</th>
<th>Qty_1090</th>
</tr>
</thead>
<tbody>
<tr>
<td>181200</td>
<td>23SEP10</td>
<td>6:21:31PM</td>
<td>disc #2</td>
<td>Other(unscheduled)</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>181201</td>
<td>23SEP10</td>
<td>6:31:33PM</td>
<td>9000 OFF</td>
<td>Other(unscheduled)</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>181202</td>
<td>23SEP10</td>
<td>6:22:33PM</td>
<td>R call #3</td>
<td>Other(unscheduled)</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>181203</td>
<td>23SEP10</td>
<td>8:56:32PM</td>
<td>R call #5</td>
<td>Other(unscheduled)</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>181204</td>
<td>23SEP10</td>
<td>5:26:09PM</td>
<td>R call #6</td>
<td>Other(unscheduled)</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>181205</td>
<td>23SEP10</td>
<td>7:06:23PM</td>
<td>R via message call</td>
<td>Other(unscheduled)</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>181206</td>
<td>23SEP10</td>
<td>7:11:30PM</td>
<td>phone without a date</td>
<td>Other(unscheduled)</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Figure 10: Sample of table Product_Complaints, which is the post-processed result of data like that shown in Figure 1.
Figure 11: Example of an out-of-trend spike suggesting an important event had occurred in the plant.

The text analytics can take it a step further and actually identify what has changed. By further categorizing the notes comprising the spike, it was identified that the complaint related to a homogenization issue leading to milk fat separation in the bottle, as illustrated in Figure 12.
The issue was immediately resolved, as is apparent by the sharp attenuation in the volume of complaints logged the following week.

This example highlights the extraordinary power of analytics applied to unstructured data. The interpretation of such data has been reduced to a quick review of a visual trend chart, making it highly likely that important issues will be spotted quickly and managed properly.
Authors
Bruce Bedford, PhD
Vice President, Marketing Analytics and Consumer Insights
Oberweis Dairy

Prior to joining Oberweis, Bedford was with a privately held analytic consultancy serving SMB clients in the manufacturing and retail industry segments. He later transitioned to a marketing management role, where he developed various product and service offerings in the process monitoring and control space. Additionally, Bedford has held executive marketing positions in the speech recognition software industry, where he built and applied data analytics models to predict consumer behavior.

He holds MS and PhD degrees in engineering from Northwestern University, an MBA from Northwestern's Kellogg School of Management and a BS in engineering from the University of California, San Diego. Bedford is also a certified Base SAS programmer.


Ben Zenick
VP Consulting Services
Zencos Consulting

Ben Zenick is vice president of consulting services and co-founder of Zencos, a business intelligence consulting and services company based in Cary, NC. Under Ben’s leadership and vision Zencos has become an elite SAS Gold Alliance partner and Inc. 500 company, recognized as one of the fast growing privately-held companies in the US. With a strong background in designing, building and deploying extensive data warehouse and business intelligence solutions, Ben ensures that Zencos team members consistently exceed customers’ expectations through efficient and intelligent execution and delivery.

Ben’s management of public sector projects has resulted in several of Zencos’ clients receiving recognition including InfoWorld’s 2008 Top 100 IT Projects. Ben has authored several white papers and presented at multiple industry events including SAS Global Forum. Ben received his BS in Business Systems from the University of North Carolina at Wilmington. Prior to starting Zencos, Ben worked at SAS as a solutions integration specialist.