Paper 11849-2016

Visual Analytics and SAS/ACCESS® Interface for Hadoop: Improving Efficiency and Increasing Analyst Satisfaction

Rosie Poultney, 89 Degrees

ABSTRACT

Over the past two years, the Analytics group at 89 Degrees has completely overhauled the toolset we use in our day-to-day work. We implemented SAS® Visual Analytics, initially to reduce the time to create new reports, and then to increase access to the data so that users not familiar with SAS® create their own explorations and reports. SAS Visual Analytics has become a collaboration tool between our analysts and their business partners, with proportionally more time spent on the interpretation. We show an example of this in this presentation.

Flush with success, we decided to tackle another area where processing times were longer than we would like, namely weblog data. We were treating weblog data in one of two ways: (1) creating a structured view of unstructured data by saving a handful of predefined variables from each session (for example, session and customer identifiers, page views, time on site, and so on), or (2) storing the granular weblog data in a Hadoop environment and relying on our data management team to fulfill data requests. We created a business case for SAS/ACCESS® Interface for Hadoop, invested in extra hardware, and created a big data environment that could be accessed directly from SAS by the analysts. We show an example of how we created new variables and used them in a logistic regression analysis to predict combined online and offline shopping rates using online and offline historic behavior.

Our next dream is to bring all of that together by linking SAS Visual Statistics to a Hadoop environment other than the SAS LASR® Analytic server. We share our progress, and hopefully our success, as part of the session.

INTRODUCTION

89 Degrees, a SAS Partner, is a leading provider of customer engagement agency services and technology products designed to enable greater analytic insight and higher marketing ROI. We are committed to helping our clients make full use of their data through improved collaboration, tools and data access.

In this paper, we share our approach to assessing the fit of new products on behalf of our clients. Specifically, we discuss SAS Visual Analytics and SAS ACCESS to Hadoop. You will find this paper useful if you are considering choosing SAS Visual Analytics for your organization, or are a SAS Visual Analytics user.

At time of writing, we are using SAS 9.4 and SAS Visual Analytics 7.1.Please note, the data shown in the examples is from a dummy data set, however the business questions are real and the results are in line with those we have seen across clients.

INCORPORATING VISUAL ANALYTICS

VISUAL ANALYTICS ON A NON-DISTRIBUTED SYSTEM

We started to evaluate SAS Visual Analytics in late 2013 as a replacement for our in-house reporting solution, built using SAS Proc Report and delivered via a custom web portal. SAS Visual Analytics had been launched earlier in the year and promised a shorter development time for reports combined with better graphics and filtering, and an improved distribution system.

We set up a non-distributed instance of SAS Visual Analytics using Amazon Web Services (AWS). This allowed us to use the application without having to invest in hardware, and gave us the numbers we needed for a business case. Further details of the environment can be found in Pasion and Aanderud [1].

Data were summarized using a batch process and transferred to a staging area using SFTP and RSYNC. From here files were directly picked up by the AutoLoad process. This was necessary since the server running SAS Visual Analytics was not connected to our systems. We initially allowed our internal Analytic team to import data using Designer Explorer via a web browser but found the connection problematic with files around 4GB or larger. Figure 1Figure 1See Figure 1 for an overview of the load process

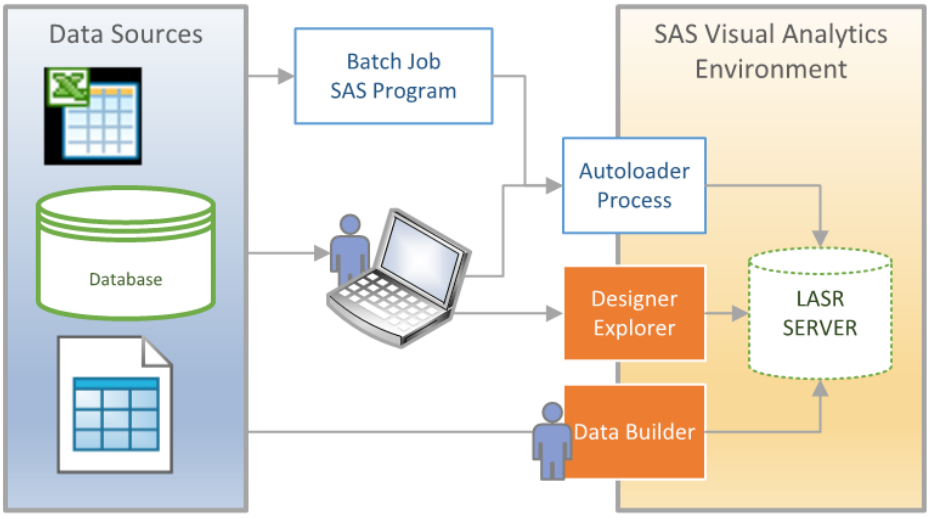

Figure 1: Data loading options using AWS

We knew we needed to build a community of users and found a consultant to provide a training course and ongoing support. A core team of analysts and BI specialists dedicated time each week to learn the functionality. Our Friday mornings turned into a show and tell that rivalled grade school!

In Q2 of 2014, we moved our first client onto SAS Visual Analytics, followed by two more in Q3 and Q4 of that same year. We transferred existing reports into SAS Visual Analytics, creating new tables and graphics. You will find, as we did, that report users want to extract data at some point, and your reports will be adopted faster if you make this easy and flexible. One of our clients was accustomed to extracting data with fields as rows rather than columns, and had built their other processes using this format. We adapted our original Proc Report codes and created stored processes within the reports.

The non-distributed system has file size constraints and we learned to keep our files to 8GB or smaller. The Analytic team realized that so many of the questions posed to us began with the words “How many customers….”, causing us to build summary views of customers, regularly updated, and containing as many variables as we could include to keep within our 8GB file size.

You can also reduce the size of files by using formats to add the text descriptions to products and product hierarchies, customer segments, store names, etc, and by calculating flags and variables within Visual Analytics. In some cases, we found we could reduce the file size by almost 50%.

MOVING TO A DISTRIBUTED SYSTEM

We have a 4-node distributed system, each with 16 cores running in a virtualized environment using the Linux operating system. Resources are shared among LASR servers and security is governed using SAS metadata and the operating system. The system supports around 100 accounts, mostly viewers, at multiple clients viewing 100 standard reports, plus internal and client users creating exploration and reports.

We copied much of the process and data still loads via the AutoLoad process, regardless of whether it is productionized data or one-off files. There are separate autoloader processes for each client and files will only load if they match the correct nomenclature. Our key concern is to keep our client data in separate areas, and this will not be an issue if your SAS Visual Analytics implementation is internal to your organization.

We have added HDFS (Hadoop Distributed File System) which we use to reload data. Once we move to the next release of SAS Visual Analytics, we will take advantage of the improved permissions and load scheduled data through HFDS, keeping the Autoload for ad-hoc files.

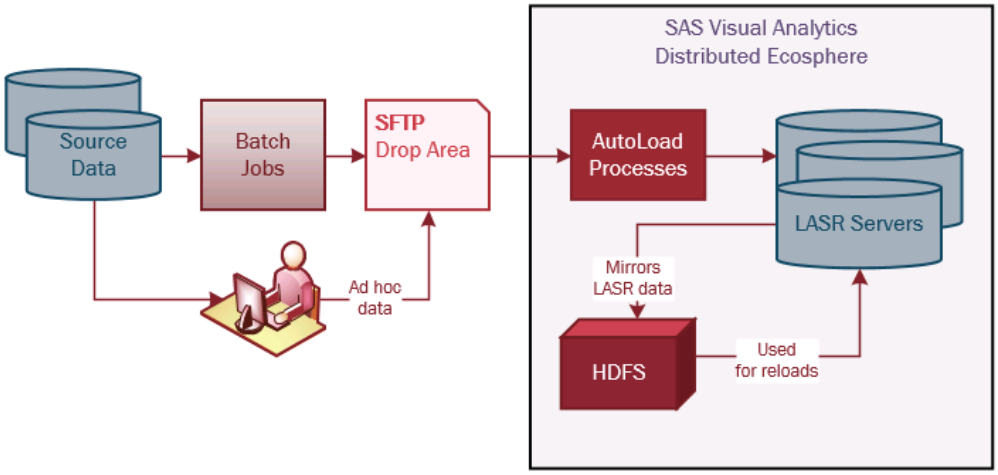

Figure 2: Data loading options using distributed system

IMPROVED COLLABORATION

As the analytics team grew more familiar with SAS Visual Analytics, we found it became more integrated in our ways of working. Where previously we had spent meetings taking notes for follow-up analysis, we were now projecting Visual Analytics onto a screen and discussing implications of the data. Projects began to move faster and we had less reruns due to miscommunication of assumptions.

At their request, we trained our business partners to create their own reports. Now our meetings frequently start with those partners sharing their reports and asking for more complex analysis. By removing the barrier to data access, we have spread the analytic load across the company and have focused analytic resources on interpretation and strategic business questions.

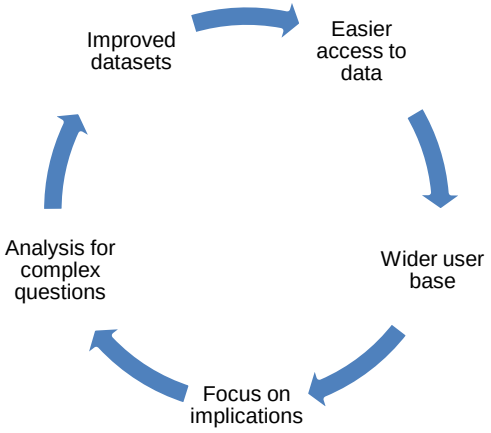

Figure 3: how increased usage leads to better insights

ENABLING USERS TO BUILD SIMPLE REPORTS

Each of our clients has standard reports and dashboards containing counts and key customer metrics overall and split by segmentations, flags, demographics and geographies. These answer many questions.

However, we realized we needed to support a highly-numerate user community used to manipulating data in spreadsheets and keen to build reports and explorations on existing tables.

We have an automated process that creates and loads customer summary tables into SAS Visual Analytics. The tables typically contain between 50 and 80 variables of the following types:

- Transaction summaries, such as sales, visits, items in the current and prior 12 months. These are split further by department, channel (online / in-store), or other relevant classifications
- Customer segmentations, model scores, geographies, preferred stores
- Customer behavior summaries, opt-in flags for different media, DM / emails received, opened, clicked, web visits, social mentions
- Demographics

We find these tables increase the number of people who feel confident creating their own reports. We provide a half-day training course covering the basics:

- Tables, bar charts, line graphs and other data objects
- Creating distinct counts, the difference between calculated and aggregated variables, custom categories, custom sorts
- Exporting data

These users, marketers rather than analysts, are familiar with the standard reports and have been part of interactive analytic sessions. They are self-selecting, often using SAS Visual Analytics as a data summary and extraction tool to answer questions that previously had to be passed to a different team.

EXAMPLE: COLLABORATION USING INTERACTIVE DATA SESSIONS

When we work with customer data from a new client, we need to share insights with our business partners quickly to enable them to focus on the questions posed by the client. Using SAS Visual Analytics, we are able to get them familiar with the data and thinking about how customers interact with the client.

This example uses simulated data for a fashion retailer and shows how we can create a basic RFM segmentation (Recency, Frequency, Monetary Value) which will inform further analyses, or be used as a sampling frame for primary research.

We loaded a file containing summary variables for 2 million customers in SAS Visual Analytics. The file contained a user ID, ZIP code and State, customer source, total items, visits and sales for the current and previous years, and current year spend in each of nine departments.

The analyst spent a day exploring the data, creating new calculated and aggregated variables, and preparing for an interactive session. Our aim was to get everyone on the team discussing the implications of the data, and to minimize the time spent restating the data.

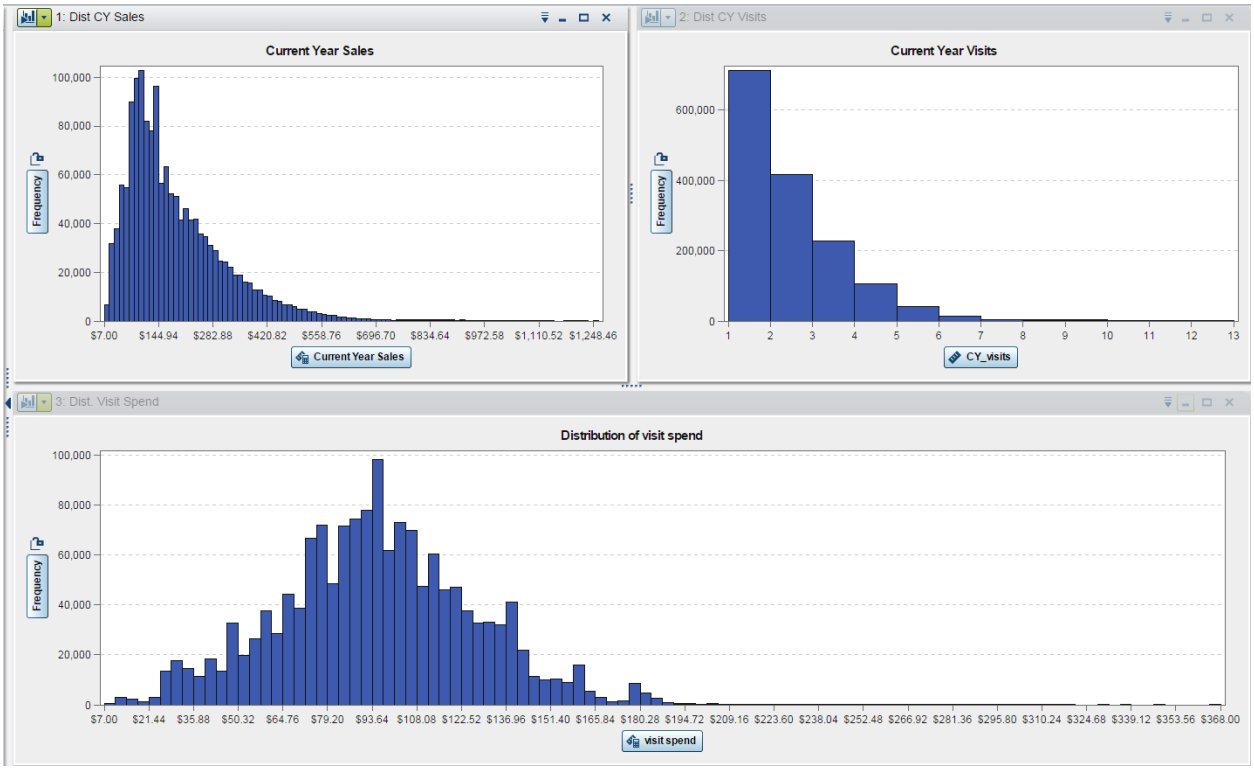

Figure 4: Histograms of three key variables created using the automated charting function in Explorer

These distributions were an immediate timesaving. Previously, we would have coded separate SAS procedures, extracted the data into Excel and created charts. However, the larger benefit is the power to make changes to the data as part of the meeting, thus speeding up the project and saving on rework. We are often asked to repeat the analyses for subsets of the data, for example:

- Remove customers with an annual spend over $1,000
- Provide the distribution for customers joining via an in-store event
- Exclude States that are new markets .

Ahead of the meeting, we created spend and visit bands using thresholds suggested by the data. These were used as a starting point for the analysis and were easy to edit. We created visit bands using a New Custom Category and spend bands using nested IF…ELSE operators in a Calculated Item.

We provided key spend metrics for the different levels of spend and visit. In the session, we were able to change the thresholds in variables shown in Figure 5. By setting up these metrics ahead of time, we were able to keep the project momentum while allowing for the potential to change what had been created. If we had presented the spend metrics in PowerPoint, we would have had to repeat the work and schedule another meeting.

Finally, we classified the combinations of spend and visits into RFM segments using New Custom Category, Figure 6. To do this, we formed a Calculated Item by concatenating spend and visit bands. Note that first you have to create a Custom Category from spend band.

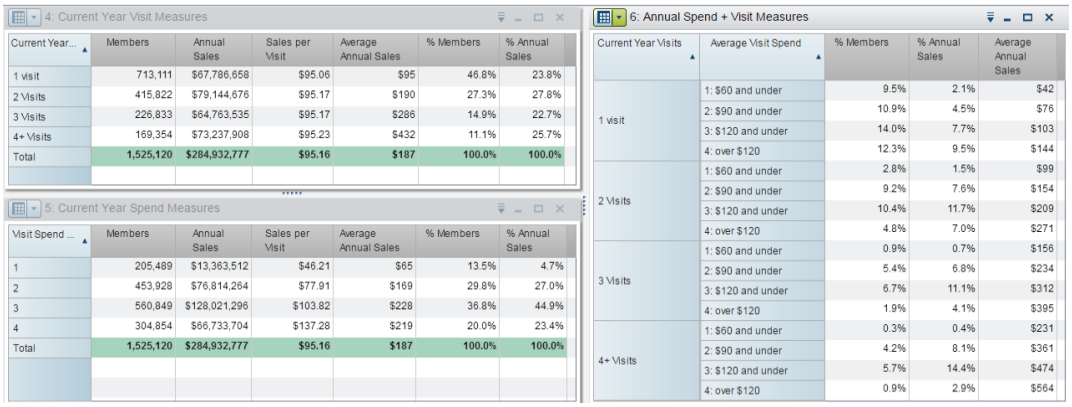

Figure 5: Customer metrics by spend and visit bands

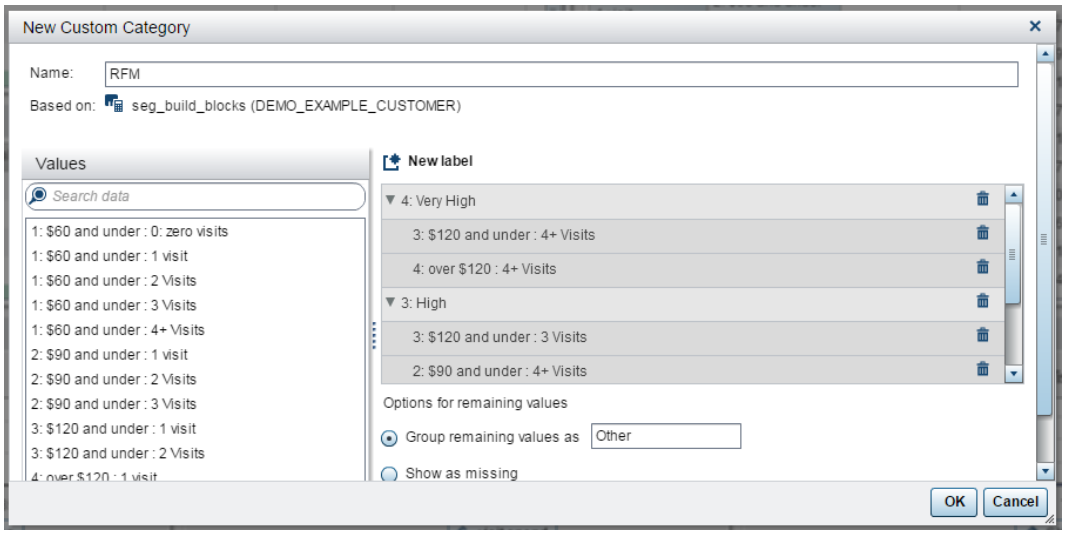

Figure 6: Using New Custom Category to create RFM segments

The RFM segmentation, Figure 7, shows that the top two groups (High and Very High) account for 24.2% of customers and for 47.6% of sales.

We emailed the report link after the meeting so participants could try changing the inputs. This approach allowed the team to focus on the strategic questions earlier in the project

- Which stores do my best customers shop at? And do they have favorite store associates?
- What are the retention rates of the different groups? And how should I incentivize them?
- How do they respond to email? Who visits the website?

9: RFM segments

RFM	Members	Annual Sales	Sales per Visit	Average Annual Sales	% Members	% Annual Sales
1: Low	568,502	$45,011,454	$73.63	$79	37.3%	15.8%
2: Medium	587,215	$104,450,768	$95.57	$178	38.5%	36.7%
3: High	268,411	$86,304,545	$103.97	$322	17.6%	30.3%
4: Very High	100,992	$49,166,010	$106.89	$487	6.6%	17.3%
Total	**1,525,120**	**$284,932,777**	**$95.16**	**$187**	**100.0%**	**100.0%**

Figure 7: Customer metrics by RFM group

TOP 5 TIPS FOR IMPROVED COLLABORATION

1. Create standard reports to get people using the tool.
2. Understand how they will use the reports.
    a. If users are extracting multiple tables, add columns into a single table.
    b. If the report is a top-level dashboard, focus on changes and targets for a few key metrics.
3. Integrate Visual Analytics into the analytic process. Present your findings using Visual Analytics instead of PowerPoint or Excel, and encourage logins by emailing reports with comments added in the email.
4. Understand the hot topics for your audience, and build in the ability to filter on these by calculating metrics as aggregated variables within Visual Analytics.
5. Build a system that can scale as needed.

USING HADOOP TO IMPROVE ACCESS TO WEB DATA

Hadoop is a file storage system that allows massively parallel processing in virtual memory on commodity hardware enabling you to perform data-heavy analytics quickly and at a much lower cost compared to the custom hardware that used to be necessary. The SAS Whitepaper *Bringing the Power of SAS® to Hadoop* [2] is an excellent summary.

A number of our clients sell high-ticket items that are typically researched online prior to purchase. Our analyses have shown that customers who are engaged with email and have recent website activity have a higher probability of shopping than their lookalikes based solely on historic transaction data. We have also seen that some behavior, such as multiple views of the same product or reviewing installation guides, are strong indicators of future purchase. If you can identify these indicators, you can trigger communications to improve your chances of staying in the customer's consideration set and ultimate purchase.

However web data isn't called big data for nothing! One client supplies us with 20GB of data per day and growing. Like many organizations, we dealt with large unstructured data by forcing it into a smaller structured form. Analysts could use session-level summaries, containing 25 to 30 measures, but had no access to the individual URL-level data.

OUR JOURNEY TO HADOOP + SAS/ACCESS

Figure 8 outlines the steps we took on the journey:

- In the original process, extracts of data at the individual URL level had to be scheduled with our data team.
- In 2014, we trialed using Hadoop as data storage, using AWS to host our Proof of Concept. We proved that we could store and access URL-level data and designed a scalable solution to be hosted in our data center. Analysts had direct access using HiveQL, but had to move data to a SAS environment for analysis.
- Our new solution uses SAS/ACCESS to connect directly to Hadoop using a LIBNAME statement, with no need to code in HiveQL

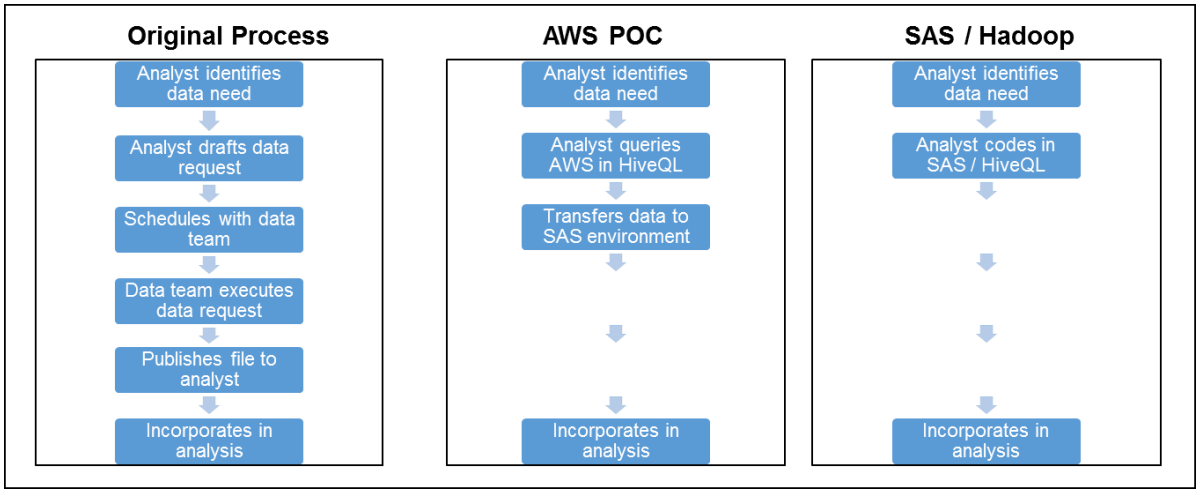

Figure 8: Comparison of analytic briefing processes

Although the move to Hadoop was initially driven by a need for cost-effective storage, SAS/ACCESS to Hadoop has allowed us to use URL-level data in more analyses. There are less people involved, less data being transferred across networks, shorter elapsed time on projects and an Analytics team focusing on answering business questions using a new technology.

EXAMPLE: INCLUDING WEB BEHAVIOR IN SHOPPER MODELS

This example uses the simulated data for the fashion retailer. We have a set of one million customers who have clicked through from an email to the website and for whom web behavior can now be linked to online and in-store purchases. We want to understand the online behavior of customers who buy upscale outerwear and how that might differ from other site visitors.

We hypothesize that likelihood to purchase is a function of the following:

- Browsing for the specific products, category, or at inspiration/look books (Hadoop + SAS/ACCESS)
- Engagement with the brand / responsive to messaging (campaign responses, number of web sessions)
- Previous purchasers (standard RFM measures, category purchasers)

We merged the results of the code in Figure 9: example of the SAS/ACCESS source code with data created using the SQL procedure in SAS BASE. We added a flag for purchase in the post period and built a logistic regression model using the LOGISTIC procedure. Even allowing for historic behavior and general levels of engagement, customers viewing the look book and specific products were four and seven times more likely, respectively, to buy those products in the subsequent month.

Now that we have reduced the time to get the data, we can replicate the analysis for categories where previously the price point wouldn't have justified the expense.

```
libname hdplib hadoop subprotocol=hive2 port=x server=”x" user=x password=x schema=x;

data weblog;
set hdplib.OUTWEAR_weblogs (keep=userid page_url );
if index(upcase(page_url), '/US/EN/CATALOG/CATEGORIES/DEPARTMENTS/OUTWEAR') then
view_outwear=1;  else view_outwear=0;
if index(upcase(page_url), '/US/EN/CATALOG/PRODUCTS/LOOKBOOK') then view_lookbook=1;
else view_lookbook=0;
if index(upcase(page_url), '/US/EN/CATALOG/PRODUCTS/INSPIRATION') then view_inspiration=1;
else view_inspiration =0;

proc sql;
create table user_web_view as
select userid, sum(view_outwear) as view_outwear, sum(view_lookbook) as view_lookbook ,
sum(view_inspiration) as view_inspiration
from weblog
group by userid
order by userid;
quit;
```

Figure 9: example of the SAS/ACCESS source code

CONCLUSION

The SAS Visual Analytics software and our Hadoop implementation have literally changed the way the analysts work, reducing the time taken to prepare data and allowing us to focus on the analysis. We have empowered our business partners with tools that allow them to quickly answer many questions themselves. Everyone is happier.

We still have many steps on our journey and will be working on three main goals in 2016:

- Continue to push the virtuous circle described in Figure 3: how increased usage leads to better insights thus allowing more people more access to deeper data. We will expand the available reports, add more data types and deliver more project work using Visual Analytics
- Develop our capabilities by using SAS Visual Statistics and link to a Hadoop environment other than the SAS LASR® Analytic server to streamline the process in our web behavior model
- Extend our toolset by exploring machine learning capabilities within SAS

REFERENCES

[1] Pasion, J and Aanderud, T. 2015. “Tactical Marketing with SAS® Visual Analytics – Aligning a Customer’s Online Journey with In-Store Purchases” *Proceedings of the SAS Global Forum 2015, Dallas, TX*: SAS Global Forum. Available at: http://support.sas.com/resources/papers/proceedings15/3352-2015.pdf

[2] SAS Institute White Paper. Bringing the Power of SAS® to Hadoop. https://www.sas.com/content/dam/SAS/en_us/doc/whitepaper1/bringing-power-of-sas-to-hadoop-105776.pdf

ACKNOWLEDGMENTS

Thanks are due to Tricia Aanderud for her patience in the early days when we were learning Visual Analytics, and for the support she and Zencos team continue to provide.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Rosie Poultney
89 Degrees, Inc
poultneyr@89degrees.com
http://www.89degrees.com