Paper 071-2013

Bigger Data Analytics for SAS® – Using SAS on Aster Data and Hadoop

John Cunningham, Teradata Corporation, Momolue Kiadii, Teradata Corporation

ABSTRACT

With the increased popularity of new Big Data clustered processing platforms, SAS analytics now has the opportunity to solve newer, bigger problems than ever before. This paper will focus on the evolution of Big Data analytics, the new data sources and types, new technologies involved, to achieve end to end analytic processing with SAS. Will specifically demonstrate the use of new Big Data technologies, SAS Analytics with SAS / Access for Aster, Aster SQL-MR, SQL-H to integrate end to end Big Data analytics on the Aster Discovery Platform, even from raw data files stored on Hadoop clusters.

INTRODUCTION

One of the key evolutions of Analytic and Enterprise Data Warehousing over the last 10 years has been the increasing importance of “Big Data”. Teradata has long been the leading provider in Enterprise Data Warehousing, focused on high performance, very large data repositories that can be used for a wide variety of Operational Support and Business Intelligence purposes. For many of these companies using Teradata, they built huge data warehouses collecting data on their customers, orders, inventory, manufacturing, financials, expenditures, support inquiries, shipping logistics, etc – all in an effort to gain a 360 degree perspective on their customers and their operations. For many Teradata customers, their data warehouses grew well into the petabytes, and did so in a scalable, performant fashion, for hundreds or thousands of users around the enterprise.

Teradata customers manage Petabytes, and use SAS == Big Data Analytics. Right?

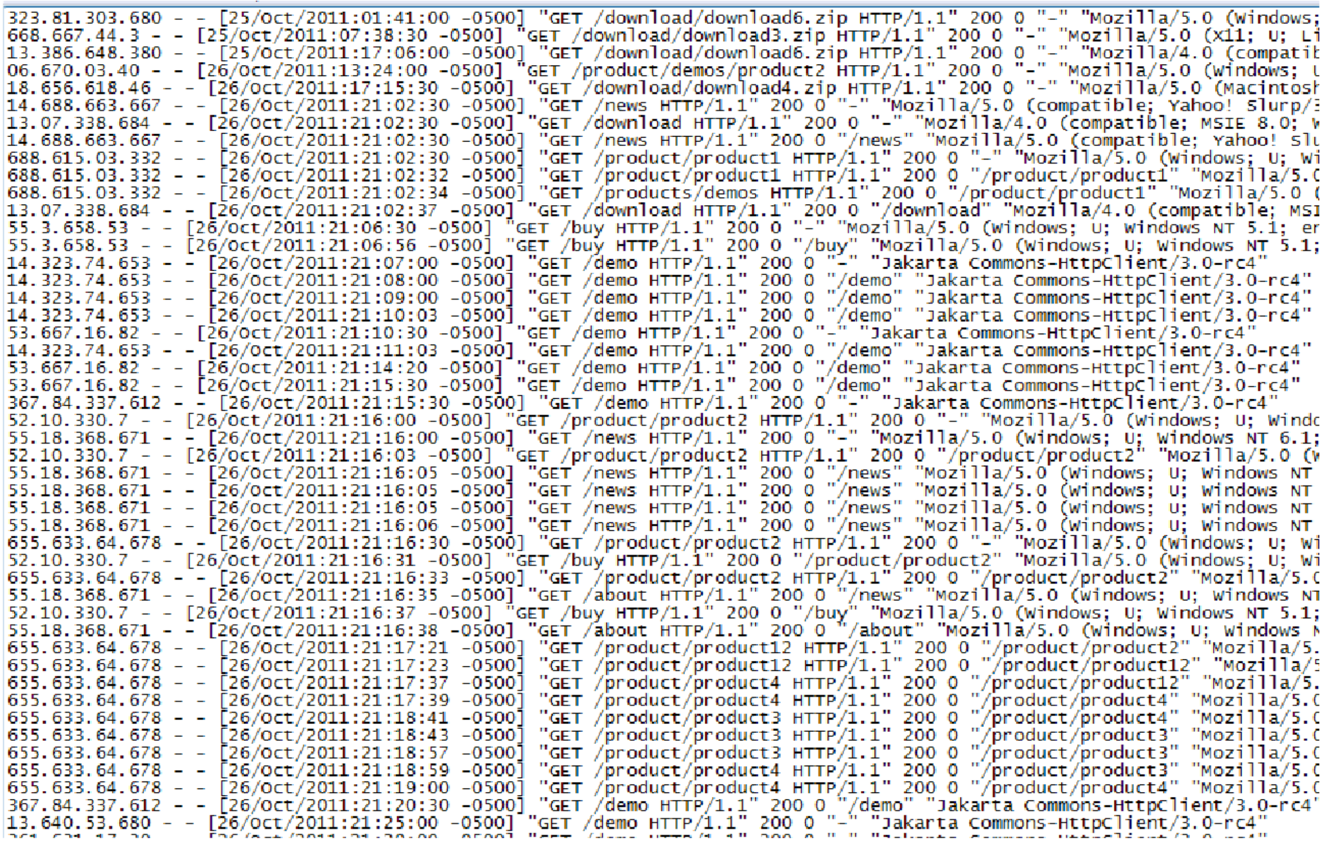
Not really. Big Data Analytics takes a twist on the drive for very large repositories of data. It’s urgency is driven by companies looking to take advantage of the growing wave of data being collected by their systems, web servers, phone switches, network data, etc., that can help companies better understand their customer behavior and processes.

As you might expect, data from these machine generated sources are enormous, tracking billions of interactions, much of which is extraneous and analytically meaningless. Until just recently, these logs were discarded because the relative cost of the disk space was higher than the marginal value found within them. Now, with the costs of Linux servers and drive space plummeting, Hadoop stands out, because it enabled users to manage very large clusters that could split inquiries into many different sub-inquiries, split in parallel across many different servers – and then returns the result that could then be reduced to the specific results.

INTRODUCE HADOOP, MAPREDUCE, AND ASTER

With the evolution of new technologies like Hadoop & Map Reduce, technicians now have a way to manage enormous server clusters to help them process their big data collections.

However, Big Data on Hadoop isn’t for everyone. As those “High Value Analytic Nuggets” are frequently buried in some complex log files, they are not typically easily accessible using a simple Query and Reporting tool. For example, the web log below is tracking web server histories of different users, different transactions, not something that we’d likely include for a business intelligence report for a business audience – may be something interesting there, but it’s not easy to find.



```
323.81.303.680 - - [25/Oct/2011:01:41:00 -0500] "GET /download/download6.zip HTTP/1.1" 200 0 "-" "Mozilla/5.0 (Windows;
668.667.44.3 - - [25/Oct/2011:07:38:30 -0500] "GET /download/download3.zip HTTP/1.1" 200 0 "-" "Mozilla/5.0 (X11; U; Li
13.386.648.380 - - [25/Oct/2011:17:06:00 -0500] "GET /download/download6.zip HTTP/1.1" 200 0 "-" "Mozilla/4.0 (compati
06.670.03.40 - - [26/Oct/2011:13:24:00 -0500] "GET /product/demos/product2 HTTP/1.1" 200 0 "-" "Mozilla/5.0 (Windows;
18.656.618.46 - - [26/Oct/2011:17:15:30 -0500] "GET /download/download4.zip HTTP/1.1" 200 0 "-" "Mozilla/5.0 (Macintos
14.688.663.667 - - [26/Oct/2011:21:02:30 -0500] "GET /news HTTP/1.1" 200 0 "-" "Mozilla/5.0 (compatible; Yahoo! Slurp/
13.07.338.684 - - [26/Oct/2011:21:02:30 -0500] "GET /download HTTP/1.1" 200 0 "-" "Mozilla/4.0 (compatible; MSIE 8.0;
14.688.663.667 - - [26/Oct/2011:21:02:30 -0500] "GET /news HTTP/1.1" 200 0 "/news" "Mozilla/5.0 (compatible; Yahoo! Sl
688.615.03.332 - - [26/Oct/2011:21:02:30 -0500] "GET /product/product1 HTTP/1.1" 200 0 "-" "Mozilla/5.0 (Windows; U; Wi
688.615.03.332 - - [26/Oct/2011:21:02:32 -0500] "GET /product/product1 HTTP/1.1" 200 0 "/product/product1" "Mozilla/5.
688.615.03.332 - - [26/Oct/2011:21:02:34 -0500] "GET /products/demos HTTP/1.1" 200 0 "/product/product1" "Mozilla/5.0
13.07.338.684 - - [26/Oct/2011:21:02:37 -0500] "GET /download HTTP/1.1" 200 0 "/download" "Mozilla/4.0 (compatible; MSI
55.3.658.53 - - [26/Oct/2011:21:06:30 -0500] "GET /buy HTTP/1.1" 200 0 "-" "Mozilla/5.0 (Windows; U; Windows NT 5.1; er
55.3.658.53 - - [26/Oct/2011:21:06:56 -0500] "GET /buy HTTP/1.1" 200 0 "/buy" "Mozilla/5.0 (Windows; U; Windows NT 5.1;
14.323.74.653 - - [26/Oct/2011:21:07:00 -0500] "GET /demo HTTP/1.1" 200 0 "-" "Jakarta Commons-HttpClient/3.0-rc4"
14.323.74.653 - - [26/Oct/2011:21:08:00 -0500] "GET /demo HTTP/1.1" 200 0 "/demo" "Jakarta Commons-HttpClient/3.0-rc4"
14.323.74.653 - - [26/Oct/2011:21:09:00 -0500] "GET /demo HTTP/1.1" 200 0 "/demo" "Jakarta Commons-HttpClient/3.0-rc4"
14.323.74.653 - - [26/Oct/2011:21:10:03 -0500] "GET /demo HTTP/1.1" 200 0 "/demo" "Jakarta Commons-HttpClient/3.0-rc4"
53.667.16.82 - - [26/Oct/2011:21:10:30 -0500] "GET /demo HTTP/1.1" 200 0 "-" "Jakarta Commons-HttpClient/3.0-rc4"
14.323.74.653 - - [26/Oct/2011:21:11:03 -0500] "GET /demo HTTP/1.1" 200 0 "/demo" "Jakarta Commons-HttpClient/3.0-rc4"
53.667.16.82 - - [26/Oct/2011:21:14:20 -0500] "GET /demo HTTP/1.1" 200 0 "/demo" "Jakarta Commons-HttpClient/3.0-rc4"
53.667.16.82 - - [26/Oct/2011:21:15:30 -0500] "GET /demo HTTP/1.1" 200 0 "/demo" "Jakarta Commons-HttpClient/3.0-rc4"
367.84.337.612 - - [26/Oct/2011:21:15:30 -0500] "GET /demo HTTP/1.1" 200 0 "-" "Jakarta Commons-HttpClient/3.0-rc4"
52.10.330.7 - - [26/Oct/2011:21:16:00 -0500] "GET /product/product2 HTTP/1.1" 200 0 "-" "Mozilla/5.0 (Windows; U; Windo
55.18.368.671 - - [26/Oct/2011:21:16:00 -0500] "GET /news HTTP/1.1" 200 0 "-" "Mozilla/5.0 (Windows; U; Windows NT 6.1;
52.10.330.7 - - [26/Oct/2011:21:16:03 -0500] "GET /product/product2 HTTP/1.1" 200 0 "/product/product2" "Mozilla/5.0 (W
55.18.368.671 - - [26/Oct/2011:21:16:05 -0500] "GET /news HTTP/1.1" 200 0 "/news" "Mozilla/5.0 (Windows; U; Windows NT
55.18.368.671 - - [26/Oct/2011:21:16:05 -0500] "GET /news HTTP/1.1" 200 0 "/news" "Mozilla/5.0 (Windows; U; Windows NT
55.18.368.671 - - [26/Oct/2011:21:16:05 -0500] "GET /news HTTP/1.1" 200 0 "/news" "Mozilla/5.0 (Windows; U; Windows NT
55.18.368.671 - - [26/Oct/2011:21:16:06 -0500] "GET /news HTTP/1.1" 200 0 "/news" "Mozilla/5.0 (Windows; U; Windows NT
655.633.64.678 - - [26/Oct/2011:21:16:30 -0500] "GET /product/product2 HTTP/1.1" 200 0 "-" "Mozilla/5.0 (Windows; U; Wi
52.10.330.7 - - [26/Oct/2011:21:16:31 -0500] "GET /buy HTTP/1.1" 200 0 "/product/product2" "Mozilla/5.0 (Windows; U; Wi
655.633.64.678 - - [26/Oct/2011:21:16:33 -0500] "GET /product/product2 HTTP/1.1" 200 0 "/product/product2" "Mozilla/5.
55.18.368.671 - - [26/Oct/2011:21:16:35 -0500] "GET /about HTTP/1.1" 200 0 "/news" "Mozilla/5.0 (Windows; U; Windows N
52.10.330.7 - - [26/Oct/2011:21:16:37 -0500] "GET /buy HTTP/1.1" 200 0 "/buy" "Mozilla/5.0 (Windows; U; Windows NT 5.1;
55.18.368.671 - - [26/Oct/2011:21:16:38 -0500] "GET /about HTTP/1.1" 200 0 "/about" "Mozilla/5.0 (Windows; U; Windows
655.633.64.678 - - [26/Oct/2011:21:17:21 -0500] "GET /product/product12 HTTP/1.1" 200 0 "/product/product2" "Mozilla/5.
655.633.64.678 - - [26/Oct/2011:21:17:23 -0500] "GET /product/product12 HTTP/1.1" 200 0 "/product/product12" "Mozilla/
655.633.64.678 - - [26/Oct/2011:21:17:37 -0500] "GET /product/product4 HTTP/1.1" 200 0 "/product/product12" "Mozilla/5.
655.633.64.678 - - [26/Oct/2011:21:17:39 -0500] "GET /product/product4 HTTP/1.1" 200 0 "/product/product4" "Mozilla/5.
655.633.64.678 - - [26/Oct/2011:21:18:41 -0500] "GET /product/product3 HTTP/1.1" 200 0 "/product/product4" "Mozilla/5.
655.633.64.678 - - [26/Oct/2011:21:18:43 -0500] "GET /product/product3 HTTP/1.1" 200 0 "/product/product3" "Mozilla/5.
655.633.64.678 - - [26/Oct/2011:21:18:57 -0500] "GET /product/product3 HTTP/1.1" 200 0 "/product/product3" "Mozilla/5.
655.633.64.678 - - [26/Oct/2011:21:18:59 -0500] "GET /product/product4 HTTP/1.1" 200 0 "/product/product3" "Mozilla/5.
655.633.64.678 - - [26/Oct/2011:21:19:00 -0500] "GET /product/product4 HTTP/1.1" 200 0 "/product/product4" "Mozilla/5.
367.84.337.612 - - [26/Oct/2011:21:20:30 -0500] "GET /demo HTTP/1.1" 200 0 "/demo" "Jakarta Commons-HttpClient/3.0-rc4"
13.640.53.680 - - [26/Oct/2011:21:25:00 -0500] "GET /demo HTTP/1.1" 200 0 "-" "Jakarta Commons-HttpClient/3.0-rc4"
```

Figure 1. Sample Web Log Data

If you are a Java Developer or a Data Scientist looking to have complete control over a Big Data domain, Hadoop opened the doors for a new class of analysis, giving the same parallel processing MPP power of larger proprietary systems, but at an entry level cost. These “Hadoop Power Users” (aka programmers) write code to parse through the detailed contents of such log files, ignoring the extraneous details to get to the relevant business components that can be used for detailed business analysis, using different programming languages like C++, Java, Python, PHP, Ruby, etc. and a growing array of Hadoop utilities. This interface enabled Hadoop programmers to gain huge level of power over these enormous computing environments, and the development platform became increasingly popular for large scale analytic processing. However, unless you were a programmer, one of the things that were missing was a simplified SQL model for querying the distributed data. There were utilities like Hive, which provided a basic SQL interface, though its interface was large rudimentary at best, with only a limited optimizer. For companies familiar with high end, high performance optimized SQL processing, or SAS users looking to leverage a simple query interface like SQL, Hive was not the solution for big data analytics.

ASTER SQL-MR – MAPREDUCE WITH A POWERFUL, SCALABLE SQL INTERFACE

Aster changes the game for SQL based Hadoop analytics.

In 2010, Aster Data introduced its new patent pending technology, SQL–Map Reduce, aka SQL-MR. It provides business users access semi-structured or multi-structured data in a simple, scalable, SQL fashion. Additionally, SQL-MR also provided an extensive framework, allowing users to incorporate Hadoop libraries into their SQL, extending the analytic depth of what is accessible via SQL.

For example, SQL-MR users could leverage functions libraries that would incorporate MPP Hadoop MapReduce models to drive for deeper insight. For example,

```
select token, sum(occurrences) as globalOccurrence
from map ( ON
select word, count(*) as occurrences
from WordOccurrences
group by word )
```

allowed for Map function to perform specialized word processing, in parallel on the Hadoop cluster, easily accessible from a SQL interface.

Other SQL-MR function soon started to come to life, including SQL-H, a specialized MapReduce function that enables high speed passes into Hadoop datasets via the Hadoop HCatalog. With SQL-H, Aster users could easily attach to Hadoop datasets, scattered over a large disk cluster, and treat them as just another SQL table, joining them easily with other database objects.

Probably the most popular Aster function available today is nPath, a specialized pattern recognition library that enables users to quickly traverse streams of data looking for patterns – and do so in the form of a SQL-MR instruction. For example, users can request a list of all traversals through a particular process path, that would help understand state changes in the overall process.

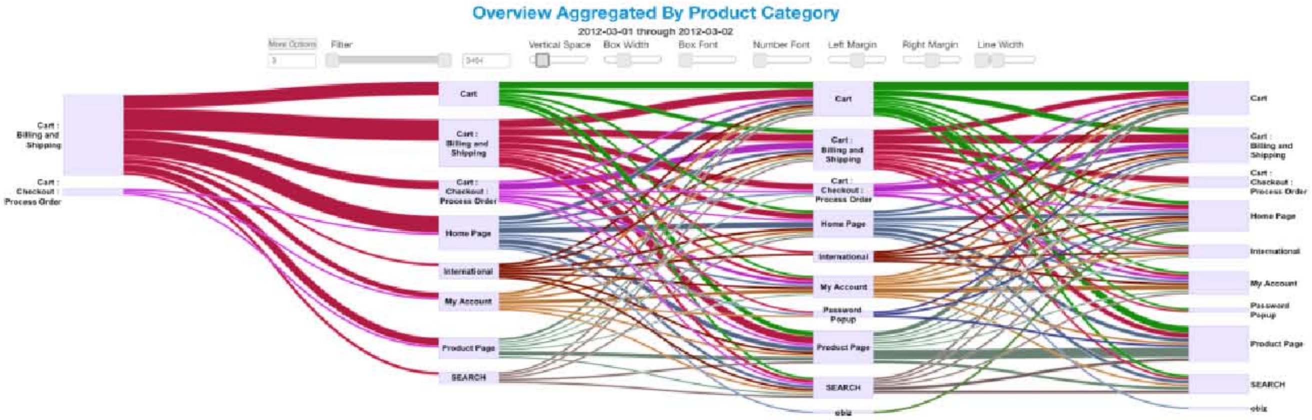


Figure 2. nPath pattern traversals for Web Sight Navigations

ASTER SQL-MR & SAS TOGETHER – THE BEST OF BOTH WORLDS

Aster and SAS together enables SAS oriented analytic users to access the full capabilities of the SAS environment, but leverage integrated capabilities to manage the underlying Aster Data.

Back in 2010, SAS too saw the promise of Aster, and was one of the first partners to develop new product specifically using the SQL-MR framework – SAS Scoring Accelerator for Aster. Like similar products for other DBMS platforms, SAS Scoring Accelerator enabled SAS users to run SAS Enterprise Miner scoring models on Aster, inside the database. Their implementation for Aster however, was unique, as it utilized the SQL-MR Framework to support this.

```
select * from
  sas_score(
    on mytable
    sas_code('hmeq.sas')
    format_xml('fmt.xml')
  );
```

In this case, SAS leveraged the SQL-MR function library to provide a Score function, reading as input SAS Scoring Code “HMEQ.SAS”, processed on a particular table.

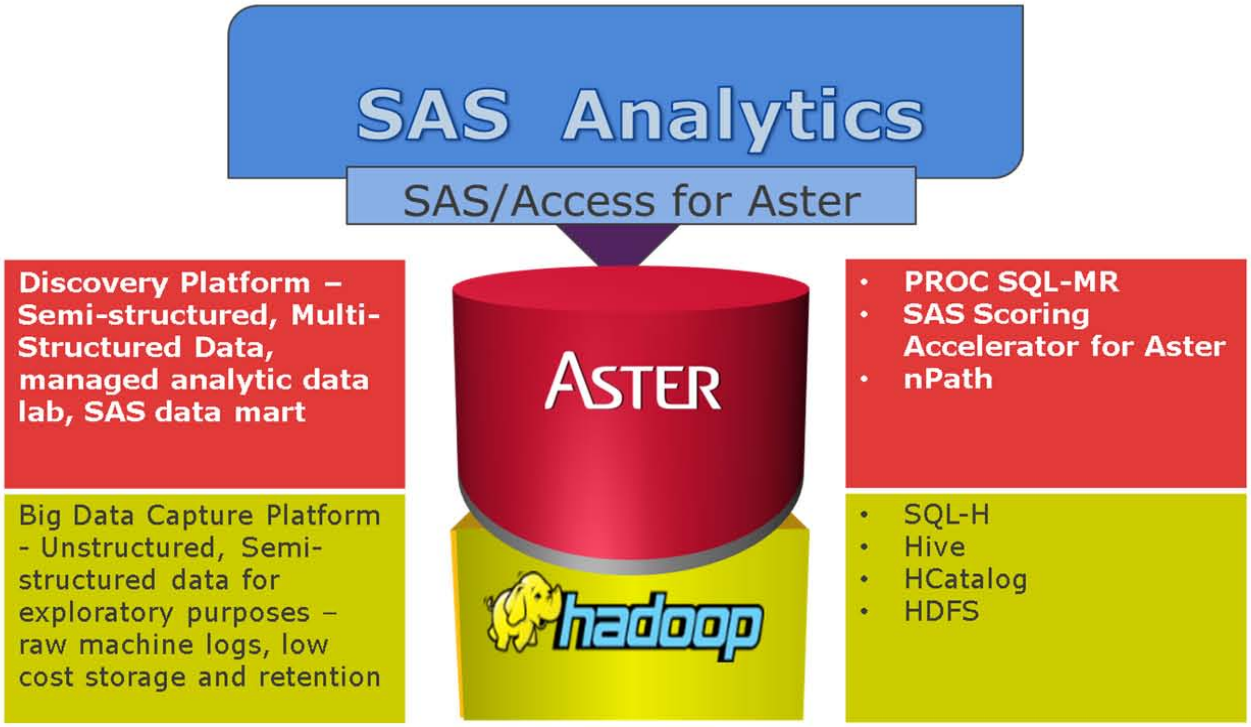

SAS Analytics
SAS/Access for Aster
Discovery Platform – Semi-structured, Multi-Structured Data, managed analytic data lab, SAS data mart
ASTER
• PROC SQL-MR
• SAS Scoring Accelerator for Aster
• nPath
Big Data Capture Platform - Unstructured, Semi-structured data for exploratory purposes – raw machine logs, low cost storage and retention
hadoop
• SQL-H
• Hive
• HCatalog
• HDFS

Figure 3. Aster Big Data Ecosystem with Hadoop

Further extending this model, the two primary interfaces for SAS users to take full advantage of the integrated Aster / Hadoop big data environment:

- SAS/Access for Aster nCluster provides the generalized interface to simplify data access for data managed in both Aster and in an underlying Hadoop cluster. Most SAS DBMS users are familiar with the LIBNAME reference that abstracts the tables within a database to look like SAS datasets. For Aster, SAS Access provides this level of integration, but also provides unique access to the SQL-MR query environment, via PROC SQL instructions. And, to simplify the basic analytic processes, SAS/Access also provides an automated mechanism to perform in-database analytic processes, primarily focused around demographic procedures, like PROC FREQ, RANK, and MEANs.

SAS Scoring Accelerator for Aster nCluster provides an integrated interface to publish scoring models from SAS, down into the Aster SQL-MR interface, to be executed across the cluster as a SQL-MR model.

ANALYTIC SCENARIO – USE SAS/ASTER TO ANALYZE DATA IN HADOOP

In this demo scenario, we will use SAS/Aster to pull raw data from Hadoop data file, join with pre-processed data in Aster database, then process with Aster SQL-MR and SAS Analytics functions.

The Environment –

- Hadoop cluster storing large volume of web data stored on system file
- Aster nCluster storing pre-processed weblog data in database table

The Process –

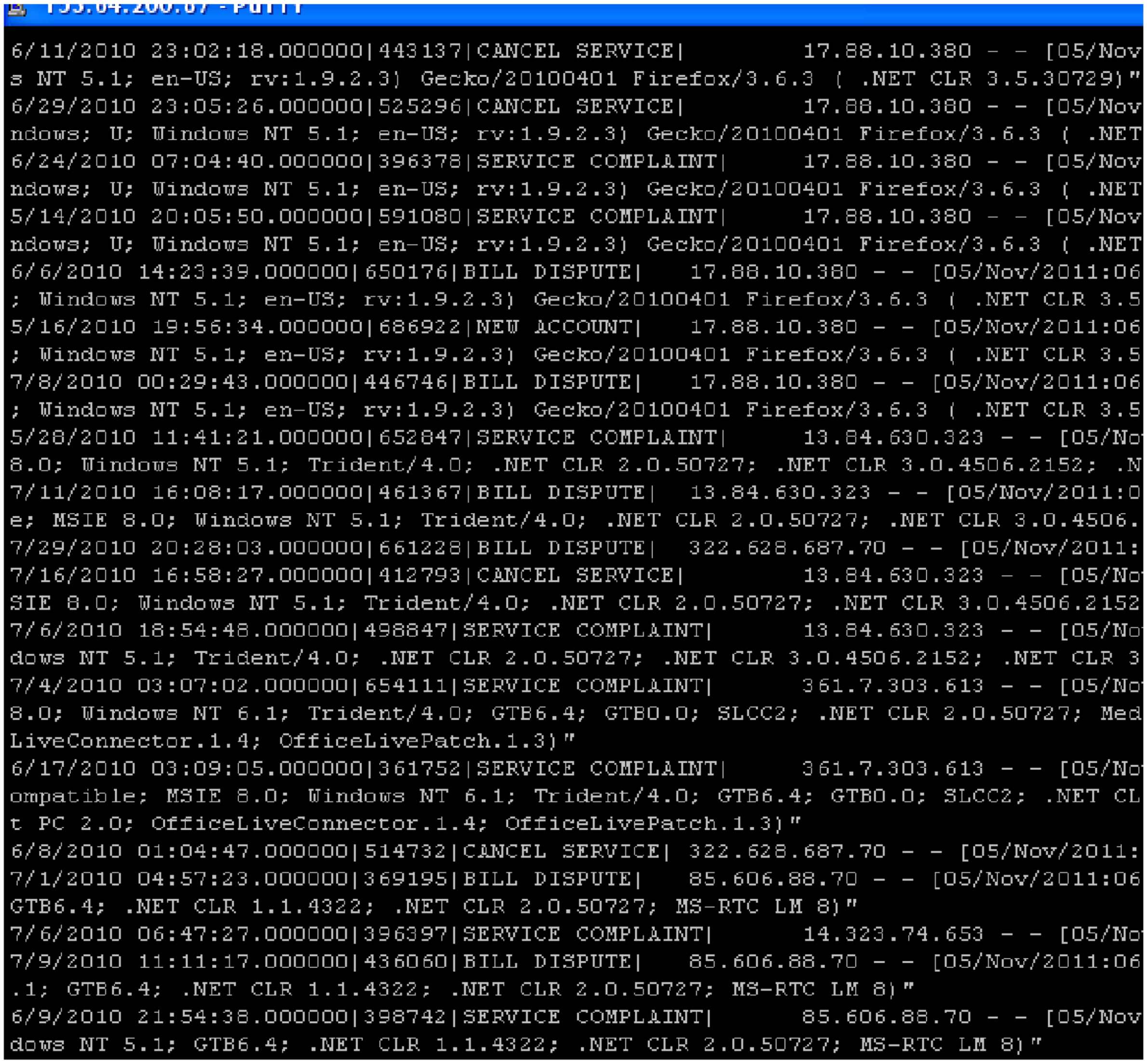
- Real working code examples all submitted from within SAS
- A 2-way join between Hadoop and Aster
- Execute Aster SQL-MapReduce functions (Sessionize, nPath)
- In_Database SAS Analytics on dataset in Aster

STEP 1

Use Aster SQL-H functionality to create an HCatalog table structure for the raw datafile on the Hadoop system

```
PROC SQL;
    connect to hadoop (
           server=“hadoop.asterdata.com“
           port=9083
           schema=default
           user=hdpuser
           password=hdpuser);

    CREATE TABLE hdp_data(
           datestamp string,
           customer_id int,
           web_click string,
           web_stream string)
    row format delimited fields terminated by ‘|’
    stored as TEXTFILE
    location ‘/apps/hive/warehouse/hdp_data’;
QUIT;
```



```
6/11/2010 23:02:18.000000|443137|CANCEL SERVICE|        17.88.10.380 - - [05/Nov
s NT 5.1; en-US; rv:1.9.2.3) Gecko/20100401 Firefox/3.6.3 ( .NET CLR 3.5.30729)"
6/29/2010 23:05:26.000000|525296|CANCEL SERVICE|        17.88.10.380 - - [05/Nov
ndows; U; Windows NT 5.1; en-US; rv:1.9.2.3) Gecko/20100401 Firefox/3.6.3 ( .NET
6/24/2010 07:04:40.000000|396378|SERVICE COMPLAINT|     17.88.10.380 - - [05/Nov
ndows; U; Windows NT 5.1; en-US; rv:1.9.2.3) Gecko/20100401 Firefox/3.6.3 ( .NET
5/14/2010 20:05:50.000000|591080|SERVICE COMPLAINT|     17.88.10.380 - - [05/Nov
ndows; U; Windows NT 5.1; en-US; rv:1.9.2.3) Gecko/20100401 Firefox/3.6.3 ( .NET
6/6/2010 14:23:39.000000|650176|BILL DISPUTE|   17.88.10.380 - - [05/Nov/2011:06
; Windows NT 5.1; en-US; rv:1.9.2.3) Gecko/20100401 Firefox/3.6.3 ( .NET CLR 3.5
5/16/2010 19:56:34.000000|686922|NEW ACCOUNT|   17.88.10.380 - - [05/Nov/2011:06
; Windows NT 5.1; en-US; rv:1.9.2.3) Gecko/20100401 Firefox/3.6.3 ( .NET CLR 3.5
7/8/2010 00:29:43.000000|446746|BILL DISPUTE|   17.88.10.380 - - [05/Nov/2011:06
; Windows NT 5.1; en-US; rv:1.9.2.3) Gecko/20100401 Firefox/3.6.3 ( .NET CLR 3.5
5/28/2010 11:41:21.000000|652847|SERVICE COMPLAINT|     13.84.630.323 - - [05/No
8.0; Windows NT 5.1; Trident/4.0; .NET CLR 2.0.50727; .NET CLR 3.0.4506.2152; .N
7/11/2010 16:08:17.000000|461367|BILL DISPUTE|  13.84.630.323 - - [05/Nov/2011:0
e; MSIE 8.0; Windows NT 5.1; Trident/4.0; .NET CLR 2.0.50727; .NET CLR 3.0.4506.
7/29/2010 20:28:03.000000|661228|BILL DISPUTE|  322.628.687.70 - - [05/Nov/2011:
7/16/2010 16:58:27.000000|412793|CANCEL SERVICE|        13.84.630.323 - - [05/No
SIE 8.0; Windows NT 5.1; Trident/4.0; .NET CLR 2.0.50727; .NET CLR 3.0.4506.2152
7/6/2010 18:54:48.000000|498847|SERVICE COMPLAINT|      13.84.630.323 - - [05/No
dows NT 5.1; Trident/4.0; .NET CLR 2.0.50727; .NET CLR 3.0.4506.2152; .NET CLR 3
7/4/2010 03:07:02.000000|654111|SERVICE COMPLAINT|      361.7.303.613 - - [05/No
8.0; Windows NT 6.1; Trident/4.0; GTB6.4; GTB0.0; SLCC2; .NET CLR 2.0.50727; Med
LiveConnector.1.4; OfficeLivePatch.1.3)"
6/17/2010 03:09:05.000000|361752|SERVICE COMPLAINT|     361.7.303.613 - - [05/No
ompatible; MSIE 8.0; Windows NT 6.1; Trident/4.0; GTB6.4; GTB0.0; SLCC2; .NET CL
t PC 2.0; OfficeLiveConnector.1.4; OfficeLivePatch.1.3)"
6/8/2010 01:04:47.000000|514732|CANCEL SERVICE| 322.628.687.70 - - [05/Nov/2011:
7/1/2010 04:57:23.000000|369195|BILL DISPUTE|   85.606.88.70 - - [05/Nov/2011:06
GTB6.4; .NET CLR 1.1.4322; .NET CLR 2.0.50727; MS-RTC LM 8)"
7/6/2010 06:47:27.000000|396397|SERVICE COMPLAINT|      14.323.74.653 - - [05/No
7/9/2010 11:11:17.000000|436060|BILL DISPUTE|   85.606.88.70 - - [05/Nov/2011:06
.1; GTB6.4; .NET CLR 1.1.4322; .NET CLR 2.0.50727; MS-RTC LM 8)"
6/9/2010 21:54:38.000000|398742|SERVICE COMPLAINT|      85.606.88.70 - - [05/Nov
dows NT 5.1; GTB6.4; .NET CLR 1.1.4322; .NET CLR 2.0.50727; MS-RTC LM 8)"
```

Figure 4. Raw data file in Hadoop (/apps/hive/warehouse/hdp_data)

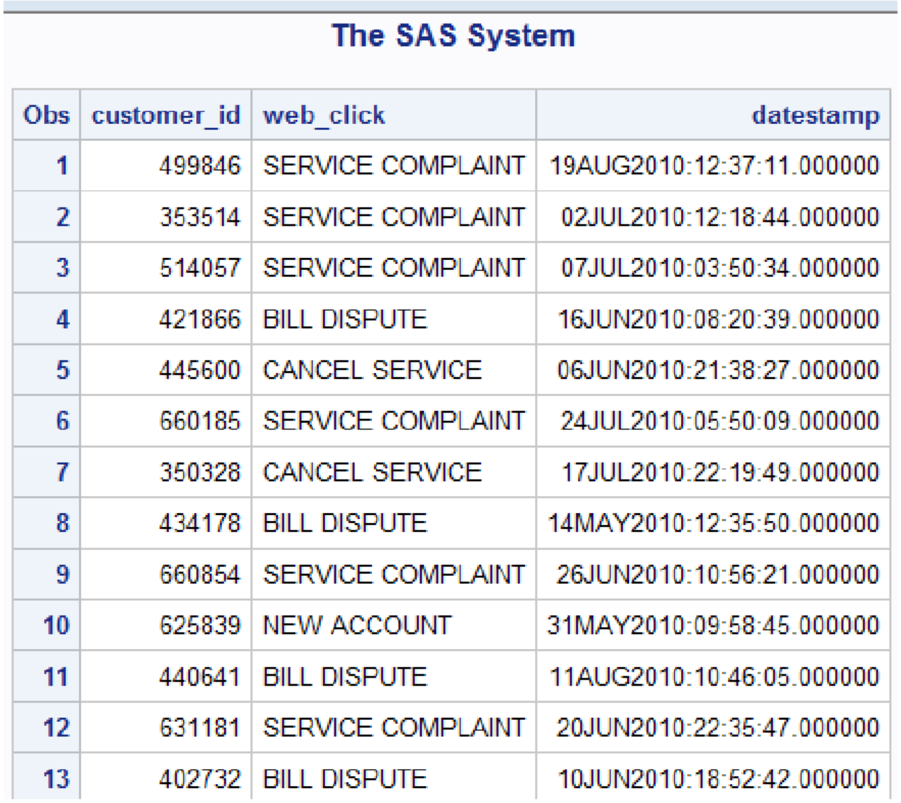
STEP 2

Use Aster SQL-MR load function to create view accessing the external data in Hadoop combined with data in Aster database table

```
PROC SQL;
    connect to aster (
           server=”aster.asterdata.com”
           port=2406
           schema=public
           user=beehive
           password=beehive);

    CREATE VIEW combined_data_view AS
           SELECT customer_id,
                  web_click :: character varying as action,
                  datestamp :: timestamp without time zone as datestamp
           FROM load_from_hcatalog (ON      mr_driver
                                      server ('aster.asterdata.com')
                                      port ('9083')
                                      dbname ('default')
                                      tablename ('hdp_data')
                                      username ('hdpuser')       );
           UNION ALL
           SELECT customer_id,
                  web_click,
                  datestamp
           FROM ast_data_table;

PROC PRINT DATA=comdined_data_view;
RUN;
```



The SAS System

Obs	customer_id	web_click	datestamp
1	499846	SERVICE COMPLAINT	19AUG2010:12:37:11.000000
2	353514	SERVICE COMPLAINT	02JUL2010:12:18:44.000000
3	514057	SERVICE COMPLAINT	07JUL2010:03:50:34.000000
4	421866	BILL DISPUTE	16JUN2010:08:20:39.000000
5	445600	CANCEL SERVICE	06JUN2010:21:38:27.000000
6	660185	SERVICE COMPLAINT	24JUL2010:05:50:09.000000
7	350328	CANCEL SERVICE	17JUL2010:22:19:49.000000
8	434178	BILL DISPUTE	14MAY2010:12:35:50.000000
9	660854	SERVICE COMPLAINT	26JUN2010:10:56:21.000000
10	625839	NEW ACCOUNT	31MAY2010:09:58:45.000000
11	440641	BILL DISPUTE	11AUG2010:10:46:05.000000
12	631181	SERVICE COMPLAINT	20JUN2010:22:35:47.000000
13	402732	BILL DISPUTE	10JUN2010:18:52:42.000000

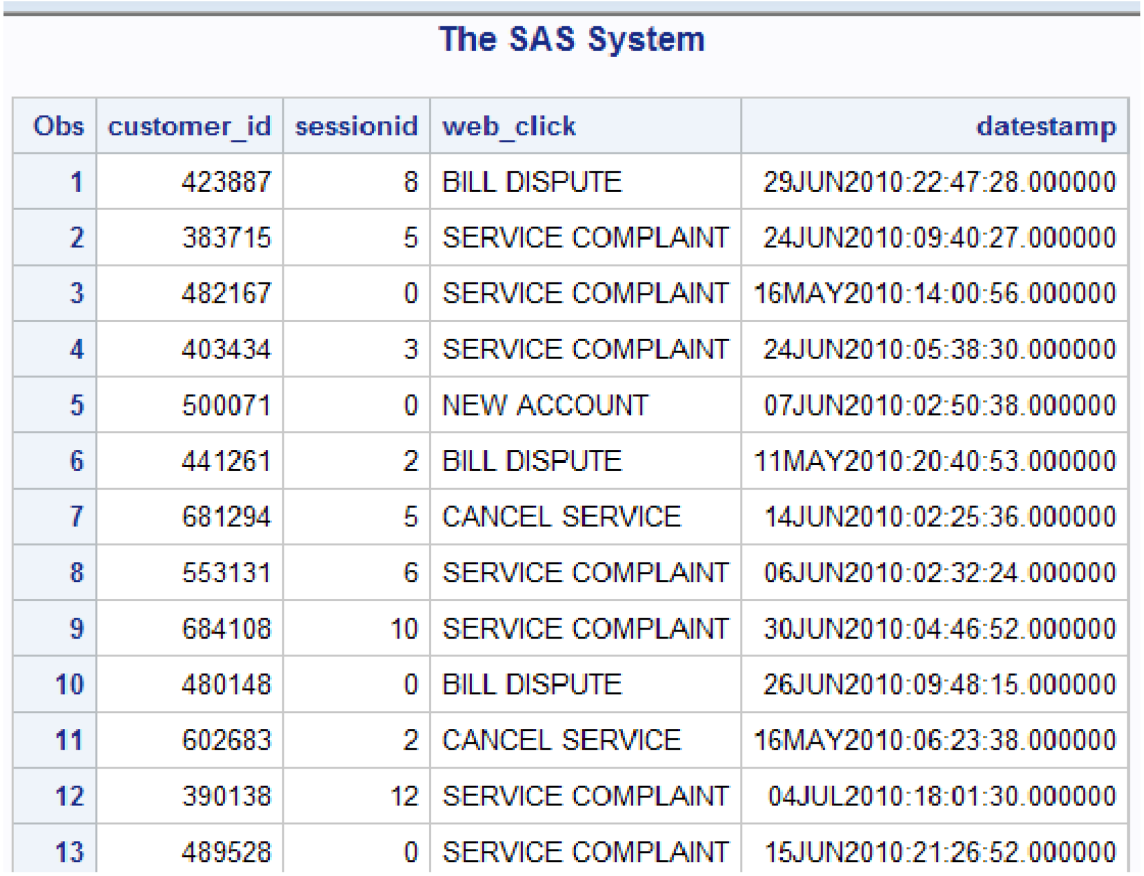
Figure 5. Sample listing of Aster view from SAS – combined Hadoop and Aster data

STEP 3

Use Aster SQL-MR Sessionize function to map each click in the combined data stream to a unique session identifier

```
        CREATE VIEW sessionize_data_view AS
                SELECT customer_id, sessionid, web_click, datestamp
                FROM Sessionize (
                        ON combined_data_view
                        PARTITION BY customer_id
                        ORDER BY datestamp
                        TIMECOLUMN ('datestamp')
                        TIMEOUT (60) );

   PROC PRINT DATA=sessionize_data_view;
   RUN;
```



The SAS System

Obs	customer_id	sessionid	web_click	datestamp
1	423887	8	BILL DISPUTE	29JUN2010:22:47:28.000000
2	383715	5	SERVICE COMPLAINT	24JUN2010:09:40:27.000000
3	482167	0	SERVICE COMPLAINT	16MAY2010:14:00:56.000000
4	403434	3	SERVICE COMPLAINT	24JUN2010:05:38:30.000000
5	500071	0	NEW ACCOUNT	07JUN2010:02:50:38.000000
6	441261	2	BILL DISPUTE	11MAY2010:20:40:53.000000
7	681294	5	CANCEL SERVICE	14JUN2010:02:25:36.000000
8	553131	6	SERVICE COMPLAINT	06JUN2010:02:32:24.000000
9	684108	10	SERVICE COMPLAINT	30JUN2010:04:46:52.000000
10	480148	0	BILL DISPUTE	26JUN2010:09:48:15.000000
11	602683	2	CANCEL SERVICE	16MAY2010:06:23:38.000000
12	390138	12	SERVICE COMPLAINT	04JUL2010:18:01:30.000000
13	489528	0	SERVICE COMPLAINT	15JUN2010:21:26:52.000000

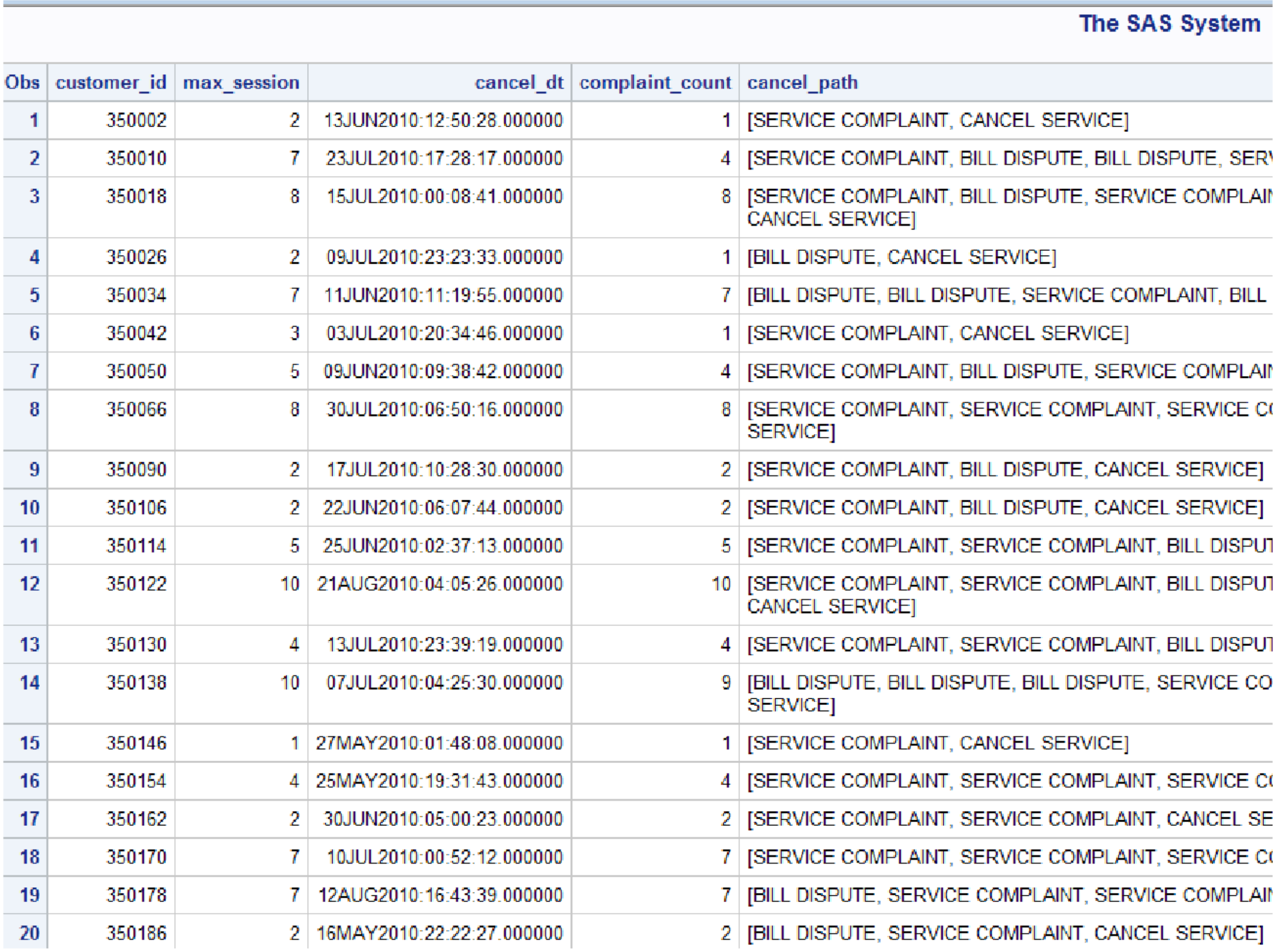
Figure 6. Sample listing of Aster view from SAS – sessionized data

STEP 4

Next, use Aster SQL-MR nPath function to show the complete progression of customers starting with complaints through cancellation – the combined sessionize data is analyzed in the nPath function

```
    CREATE TABLE path_to_cancel
    DISTRIBUTE BY HASH (cancel_path)
    AS
    SELECT
           customer_id,
           max_session,
           complaint_count,
           cancel_dt,
           cancel_path
    FROM nPath    (
    ON sessionized_data_view
    PARTITION BY customer_id
    ORDER BY datestamp
    MODE (NONOVERLAPPING)
    SYMBOLS (
    web_click in (‘BILL DISPUTE’, ‘SERVICE COMPLAINT’) AS COMPLAINT,
    web_click = ‘CANCEL SERVICE’ AS CANCEL
    )
    PATTERN ('COMPLAINT+.CANCEL')
    RESULT (
           FIRST (customer_id OF COMPLAINT) AS customer_id,
           MAX (sessionid OF ANY (COMPLAINT, CANCEL)) AS max_session,
           COUNT (web_click OF COMPLAINT) AS complaint_count,
           LAST (datestamp OF CANCEL) AS cancel_dt,
           ACCUMULATE (web_click OF ANY (COMPLAINT, CANCEL)) AS cancel_path
)
    ) n;

PROC PRINT DATA=path_to_cancel;
RUN;
```



The SAS System

Obs	customer_id	max_session	cancel_dt	complaint_count	cancel_path
1	350002	2	13JUN2010:12:50:28.000000	1	[SERVICE COMPLAINT, CANCEL SERVICE]
2	350010	7	23JUL2010:17:28:17.000000	4	[SERVICE COMPLAINT, BILL DISPUTE, BILL DISPUTE, SER
3	350018	8	15JUL2010:00:08:41.000000	8	[SERVICE COMPLAINT, BILL DISPUTE, SERVICE COMPLAI CANCEL SERVICE]
4	350026	2	09JUL2010:23:23:33.000000	1	[BILL DISPUTE, CANCEL SERVICE]
5	350034	7	11JUN2010:11:19:55.000000	7	[BILL DISPUTE, BILL DISPUTE, SERVICE COMPLAINT, BILL
6	350042	3	03JUL2010:20:34:46.000000	1	[SERVICE COMPLAINT, CANCEL SERVICE]
7	350050	5	09JUN2010:09:38:42.000000	4	[SERVICE COMPLAINT, BILL DISPUTE, SERVICE COMPLAI
8	350066	8	30JUL2010:06:50:16.000000	8	[SERVICE COMPLAINT, SERVICE COMPLAINT, SERVICE C SERVICE]
9	350090	2	17JUL2010:10:28:30.000000	2	[SERVICE COMPLAINT, BILL DISPUTE, CANCEL SERVICE]
10	350106	2	22JUN2010:06:07:44.000000	2	[SERVICE COMPLAINT, BILL DISPUTE, CANCEL SERVICE]
11	350114	5	25JUN2010:02:37:13.000000	5	[SERVICE COMPLAINT, SERVICE COMPLAINT, BILL DISPUT
12	350122	10	21AUG2010:04:05:26.000000	10	[SERVICE COMPLAINT, SERVICE COMPLAINT, BILL DISPUT CANCEL SERVICE]
13	350130	4	13JUL2010:23:39:19.000000	4	[SERVICE COMPLAINT, SERVICE COMPLAINT, BILL DISPUT
14	350138	10	07JUL2010:04:25:30.000000	9	[BILL DISPUTE, BILL DISPUTE, BILL DISPUTE, SERVICE CO SERVICE]
15	350146	1	27MAY2010:01:48:08.000000	1	[SERVICE COMPLAINT, CANCEL SERVICE]
16	350154	4	25MAY2010:19:31:43.000000	4	[SERVICE COMPLAINT, SERVICE COMPLAINT, SERVICE C
17	350162	2	30JUN2010:05:00:23.000000	2	[SERVICE COMPLAINT, SERVICE COMPLAINT, CANCEL SE
18	350170	7	10JUL2010:00:52:12.000000	7	[SERVICE COMPLAINT, SERVICE COMPLAINT, SERVICE C
19	350178	7	12AUG2010:16:43:39.000000	7	[BILL DISPUTE, SERVICE COMPLAINT, SERVICE COMPLAI
20	350186	2	16MAY2010:22:22:27.000000	2	[BILL DISPUTE, SERVICE COMPLAINT, CANCEL SERVICE]

Figure 7. Sample listing of Aster view from SAS – results from nPath

The nPath function allows you to perform regular pattern matching over a sequence of rows.

With it, you can find sequences of rows that match a specified pattern and then extract information from the matched PATTERNs using SYMBOLs that represent the matched rows in the pattern.

In this example, the sequence of rows is the customer transaction data - represented by the PARTITION BY customer_id clause in the nPath statement.

The SYMBOLs represent web_click column values 'BILL DISPUTE' or 'SERVICE COMPLAINTS' as COMPLAINT and 'CANCEL SERVICE' as CANCEL.

The PATTERN is defined as one or more COMPLAINTs followed by exactly one CANCEL (COMPLAINT+.CANCEL) and only rows matching this PATTERN are processed.

The RESULTs are the derived/aggregated output values for each matched PATTERN in the partition of customer transactions (In other words, nPath generates one output row per PATTERN match).

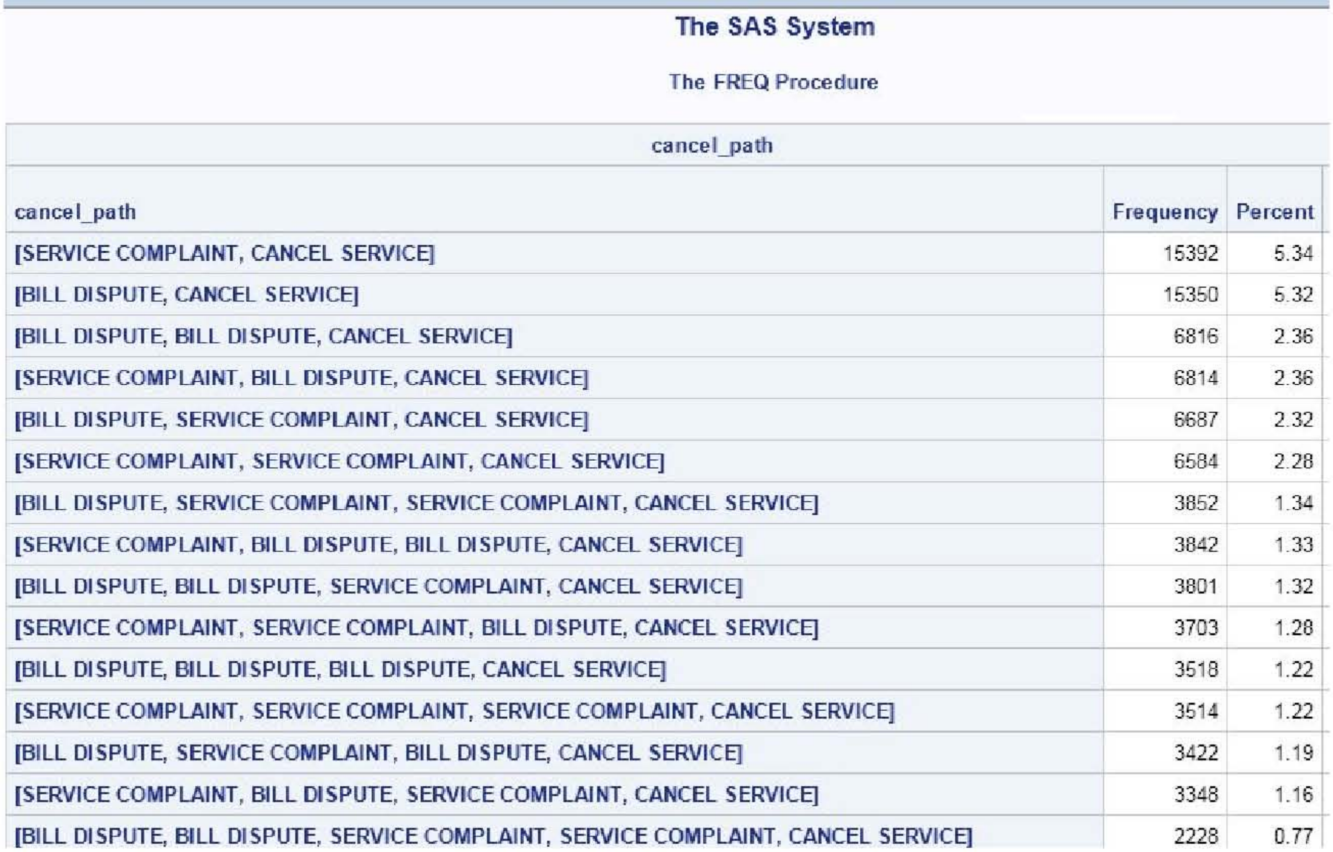
With this, we can see that nPath has walked all of the customer transactions, building a path for each customer that begins with one or more COMPLAINT request and ultimately ends in a CANCEL SERVICE request. From here, we can use this data to figure out which paths are more prominent, and act accordingly.

STEP 5

Finally with SAS Access for Aster, we can do further data analytics on the resulting data set from Aster using SAS In-Database Analytics function

- SAS Analytics PROC FREQ function to get top 10 cancellation paths

```
PROC FREQ
    DATA = path_to_cancel
    ORDER=FREQ;
    TABLES cancel_path ;
RUN;
```



The SAS System

The FREQ Procedure

cancel_path

cancel_path	Frequency	Percent
[SERVICE COMPLAINT, CANCEL SERVICE]	15392	5.34
[BILL DISPUTE, CANCEL SERVICE]	15350	5.32
[BILL DISPUTE, BILL DISPUTE, CANCEL SERVICE]	6816	2.36
[SERVICE COMPLAINT, BILL DISPUTE, CANCEL SERVICE]	6814	2.36
[BILL DISPUTE, SERVICE COMPLAINT, CANCEL SERVICE]	6687	2.32
[SERVICE COMPLAINT, SERVICE COMPLAINT, CANCEL SERVICE]	6584	2.28
[BILL DISPUTE, SERVICE COMPLAINT, SERVICE COMPLAINT, CANCEL SERVICE]	3852	1.34
[SERVICE COMPLAINT, BILL DISPUTE, BILL DISPUTE, CANCEL SERVICE]	3842	1.33
[BILL DISPUTE, BILL DISPUTE, SERVICE COMPLAINT, CANCEL SERVICE]	3801	1.32
[SERVICE COMPLAINT, SERVICE COMPLAINT, BILL DISPUTE, CANCEL SERVICE]	3703	1.28
[BILL DISPUTE, BILL DISPUTE, BILL DISPUTE, CANCEL SERVICE]	3518	1.22
[SERVICE COMPLAINT, SERVICE COMPLAINT, SERVICE COMPLAINT, CANCEL SERVICE]	3514	1.22
[BILL DISPUTE, SERVICE COMPLAINT, BILL DISPUTE, CANCEL SERVICE]	3422	1.19
[SERVICE COMPLAINT, BILL DISPUTE, SERVICE COMPLAINT, CANCEL SERVICE]	3348	1.16
[BILL DISPUTE, BILL DISPUTE, SERVICE COMPLAINT, SERVICE COMPLAINT, CANCEL SERVICE]	2228	0.77

Figure 8. PRC FREQ results on cancel path data

CONCLUSION

With Aster Data, SAS users get access to big data technology to deal with large scale analytic problems, on both the Aster DBMS, and Apache Hadoop. With this model, SAS users get a powerful platform to analytics on very large input log files and datasets. With intelligent use of SQL-MR instructions, SAS users can leverage the power of the SQL-MR engine to perform large scale analytic processing on Aster and Hadoop data, in database, in cluster.

They get the advantages of –

- Simplified SQL Query Model from PROC SQL

- Economical Hadoop storage model for large input data and log files
- Extensible Aster SQL-MR function model for advanced analytic functions, embeddable into SAS code

REFERENCES

- Franks, Bill – “Taming The Big Data Tidal Wave: Finding Opportunities in Huge Data Streams with Advanced Analytics”, 2012, Hoboken, NJ: John Wiley and Sons
- Teradata Corporation – “Teradata Aster Analytics Foundation Users Guide”, 2013, Miamisburg OH, Teradata Corporation

CONTACT INFORMATION

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

John Cunningham
Teradata Corporation
Danville, CA 94526

925 552 7124
john.cunningham@teradata.com

Momolue Kiadii
Teradata Corporation.
100 SAS Campus Drive
Cary, NC 27513
919-531-3215
momolue.kiadii@teradata.com