EVALUATION OF TERADATA ASTER

BY

Nancy Abramson - nabramson@ieee.org

Paul Balas - 303computing@gmail.com

EXECUTIVE SUMMARY

Today we have more data than we can analyze. Rapid automated data analysis will turn the data into insight in order to generate business, increase market share, and improve user experience. Data science combines the understanding of databases, statistics and business knowledge in order to better understand customer behaviors while keeping supporting systems efficient and healthy.

Here are some examples of how data can drive value:

1. Tactical optimization – measure to improve marketing campaigns and business processes
2. Predictive analytics – anticipate future demand to allocate resources
3. Prescriptive analytics – know the best changes to make improve customer experience
4. Recommendation engines – customize recommended show list
5. Automated decision engines – identify rogue set top boxes

Today's modern data science toolkits divide into three ends of the complexity spectrum:

1. Integrated Analytics Platforms with RDBMS back-ends
2. Integrated analytics platforms + hadoop back-ends
3. Open source platforms incorporating hadoop + flume/sqoop + hive/impala/sparq…

There is a tradeoff between a more expensive and integrated analytics platform vs. an open source environment composed of multiple tools. Both techniques have pros and cons. The direction your company takes will largely depend on two factors:

1. Do you already have a bias and investment towards a vendor offering these capabilities, or
2. Do you have a commitment to open-source solutions

This is often the leading factor in how your company will proceed, and if you've done your business case with demonstrable ROI and an expected business outcome, the software cost nor people expense should not be a major factor on the proprietary vs. open-source decision. If your ROI is too thin, you may have the wrong business case to justify the investment.

One of the tool sets available is Teradata Aster which is a traditional SQL database that uses Hadoop. This article reviews Teradata Aster in supporting data science and how well it can bring in a variety of datasets, manage the data once it comes in, model the data, and tell the data story. Also included is an end to end example of a business problem solved. The example problem to solution time was done within a week including an automated daily dashboard.

Teradata Aster does allow for rapid analysis of multiple datasets in a familiar database environment. It utilizes the power and economics of Hadoop with the enterprise ready database. Aster can be used reduce the time to production of a traditional data warehouse by bringing the analysis to the data instead of having to lift the data over the network. Rapid iterations can be done large data sets to find the data warehouse dimensions and facts that have the greatest business impact.

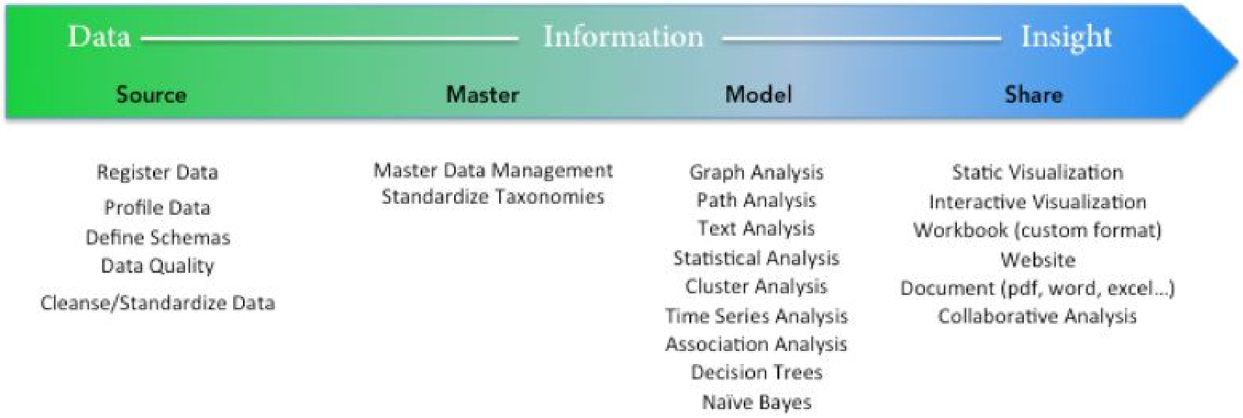


Figure 1 –Data Science Transformation Workflow

Data Steps to Insight

There are four steps in in the lifecycle of transforming data into insight as shown in Figure 1.

Step One – Identify and Understand Data Source

This is the process of getting data of interest into the system. Most data the data scientist is interested in analyzing is not in a pristine state ready for analysis. So part of the process for sourcing data is to understand it and to clean it up.

In order to get outside data into Aster, inserting data was benchmarked using traditional ETL tools and the in-place Aster loader. Aster tables using delimited files in-place

Aster's ncluster_loader command line tool screams at 100,000 rows/s. For the same data, the inserts average 3 rows/s using Aster's JDBC connection over the network within Talend and Pentaho ETL. Aster's ncluster_loader does not support updates or deletes. Much like Hadoop, the best way to complete updates and deletes with ncluster_loader is to create a new table with the new and updated rows. Updates and deletes maintain 3 rows/s within ETL tools using the Aster JDBC connector. Even factoring in the network overhead for the ETL versus the in-place inserts, the JDBC is very slow. JDBC has a limit of less than a million rows per day because of processing time. Although streaming of data is not supported within Teradata Aster for real-time analysis, the Hadoop platform does have Kafka which can be used for streaming.

Teradata Aster does not provide profiling tools nor stored procedures to support understanding the data. Teradata does include an additional application Revelytix Loom which can profile and manage the data lake. It has a gui interface that is independent of Aster. Teradata Loom is an enterprise – not open source – application but is free with Teradata's Hortonworks and Cloudera Hadoop extensions.

Some best practices include using consistent column names and data types. This supports integrating multiple data sources into common column names such as customer id. The structured environment that Aster produces is conducive to metadata management because it imposes a structure on the data with a schema.

Step Two – Master Data Management

Mastering data is all about lining up the data into standard attributes and hierarchies (taxonomies). It's a requirement to make sure that we don't count a thing (like a person) twice. For example, a person may be known by social security number in one data source and by a name and address in another data source. The system needs to make sure that both these records are related to a single individual. This goes for many other things in the world we identify uniquely, like states or electronic devices. Also, we also need to make sure that we lookup incoming data values and line them up to the taxonomies they belong to and at the appropriate level. For example, sometimes you might see a state abbreviated by CO or spelled out by its name – Colorado. Either of those values 'rolls-up' to the United States under a Geo-Political Taxonomy.

Master Data Management best practices need to be used with Aster Teradata because Aster does enforce data management beyond traditional relational database SQL commands. Teradata Aster is based on PostgreSQL which allows for an experienced SQL and relational developer to leverage those skills. The user need to have a solid understanding of the business in order to ask targeted business questions for understanding and blending data.

Aster is intended for rapid development so it intentionally does not have Master Data Management which is normally done when the discovery is more mature and moves into the data warehouse. Initial harmonization can be done using Aster PostgreSQL and advanced analytic tools which can be used to jump start the data warehouse master data management.

Step Three – Model Data for Better Understanding of the Business

This is the part of the process where the data scientist takes the data and uses various statistical and data mining techniques to derive understanding identify patterns, and make predictions. At this point, data needs to be in great shape or results will be misleading. By aggregating and showing the data patterns, the business can understand which levers they have to influence behaviors and understand what the future holds.

Aster has an easy to use set of over 50 data mining algorithms that uses the Hadoop mapreduce engine. By using mapreduce, it is powerful enough to analyze entire data sets without the need to sample. The syntax for the algorithms are a natural SQL-like extension within Aster as shown with the following sample syntax:

```
SELECT [ ALL | DISTINCT [ ON ( expression [, ...] ) ] ]
* | expression [ [ AS ] output_name ] [, ...]
FROM sqlmr_function_name
( on_clause
function_argument
) [ [ AS ] alias ]
[, ... ]
[ WHERE condition ]
```

This handful of SQL lines are in stark contrast to the hundreds of lines of Java code and library knowledge to support Map Reduce functions.

The Aster algorithms include:

- Unsupervised classification
    - k means clustering

    - k nearest neighbor outlier detection
    - Principal Component Analysis (PCA)
    - Vector Distance

- Supervised classification
    - Naïve Bayes
    - Generalized linear model
    - LDA
    - Random Forest
    - Support Vector Machines (SVM)

- Predictive analytics
    - Regression
    - Generalized Linear Mode

- Recommendation
    - Basket generator
    - Collaborative filtering

STEP FOUR – TELLING THE DATA STORY

Presenting a clear story that gets the point across for the analysis. Good insights should be obvious. Also, people often collaborate on an analysis. Providing a mechanism for people to share ideas is a critical component to a good system.

Aster brings the analysis to the data so instead of having to lift the data over the network to be analyzed somewhere else. Rapid iterations can be done large data sets which can produce results with multiple dimensions that reduce the time to value.

The display dashboard and SQL interface make it easy to connect to traditional reporting tools such as Tableau or embed the URL in a website for a standalone presentation. Best practices for connecting to Tableau include having a wide and short table at the highest grain possible for drilling down and pivoting the data.
Teradata has a presentation tool AppCenter that allows for advanced display of data that users can use for ad hoc analysis. AppCenter can also combine SQL together to be shared with other users. The does accommodate that Aster does not have stored procedures. AppCenter and Aster support a variety of path analysis graphics such as Sankey, chord, and sigma diagrams.

AppCenter also allows for operations to be simplified for agile business changes.

Business Problem Example

The example given uses a unique Teradata Aster algorithm to find patterns within two different log data sources. The business requested information on how many clicks from the website it took before an error occurred in the Apache Tomcat log. This is important to the business because common user click paths can be used to:

1. Automatically adjust the website to re-route customers away from the errors
2. Target code blocks for programmers to debug
3. Replicate test cases for testing once fixes are in place

Website event data is semi-structured because the events are recorded in logs much like the frames in a video real. Each row of data is a snapshot of what is happening at the moment with all the rows needed to understand the complete user experience with server information. The process for collapsing the timestamped rows by sessionizing and path analysis is used to aggregate the sessions into comparative examples for deeper analysis.

Step One – Identify and Understand Data Source - Web Paths

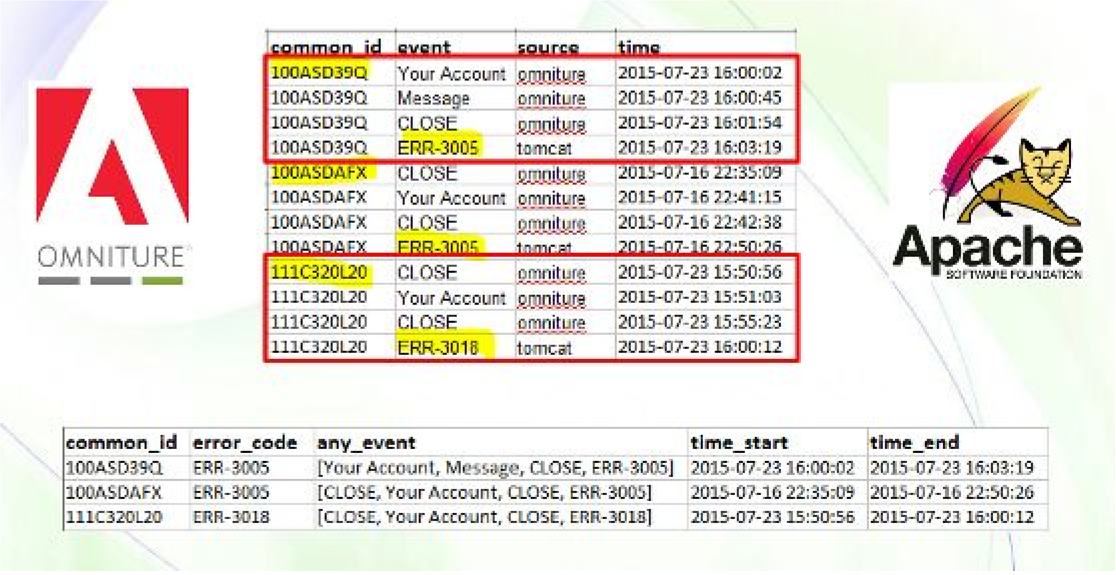
The identification of the data source bounds what any potential data solution can encompass. Ultimately the data solution is only as good as the data source which requires in depth understanding of the business and the quality and reliability of the delivered data. The data for this case is timestamped with additional codes that reveal each web experience from beginning to end for an account.

Step Two – Master Data Management

CLOSE-OK and CLOSE values are transformed to just CLOSE values because they have the same user function but two different machine codes. This allows for aggregating the values together generate a larger path count to create a stronger signal for the user experience.

Step Three – Model Data for Better Understanding of the Business

The data series needs to be understood in order to create a pattern for identifying the series that leads to an error and making sure that it is in a session that is in a reasonable amount of time. That is, the user did not leave and come back and had an error showing unrelated key click paths. Here is the sample data merged from the multiple data sources and aggregated with path analysis.



common_id	event	source	time
100ASD39Q	Your Account	omniture	2015-07-23 16:00:02
100ASD39Q	Message	omniture	2015-07-23 16:00:45
100ASD39Q	CLOSE	omniture	2015-07-23 16:01:54
100ASD39Q	ERR-3005	tomcat	2015-07-23 16:03:19
100ASDAFX	CLOSE	omniture	2015-07-16 22:35:09
100ASDAFX	Your Account	omniture	2015-07-16 22:41:15
100ASDAFX	CLOSE	omniture	2015-07-16 22:42:38
100ASDAFX	ERR-3005	tomcat	2015-07-16 22:50:26
111C320L20	CLOSE	omniture	2015-07-23 15:50:56
111C320L20	Your Account	omniture	2015-07-23 15:51:03
111C320L20	CLOSE	omniture	2015-07-23 15:55:23
111C320L20	ERR-3018	tomcat	2015-07-23 16:00:12

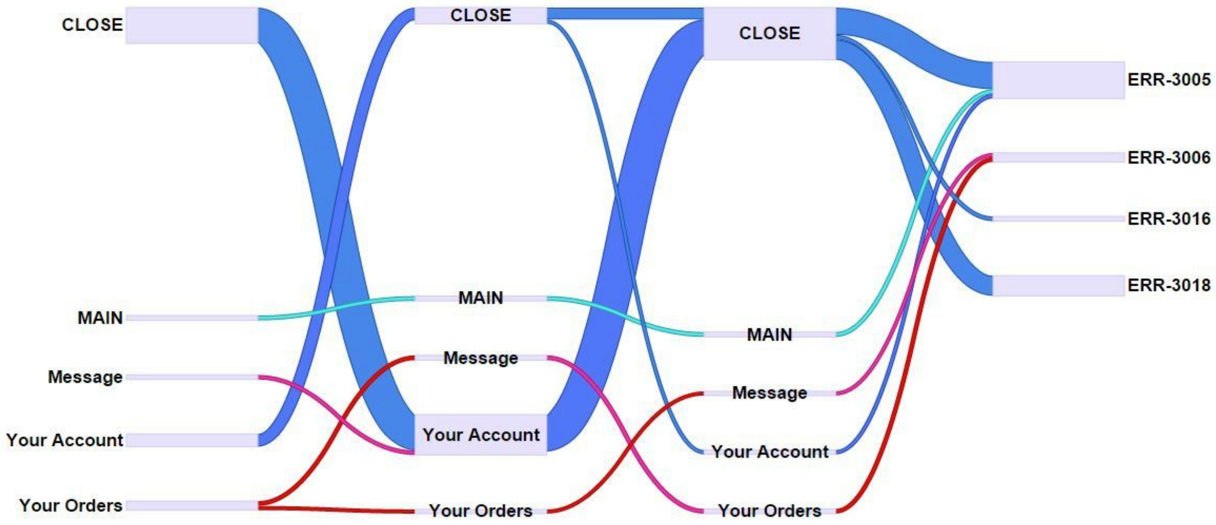
common_id	error_code	any_event	time_start	time_end
100ASD39Q	ERR-3005	[Your Account, Message, CLOSE, ERR-3005]	2015-07-23 16:00:02	2015-07-23 16:03:19
100ASDAFX	ERR-3005	[CLOSE, Your Account, CLOSE, ERR-3005]	2015-07-16 22:35:09	2015-07-16 22:50:26
111C320L20	ERR-3018	[CLOSE, Your Account, CLOSE, ERR-3018]	2015-07-23 15:50:56	2015-07-23 16:00:12

Step Four – Telling the Data Story

The Aster AppCenter allows for automation and graphically presenting results.

Sankey diagrams are typically used to visualize energy or material or cost transfers between processes. They can also visualize the energy accounts or material flow accounts on a regional or national level.

Sankey diagrams put a visual emphasis on the major transfers or flows within a system. They are helpful in locating dominant contributions to an overall flow. Often, Sankey diagrams show conserved quantities within defined system boundaries.



Conclusion

Aster is moderately easy to learn on the technical continuum with point and click BI on demand being the easiest and Java Map Reduce being the most complex. Teradata Aster is based on PostgreSQL with SQL like Map Reduce extensions. It is easy to use for someone familiar with ANSI SQL. The user does need to know how to clean the and how to analyze the data without prompts. It does not have a GUI interface like other tools such as WEKA for supporting profiling and algorithm decisions. As a result, one data scientist with a solid understanding of the business can run an average ad-hoc process end to end in a week.

Other solutions typically require multiple skills/people to accomplish the same outcome with a higher degree of system complexity. So Teradata Aster’s real value proposition is simplicity in answering high-value and complex business questions.