Big Data Benchmarking – Data Model Proposal

Ahmad Ghazal and Alain Crolotte
Teradata Corporation
ahmad.ghazal@teradata.com , alain.crolotte@teradata.com

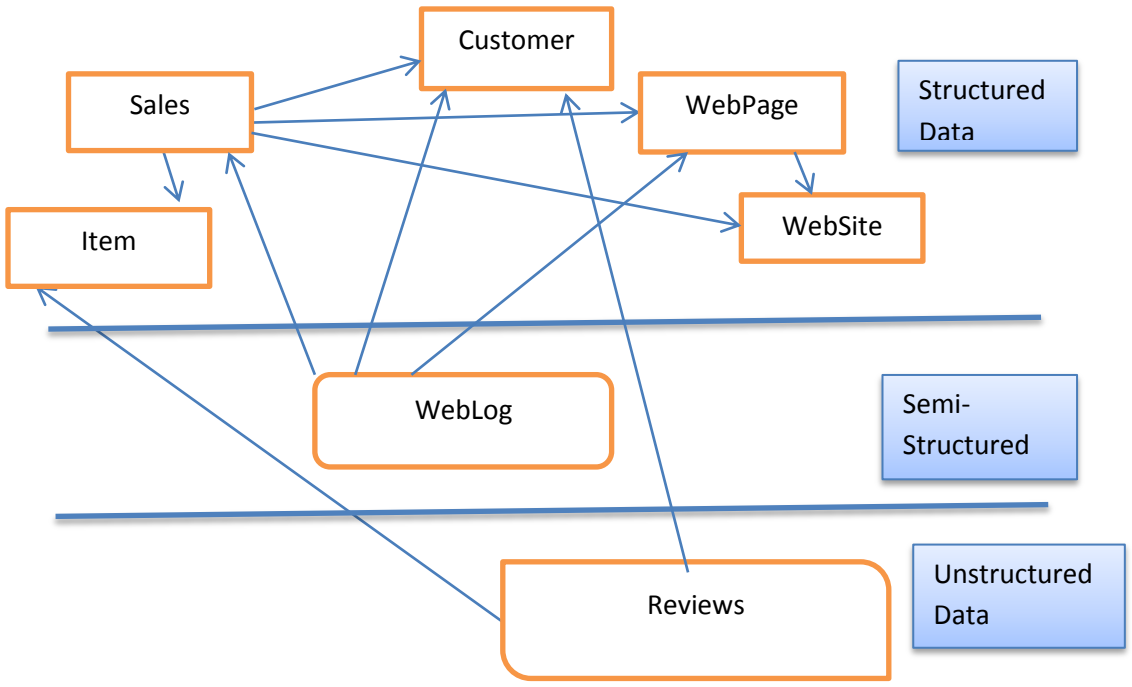
Summary –Big Data originates from what is called structured, unstructured and semi-structured data sources. Structured data are typically relational data from transaction processing or decision support systems. Text, video and audio are common examples of unstructured data. Semi-structured data is a form of structured data that does not conform to relational data models. Examples of semi-structured data are XML, weblogs and sensor information. We believe that a Big Data benchmark should encompass these three different sources of data and link these data together and allow queries mining these data separately but also together.

While we believe that the data generation for these data should be ensure that the same data is produce on all platforms, we do not believe that a particular way to answer the queries should be imposed as new techniques do not use a common language such as SQL. We also believe that a single business subject should be utilized so that a common thread can be used for all three types of data.

To make our Big Data Benchmark requirements more concrete we are proposing a data model based on the retail business model originally defined by the TPC-DS model. The retail model in TPC-DS is well-understood and rich in customer data. We also think the TPC-DS is more suited for Big Data benchmarks than other TPC models like TPC-H. The reason is that star schema is a common model for the structured part of analytics and it is more natural and easier to extend it to include the semi-structured and unstructured data.

The remainder of this document shows he details of our proposed data model. Also, we briefly discuss few examples of potential queries for the benchmark to illustrate our data model.

Data Model – The following figure portrays our proposed data model. In the sequel we go through the three types of data and show how benchmark requirements can be defined in specific terms.



The structured part of it is adopted from the TPC-DS model as mentioned in in the summary above. For semi-structured data we plan on using Web logs. The data structure is based on user clicks on the retail company website that captures who is making the clicks and pages visited in these clicks. As shown in the above figure, the clicks data is related to Sales, Customer and WebPage tables since it captures users who made the clicks, what pages are visited and if the clicks led to actual orders. The model also captures clicks of guest users who are not registered with the retail website.

As unstructured data which typically involves free text we chose the area of customer reviews. This are is very important area in commerce social media as it has been demonstrated that other people opinions are a key factor in making purchase decisions even when these people are strangers. The content of this data will be reviews on specific items contained in the structured area and other items as well.

Workload - Generating data with different scaling for the structured data could be based on the *TPC-DS specification.* The amount of Web logs generated will be based on the number of Customers, Sales and WebPage from the structured data area. The data will be generated according to fixed proportions but the distribution will be skewed. The skew will be controlled through the use of a small amount categories reflecting usage and purchase habits. Some of the data will have missing customer and/or Sales information to reflect guest users and clicks that did not lead to sales will also be present and in fixed proportion. The review data generation is similar to the Weblogs. The distribution is driven by the Item and Customer tables. Guest users can also place reviews on Items.

Note that all the proposed data generation above for all the three sources does not require a specific file system for implementing the benchmark. Relational tables and text files used in HDFS (Hadoop distributed file system) are just examples that can capture such data.

Queries -This proposal main focus is on the data model. However, we include description of few analytic queries to illustrate the richness and adequacy of the data model. The list below contains sample queries that cover different combinations of our proposed model various sources.

1. Purchase patterns.
   a) Find baskets of category purchases that happen in same transaction or multiple transactions for the same customer. Parameters here could be product category, date range of transactions and customer id.
   b) Find category purchases that happen after purchase of a specific category. For example, find out what customers purchase after a TV purchase.
   c) Find timing of related purchases. For example, find how long it took customers who bought a computer to purchase a printer.
2. Click analysis.
   a) Sessionizing analysis. Group clicks into sessions based on same user with different clicks for a specific period of time.
   b) Most frequent pages visited and users with most visits. This could be for a date range or by hour, day or week.

c) Most popular paths (sequence of clicks). This could be done for all users or by user. Also, can be done for clicks that lead to an actual sale transaction.

3. Product reviews.
   a) Find non-guest users who did the most reviews.
   b) Find products that got the most reviews.
   c) Find most frequently mentioned keywords in reviews.
   d) Sentiment analysis polarity which refers to the degree of positive or negative intention or mood in a reviews. This could be done by user and/or product.
   e) Combinations of the above. For example, find polarity of prolific reviewers. This could be used to detect bogus or paid reviewers.

In the foreseeable future, we do not anticipate query language to be a requirement for Big Data benchmark proposals. However, SQL (including the use of UDF), Hadoop MapReduce and its eco systems like HIVE and Pig are potential languages/interfaces for queries in this space.