Value of Data Transformation

Sean Kandel, Co-Founder and CTO of Trifacta
Organizations today generate and collect an unprecedented volume and variety of data. At the same time, the adoption of data-driven decision-making enhances the urgency for rapid, agile analysis of these data sets. As technologies for storing, processing and visualizing data have improved dramatically in the last decade, data cleaning and manipulation have increasingly become the bottleneck in gaining insight and ultimately improving business processes.

The Need For Data Transformation

According to Gartner, 64% of large enterprises plan to implement a Big Data project in 2014, but 85% of the Fortune 500 will be unsuccessful in doing so. Such projects fail because of challenges in both human resources and the analysis workflow.

OVERBURDENED, UNDER-UTILIZED ANALYTIC TEAMS
Finding and hiring the right mix of statistics, programming, and domain expertise is difficult. But once those talented people are on board, their skills are often wasted as they become bogged down in low-level data cleansing tasks or cannot quickly access the data they need.

THE BURDEN ON IT TEAMS
In many organizations, a business unit within IT such as the data warehousing group is responsible for managing and disseminating data resources. As the volume and variety of data increases, IT teams struggle to meet the growing demand from business units. The adoption of new technology stacks and practices that work with unstructured data compound the problem. For instance, IDC predicts the market for Hadoop software will increase by an order of magnitude by 2018.

UNPRODUCTIVE DATA ANALYSTS
Skilled data analysts and data scientists are a scarce and valuable resource. The McKinsey Global Institute predicts that demand for deep analytical talent could be 50-60% greater than its supply by 2018 and that there will be a shortage of approximately 1.5 million data analysts and data-savvy managers. Yet many analysts spend an inordinate amount of time manipulating data—searching for relevant data, diagnosing data quality issues and reformatting data before performing meaningful analysis—instead of exercising their specialty. Some estimate that such data cleansing and prep tasks constitute 50-80% of the development time and cost in data warehousing and analytics projects.

BUSINESS ANALYSTS SHUT OUT
Business analysts within a number of business units — including marketing, sales, and product — are adopting BI tools. However, these tools often assume that source data arrives in a particular format. Because data formatting tasks today rely on low-level programming, many business users rely on IT staff or data analysts to access the data they need. Unaligned priorities and communication overhead result in costly turnaround times and impede progress on downstream analytic tasks. The analysts with the deepest domain and business expertise are needlessly shut out from performing analysis.

Unusable Data
Inaccessible data and poor data quality stymie data analysis and prevent organizations from fully taking advantage of the data they collect.
POOR DATA UTILIZATION
Through activity and system logs, third-party APIs and vendors, and other publicly available data, organizations have access to an increasingly large and diverse set of data sources. However, the prohibitive cost of data transformation leaves much of this data dormant in “data lakes.” In many cases, analysts do not even attempt to test certain hypotheses or address potential questions. Instead of performing ad hoc analytic tasks, analysts are limited to working with static reports or fixed subsets and views of data.

POOR DATA QUALITY
Data sets regularly contain missing, extreme, duplicate or erroneous values that can undermine the results of analysis. These anomalies come from various sources, including human data entry error, inconsistencies between integrated data sets, and sensor interference. Flawed analyses due to dirty data are estimated to cost billions of dollars each year. Discovering and correcting data quality issues can also be costly.

Data Transformation Defined
In 2012, we conducted interviews with 35 data analysts from 25 organizations across a variety of sectors, including healthcare, retail, marketing and finance. Below we summarize some of the more common and challenging Data Transformation tasks based on this study and more recent conversations with hundreds of data analysts.

DISCOVERY
Finding relevant data distributed across multiple databases, database tables or files is often a significant bottleneck. Organizations often lack sufficient documentation or search capabilities to enable efficient identification of desired data. Instead, analysts relied on their colleagues: they often asked database administrators or others for help. In some cases, the administrator who set up the data may have already left the company.

The difficulties in discovering data are compounded by the difficulty of interpreting certain fields in a dataset. Many analysts describe situations in which fields are coded and require lookups against external tables. Foreign key definitions help identify the appropriate table to perform lookups, but these definitions were often missing in relational databases and non-existent in other types of data stores. Even without coding, missing units or metadata created ambiguity. For instance, one analyst noted that many date-time fields were stored without time zone information. The analysts had to reconstruct time zones from corresponding geographic information.

In many cases, schema drift leads to redundant columns. One company we interviewed had a database table containing four columns containing job titles for its users. These columns evolved over time, were often conflicting and there was no documentation describing which column was up-to-date or appropriate to use.

Verifying Assumptions
Analysts make assumptions during analysis that inform the types of transformations they use, how they sample data and which models are appropriate. Common assumptions included “how values were distributed within an attribute” (was an attribute normally distributed?), “what values were unique” (were there duplicates?), and “how different attributes related to each other” (was X always greater than Y?). Other assumptions required domain expertise to verify.

Detecting Anomalies
Data sets may contain a number of quality issues that affect the validity of results, such as missing, erroneous or extreme values. Many analysts report issues dealing with missing data. In some cases, observations contained missing or null attributes.

Analysts reported using a number of methods for imputing missing values. One organization even had an intern build a dedicated interface for displaying and correcting missing...
values across data sources. In other cases, entire observations were missing from a data set. Missing observations were much more difficult to detect.

**STRUCTURING**
Once an analyst discovers appropriate data to use, she often needs to manipulate the acquired data before she could use it for downstream analysis. Such data wrangling, munging, or cleaning involves parsing text files, manipulating data layout and integrating multiple data sources. This process, whether managed by IT staff or by analysts, is often time-consuming and tedious.

**Ingesting Semi-Structured Data**
A number of analysts report issues processing semi-structured data. The most common example is ingesting log files. Parsing log files may require writing complex regular expressions to extract data into relevant fields. Interleaved log files containing multiple event types in a single file can further complicate parsing.

So-called “block data” is another common data format that was difficult to parse. In a block format, logical records of data are spread across multiple lines of a file. Typically one line (the “header”) contains metadata about the record, such as how many of the subsequent lines (the “payload”) belong to the record.

Data from third-party services or APIs often required a level of processing before analysis could begin. Email campaign providers, credit card processing companies, and other external services often delivered user reports in idiosyncratic formats.

Although many data sets arrived in these formats, most analysis tools do not support such semi-structured formats, preventing their use at early stages of data transformation.

**Advanced Aggregation and Filtering**
Many analysts note difficulty performing ad hoc grouping of observations, as in path or funnel analysis. One analyst at a web company investigated the sequence of actions users took before converting to a paying customer, upgrading their accounts, or canceling their accounts. The source data set was a log of user activities on the website, with each entry corresponding to a single activity by a single user. The analysts needed to group activities not only by user, but also by event time, where the time was conditional on other events in the log (i.e., prior to closest conversion). These types of queries in SQL often involve nested subqueries. Similar subqueries are necessary to write filters such as “delete all users who never upgraded their account.”

**CONTENT**
The ability to augment data sets with other data from multiple sources is crucial to many analytic workflows. However, doing so imposes a number of challenges.

**Missing Identifiers**
Identifiers useful for joining records across data sets were often missing in one or more data sets, inconsistent across data sources or incorrect for certain records. When identifiers were missing or incorrect, analysts derived new methods for linking records. Some analysts would match records using rules based on other fields that did not uniquely identify distinct records.

Analysts reported three types of inconsistency in identifiers during integration. First, identifiers used slight variations in spelling or formatting that make direct matches difficult. For instance, a patient’s first name might be stored as “John” in one record and “Jonathan” in another. Some analysts defined ad hoc rules (“fuzzy matches”) to detect similar items. The analysts then inspected the matches to verify that the records referred to the same entity.

**Inconsistent Encodings**
Second, data sources used two different encodings to represent the same identifier. For instance, a state might be identified by its full name (e.g., California) or by its as Federal Information Processing Standard (FIPS) code (e.g., 6). In this case, an analyst must find or construct a mapping between identifiers. In the third case, identifiers used inconsistent units of measurement or class definitions. Multiple analysts described attempts to consolidate their respective company’s industry codes with the North American Industry Classification System (NAICS). Others described difficulties integrating geographic data
with varied region definitions. Similarly, many data sets use overlapping conventions for financial quarters. The situation is complicated when sets of regions overlap and one standardization does not subsume the others.

These integration problems were made more difficult when the data was stored across multiple databases. In response, most analysts reported having to migrate all of the data sets into the same data processing framework. The lack of support for data integration also impedes the effective use of exploratory visualization tools. Because analysts were often unable to integrate external data to augment visualizations within a tool, they must resort to assembling data outside of the tool. One analyst noted that she spent most of her time integrating data together from disparate data sources to drive visualizations.

Benefits Of Transformation

IT teams, data analysts or scientists, and business analysts across business units can all benefit from Data Transformation. Agile Data Transformation greatly increases the probability of success in Big Data initiatives by increasing productivity, enabling a broader class of business users to work directly with data, and improving data utilization.

INCREASED PRODUCTIVITY

Effective Data Transformation systems make existing data workers more productive. IT Teams and analysts can build and manage data products and transformation scripts more effectively. Efficient Transformation search and recommendation amortizes the cost of data manipulation across teams of workers, allowing them to leverage existing work. By reducing the time to perform data cleaning and preparation, analysts can utilize more data in important analytic tasks such as feature construction, model construction and validation, exploratory visualization and content augmentation. IT teams can service more downstream customers while focusing on other core IT functions within the organization. Moreover, making Data Transformation more accessible to non-programmers enables these users to provide self-service solutions to their consumers: data and business analysts within the organization.

BRINGING BUSINESS ANALYSTS TO THE DATA

With high-level graphical Data Transformation products business users can access and manipulate data on-demand. Instead of waiting hours, days or weeks for the data they need, business analysts pull data from various formats and sources directly into downstream visualization and analytics tools. With Self-service Data Transformation, business analysts spend more time exercising their specialties: exploring hypotheses and making urgent business decisions.

IMPROVED DATA USABILITY

The success of analysis depends not only on the people who work with data, but also on the quantity and quality of data harnessed for modeling, prediction and visualization. Accelerating the time to clean and manipulate data allows domain experts to work with more data and to explore deeper relationships between data sets. Data Transformation products that couple automated routines for anomaly detection with appropriate visual summaries enable more rapid detection and response to potential data quality issues.

IMPROVED BUSINESS PROCESS

Ultimately, the goal of data analysis is not simply insight, but rather an improved business process. Successful analytics lead to product or operational changes that drive value for the organization. Such changes often necessitate new metrics and new workflows for capturing and analyzing the new process. Agile Data Transformation shortens the cycle in this process, creating more rapid adoption and analysis of evolving processes.