prepare a data set for later integration. This chapter discusses these issues and describes the data profiling processes.

Establishing Usability of Candidate Data Sources

No business intelligence (BI) and analytics program can be built without information. That information will originate in data sets coming from many different sources and providers. Few of these providers will have a stake in the success of the outcome of your BI program. It is therefore not surprising that there is an oft-quoted (although difficult to source!) statistic claiming that 70% of the effort associated with a data warehousing or data mining project is spent on data preparation. A large part of this effort involves trying to harmonize data sets from disparate sources into a single repository. To this end, those responsible for accessing, extracting, and preparing data for loading into an analytical platform must understand the particulars about the data in the source systems prior to its extraction and integration. This involves more than trusting the supplied documentation, data dictionaries, or printouts of data models.

Anyone who has gone through the drill of preparing data for analytic processing understands this. The data that we are given is seldom in a pristine form, with some data sets having do details at all. Different data sets may be more or less usable in their original states, and the value of that data may differ based on how much effort needs to be invested to ensure that the data can be consistently integrated into the data warehouse. When faced with an integration process incorporating disparate data sets of dubious quality, data profiling is the first step toward adding value to that data. Data profiling automates the initial processes of what we might call *inferred metadata resolution:* discovering what the data items really look like and providing a characterization of that data for the next steps of integration.

Preliminary data profiling prior to data integration provides a reasonable characterization of the metadata associated with a data set, which can help reduce the amount of effort required for integration. The information collected via profiling will help to automate the preparation of data for integration into a data warehouse, which then yields a significant reduction in the cost of building that data warehouse. In Jack Olson's book, *Data Quality, The Accuracy Dimension*, it is noted that a Standish Group report, ''performing data profiling at the beginning of a project can reduce the total project cost by 35%.'' If this is true, it suggests that for many data warehousing projects, the cost of a data profiling tool is dwarfed by the potential cost savings.

Data Profiling Activities

Data profiling is a hierarchical process that attempts to build an assessment of the metadata associated with a collection of data sets. The bottom level of the hierarchy

characterizes the values associated with individual attributes. At the next level, the assessment looks at relationships between multiple columns within a single table. At the highest level, the profile describes relationships that exist between data attributes across different tables.

The complexity of the computation for these assessments grows at each level of the hierarchy. The attribute-level analysis is the least burdensome, whereas cross-column analysis can actually be costly in terms of computational resources. This provides one aspect of evaluation of data profiling tools: performance.

Another important evaluation criterion is ease of use of the results. Because there is so much information that can be inferred from the data values that make up a data set, it is easy to get lost in reams of statistics and enumerations. Remember the goal of reducing the effort required to integrate data, and keep this in mind when reviewing a profile assessment.

One other item to keep in mind while profiling data is that the most significant value is derived from discovering business knowledge that has been embedded in the data itself. Old-fashioned (and currently frowned-upon) database administration (DBA) tricks, such as overloading data attributes (in lieu of adding new columns to production databases) with encoded information, carry embedded business knowledge that can be shaken out and lost during an automated data cleansing process. As an example, in one of my client's employee identifier fields, most of the values were all digits, but a large number of records had a value that was all digits except for the character "I" appended to the end of the number. Appearances of asterisks as the last character in a name field and combination codes in a single column (i.e., a single string that comprises three distinct coded values, such as "89-NY/USA") are other kinds of examples of business knowledge embedded in the data.

The existence of a reason for a rule shows one of the subtle differences between a data quality rule and a business rule. Declaring that all values assigned to a field in a database record must conform to a pattern gives us a way to validate data within a record but gives us no insight into why the value must be in that form. When we associate meaning with a data quality rule, we suddenly have a context that allows us to understand why values must conform to that format and what deviations from that format mean.

Data Model Inference

When presented with a set of data tables of questionable origin (or sometimes even with a pedigreed data set), a data consumer may want to verify or discover the data model that is embedded within that set. This is a hierarchical process that first focuses on exposing information about the individual columns within each table, then looks at any relationships that can be derived between columns within a single

columns exhibits a key relationship. The second approach is more of a semantic approach that evaluates column names to see if there is any implied relation. For example, two tables from the same data set may each have a column called ''party_addr_id,'' which might lead us to conjecture that these columns refer to the same object identifier.

Attribute Analysis

Attribute analysis is a process of looking at all the values populating a particular column as a way to characterize that set of values. Attribute analysis is the first step in profiling because it yields a significant amount of metadata relating to the data set. The result of any of these analyses provides greater insight into the business logic that is applied (either on purpose or as a by-product of some other constraints) to each column. The end product should be a list of questions about each column that can be used to determine data quality or validation constraints or even information from which some BI can be inferred.

Typically this evaluation revolves around the following aspects of a data set:

- **Range analysis**, which is used to determine if the values fit within a well-defined range;
- **Sparseness**, which evaluates the percentage of the elements populated;
- **Format evaluation**, which tries to resolve unrecognized data into defined formats;
- **Cardinality and uniqueness**, which analyzes the number of distinct values assigned to the attribute and indicates whether the values assigned to the attribute are unique;
- **Frequency distribution**, which shows the relative frequency of the assignment of distinct values;
- **Value absence**, which identifies the appearance and number of occurrences of null values;
- **Abstract type recognition**, which refines the semantic data type association with a specific attribute;
- **Overloading**, which attempts to determine if an attribute is being used for multiple purposes.

The frequency analysis also provides summarization/aggregate values that can be used to characterize the data set, including:

- **Minimum value**, based on the ordering properties of the data set;
- **Maximum value**, also based on the ordering properties of the data set;
- **Mean**, providing the average value (for numeric data);

- **Median,** providing the middle value (if this can be defined);
- **Standard deviation**, (relevant only for numeric values).

Profiling data attributes (or columns) also sheds light on details of descriptive characteristics, feeding these types of analyses:

- **Type determination,** which characterizes data type and size;
- **Abstract type recognition**, which refines the semantic data type association with a specific attribute, often depending on pattern analysis;
- **Overloading**, which attempts to determine if an attribute is being used for multiple purposes; and
- **Format evaluation**, which tries to resolve unrecognized data into defined formats.

Range Analysis

Relating a value set to a simple type already restricts the set of values that a column can take; most data types still allow for an infinite number of possible choices. During range analysis a set of values is tested to see if the values fall within a well-defined range. If so, depending on the data type, some inferences may be made. For example, if the data type is a date, the range may signify some time period for which the corresponding data is relevant. Another example might distinguish a small range of integer values that correspond to some enumerated encoding (i.e., a hook into some other reference data set).

More complex range analysis algorithms may be able to identify nonintersecting ranges within a data set as well. Consider an integer column that contains values between 0 and 9 as well as the value 99 as an error code condition. A naïve range analyzer might propose 0 through 99 as this attribute range, whereas the more sophisticated analyzer could bisect the values into two ranges, 0 through 9 and the single-valued range of 99. The more refined the distinct extant value ranges are, the easier it is for the business analyst or domain expert to recognize a meaning in those ranges, which then can be documented as attribute metadata.

Range analysis can be used in an intelligence application to explore minimum and maximum values of interest, perhaps related to customer activity and monthly sales or to prices customers are being charged for the same products at different retail locations. Another example might look for evidence of insurance fraud based on the existence of a wide range of practitioner charges for the same procedure.

Sparseness

The degree of sparseness may indicate some business meaning regarding the importance of that attribute. Depending on the value set, it probably means one of

Relationship Analysis

Cross-column or *relationship analysis* focuses on establishing relationships between sets of data. The goal of these processing stages is to identify relationships between value sets and known reference data, to identify dependencies between columns (either in the same table or across different tables), and to identify key relationships between columns across multiple tables.

Domain Analysis

Domain analysis covers two tasks: identifying data domains and identifying references to data domains. We have already discussed one virtual process of domain identification in the earlier section about Format Evaluation; discovered format specifications can represent data domains. Enumerated domains may be inferred and proposed from a value set when:

- The number of values is relatively small as compared to the context in which it is used (i.e., the number of possible values that an attribute might take is limited to a small set).
- The values are what we could call *intuitively distributed.* This means that the distribution, although not always even, will take on characteristics specific to the context. In some cases there is a relatively even distribution; in other cases there may be more weighting given to a small subset of those values.
- Other domains exist that may be derived from this domain.
- The domain is used in more than one table.
- The attribute that uses the value from the domain is rarely null.

Unfortunately, these are more guidelines than rules, because exceptions can be found for each characteristic. The brute-force method for identifying enumerated domains is to look at all possible value sets. We begin by presuming that each column in every table potentially draws its values from a defined domain. For every table, we walk through each column and select all the distinct values. This set is now a candidate domain, and we then apply heuristics to decide whether to call this set a domain. It turns out that sometimes we can make some kind of determination early in the analysis, and sometimes we have to wait until more knowledge has been gained.

Presuming that we have already started to build the domain inventory, we can see whether other data attributes make use of the same domain by analyzing how well the set of values used to populate one attribute matches the values of a known domain. The value-matching process for a specific attribute can be described using the following steps.

1. The attribute's distinct values are collected and counted.
2. The set of unique values is matched against each domain. Fast matching techniques are used for scalability.
3. For each domain, we compute three ratio values. The *agreement* is calculated as the ratio of distinct attribute values that are present in a domain to the total number of distinct values in the attribute. The *overlap* is calculated as the number of domain member values that do not appear in the attribute divided by the number of domain values. Last, we compute the *disagreement* as the number of values that appear in the attribute but are not members of the domain.
4. The domains are sorted by their agreement percentages. The highest agreement percentages are presented as likely identified domains.

When we compare an attributes value set to a known domain, there are four cases.

1. All the values used in the attribute are members of the known domain, and all the values in the domain are used in the attribute (agreement = 100%, overlap = 0%, disagreement = 0%). In this case, it is safe to say that the attribute takes its values from the known data domain.
2. All the values in the attribute are members of the known domain, but there are domain members that are not used in the attribute (agreement = 100%, overlap > 0%, disagreement = 0%). In this case, it is also likely that the attribute takes its values from the domain, but this may also indicate the attribute's use of a subdomain, which should be explored.
3. Some of the attribute values are members of the known domain, but some of the values used in the attribute are not members of the known domain (agreement < 100%, disagreement > 0%). In this case, there are two possibilities: (a) There is no real agreement between the attribute's values and the domain, in which case the search for a match should continue, and (b) The known domain may actually be a subdomain of a much larger set of values, which should be explored. The decision will probably depend on the percentages computed.
4. None of the values used in the attribute are taken from the known domain (agreement = 0%, overlap = 100%, disagreement = 100%). In this case it is probably safe to say that the attribute does not take its values from the domain.

Functional Dependency

A functional dependency between two columns, *X* and *Y,* means that for any two records *R1* and *R2* in the table, if field *X* of record *R1* contains value *x* and field *X* of record *R2* contains the same value *x,* then if field *Y* of record *R1* contains the value *y,* then field *Y* of record *R2* must contain the value *y.* We can say that attribute *Y* is determined by attribute *X.* Functional dependencies may exist between multiple

source columns. In other words, we can indicate that one set of attributes determines a target set of attributes.

A functional dependency establishes a relationship between two sets of attributes. If the relationship is causal (i.e., the dependent attribute's value is filled in as a function of the defining attributes), that is an interesting piece of business knowledge that can be added to the growing knowledge base. A simple example is a ''total_amount_charged'' field that is computed by multiplying the ''qty_ordered'' field by the ''price'' field.

If the relationship is not causal, then that piece of knowledge can be used to infer information about normalization of the data. If a pair of data attribute values is consistently bound together, then those two columns can be extracted from the targeted table and the instance pairs inserted uniquely into a new table and assigned a reference identifier. The dependent attribute pairs (that had been removed) can then be replaced by a reference to the newly created corresponding table entry.

Key Relationships

A table *key* is a set of attributes that can be used to uniquely identify any individual record within the table. For example, people databases might use a Social Security number as a key (although this is ill-advised, considering that many people do not have a Social Security number), because (presumably) no two people share the same one. If we have one table that contains a specified key field, other tables may be structured with references to the first table's key as a way of connecting pairs of records drawn from both tables. When one table's key is used as a reference to another table, that key is called a *foreign key.*

Modern relational databases enforce a constraint known as *referential integrity,* which states that if an attribute's value is used in table A as a foreign key to table B, then that key value must exist in one record in table B. There are two aspects to profiling key relationships: identifying that a key relationship exists, and identifying what are called *orphans* in a violated referential integrity situation.

A foreign key relationship exists between (table A, column x) and (table B, column y) if all the values in (table A, column x) overlap completely with the values in (table B, column y) and the values in (table B, column y) are unique. A data profiling application should be able to apply this assertion algorithmically to find foreign key relationships.

Orphans are foreign key values that do not appear in records in the targeted table. An example might be a reference in a catalog to a product that is no longer being made or sold by the company. The referential integrity constraint asserts that if the product is referenced in the catalog, it must exist in the active products database. If the data profiling tool is told that a foreign key relationship exists, it is simple to check for orphans. Even if the profiling tool has no prior knowledge about foreign

keys, it is possible to loosen the rules for identifying the foreign key relationship to find *near-foreign keys* where there are some values that would be orphans if the relationship did really exist. As in other cases, the tool can only propose these discoveries as rules, and it is up to the analyst to determine the value in the proposal.

Management Issues

When data profiling tools were first introduced, their relatively high cost was prohibitive for many implementations. However, at this point, data profiling is essentially a commodity; most data warehousing, data quality, or data integration vendors bundle profiling within their offerings, and there are a number of open source data profiling tools available as well. Some things to be aware of, though, include system performance and scope of the results.

First off, some of the algorithms used in data profiling are actually quite computationally intensive, and it is not unusual for some of the analysis to require both large amounts of computational resources (memory, disk space) and time to successfully complete. Second, because the computations are summaries of frequency analysis and counts, the results presented tend to be almost endless, with long lists of values, each of which may have appeared only once in a column. For small tables this is not really an issue. But if you start looking at large tables (greater than 1 million records, which today is really not unusual), the output can be more than overwhelming. The savvy manager needs to be aware that some expertise is required in absorbing the results of a data profiling application and know how best to use the application.

Summary

Data profiling adds significant value to the BI program when it can be used to effectively provide the archeological or forensic evidence of why specific data is the way it is. Data profiling is also useful in exposing business rules that are embedded in data, and it can help preserve information that may be scrubbed out during the data integration stages. In addition, profiling is actually directly useful in a number of BI applications, such as fraud detection, when the data analyst is familiar with the kinds of results to look for.

In this chapter we examine alternate information contexts—what kind of data is available, what its value is, and how to plan to integrate that information into your enterprise.

We will also discuss some of the more interesting issues and applications associated with the BI program: the use of demographics (i.e., nonqualitative descriptions, such as age and marital status) and psychographics (i.e., qualitative descriptions of lifestyle characteristics) for enhancement.

A lot of the bits of knowledge we can uncover through a BI application are not actionable unless we have some idea of what to do once we have discovered them. Making use of other data to tell us what to do with our knowledge not only provides insight into how to convert knowledge into dollars, it also continues to leverage our use of data in the first place. The kinds of data uses described in this chapter focus on the thought processes we can employ to help in not only answering questions, but helping in figuring out what the questions should be in the first place.

Customer Profiles and Customer Behavior

If the intent of BI and analytics is optimizing business opportunities, then most business scenarios involving revenue generation as well as managing the customer experience must reflect what can be learned from interactions with customers. There are at least three aspects to this consideration: who the customers are, how they behave, and how you can influence changes to customer behavior that benefit both you and the customer.

Customer Knowledge and Customer Profiles

The first aspect is ''customer knowledge,'' implying awareness of key customer characteristics that are relevant to your organization's business processes. Your analytical framework will most likely want to capture these characteristics in a ''customer profile'' to help in both analyzing the different archetypes of customers you have and how each of the archetypes interact with the business.

The phrase ''customer profile'' is used in similar contexts with different meanings. In one sense, a profile provides a general overview of the customer incorporating details about inherent characteristics (such as ''name,'' or ''birth date''), descriptive demographic characteristics (such as where the individual lives, whether the individual is married), preferences (such as the customer's favorite sports team), as well as analytical characteristics such as purchasing patterns or credit-worthiness.

A slightly adjusted view of a customer profile is mapped to your business and the way your business interacts with customers. In this view, there are classes of customer profiles into which each customer entity is grouped. The value of each

customer type is calculated in terms of specific variables relevant to creating value, such as the number of products purchased, or the frequency of store visits, or the variety of products bought.

Customer Behavior

Customer behavior is also a somewhat fuzzily defined term; for our purposes let's suggest that "customer behavior" models are intended to capture information about the actions a customer performs under specific circumstances.

As an example, let's say that a retail company emailed a special coupon for an in-store purchase of a particular item. There are a number of specific circumstances associated with this scenario: the presentation of the offer, the method of presentation, the time at which the offer was presented, the time the customer took an action, the timeframe associated with the offer, a particular retail location. Given this scenario, the retailer can track customer actions in relation to the circumstances—the customer ignored the offer, or took advantage of it at some specific time and location.

Developing Behavior Models

In essence, the initial objective of capturing and analyzing customer behavior is to develop models reflecting customer decision-making processes. In turn, these models are expected to help predict behaviors associated with the different customer archetypes. To continue the example, the company can link a tracking mechanism to the email campaign such as a bar code to be scanned at the point of sale. After the conclusion of the campaign, statistics can be collected about which customers responded. That data set can then be subjected to dimensional analysis based on the customer profile characteristics. This will allow some segmentation to suggest any correlation between selected variables and purchasing the product, such as showing predispositions like:

- Customers between the ages of 30 and 40
- Customers living within 2.5 miles of the retail location
- Customers with an income between $80,000 and $100,000 per year
- Customers who vacation in Florida during December

Remember, though, that correlation does not necessarily imply causation. Identifying potentially dependent variables may suggest a predisposition, but establishing the predictive nature of this suggestion requires additional research. The bottom line is that there is a need feeding the *insight* back into the process, and you should ask questions such as:

- Can the correlation be validated using additional campaigns?

One more use is examining processes within the organization that can be adjusted to help increase customer lifetime value. This can be done by focusing on optimizing the dependent variables, such as lowering the associated costs, increasing customer lifetime, or increasing the profitability.

Yet one big challenge for calculating and managing customer lifetime value is accumulating the right data needed for each of the dependent variables. As an example, let's look at one of these variables: duration of the customer lifetime. Calculating customer lifetime requires historical data detailing all customer transactions across all areas of the business. You cannot just depend on the dates of the sales transactions, especially when the sales cycle requires numerous steps each time the customer is engaged. At the same time, there may be other customer touch points that indicate engagement (such as calls to the call center), while monitoring of service usage could signal a reduction in use, signifying an imminent disengagement.

In each case, coming up with a definition for customer engagement and determining what data is necessary to confirm that a customer continues to be engaged requires both defined policies and directed data management effort. The same can be said about any of the other variables: what are the actual costs of acquisition? Does that include the costs of manufacturing the product, marketing, general administrative costs allocated to each customer? What are the ongoing servicing costs? Do those include specific service and maintenance activities for each customer, or do we also allocate part of the infrastructure charges (such as placing a new cell tower) to service costs?

Customer lifetime value is a very powerful concept that can help drive specific actions, both strategic and tactical. And the need to present the appropriate collections of data to help analysts formulate the right types of questions lends credibility with our recurring themes of clarifying business term definitions, understanding business user requirements, and ensuring quality and governance for the source data sets used for driving the data discovery and analysis.

Demographics, Psychographics, Geographics

Our customer profiles help the business know who the customers are, but as we have seen, there may be interest in knowing more than just customer names and locations. You might want to use those profiles to understand the kind of people your customers are—how old they are, what kinds of foods they like, what styles of car they drive, what their hobbies are, or how they like to have fun. Knowing this kind of information can enhance the way that products and services are marketed, especially when analyzing customer characteristics for market segmentation.

By analyzing demographic data (corresponding to the inherent characteristics) and the psychographic data (corresponding to habits, desires, preferences,

affinities) of those people who populate each market segment, the business analyst can try to formulate a profile of characteristics that model or represent each segment. Chapter 17 discusses how this segmentation is performed, as well as how new individuals are mapped into the previously defined segments. In this section we look at those characteristic details that can help describe people as a prelude to segmentation.

Demographics

Demographics represent a quantitative statistical representation of the *nonqualitative* characteristics of a human population. Demographics can include a person's age, marital status, gender, race, religion, ethnicity, income level, and such. Demographics incorporate that information about a person that is not necessarily a result of a lifestyle choice, but rather is more likely to be some attribution related to some external set of variables. For example, people do not choose their age; age is related to the difference between the date they were born and today's date. Similarly, people do not choose their race.

Typically, demographics are used to demonstrate the similarity between people. For example, a population is grouped by membership within certain characteristics (e.g., ages 18–34, 35–49, 50–75). The intent is examining the behavior of customers grouped into similar populations.

Psychographics

Psychographics refer to the quantitative representation of the *lifestyle* characteristics of a segmented population, mostly in terms of attitudes, interests, passions, and values. Whereas demographics measure those attributes that are not chosen, psychographics measure chosen attributes related to lifestyle, such as food and beverage preferences, the types of television programs someone watches, vacation location choices, chosen hobbies, leisure activities, and so on.

Psychographics are often used to differentiate people within a population. For example, psychographic information can be used to segment the population by component lifestyles based on individual behavior.

Developing the Customer Profiles

As discussed previously in this chapter, demographic and psychographic information is used in combination to enhance customer profiles. Demographic and psychographic information is frequently presented in a summarized, comparative form. An example is ''75% of males between the age of 18 and 34 have sampled at least four kinds of French wine.'' A statement like this relates a demographic population

TABLE 15.1 Example of US 2010 Census Housing and Occupancy Statistics —(*Continued*)

Subject	Number	Percent
American Indian and Alaska Native alone householder	1	0.0
Asian alone householder	0	0.0
Native Hawaiian and Other Pacific Islander alone householder	0	0.0
Some Other Race alone householder	3	0.1
Two or More Races householder	0	0.0
Subject	**Number**	**Percent**

additional demographic and psychographic data and performing these kinds of enhancements to specially attribute geographic regions to much greater detail. Fortunately, there are a number of companies that package and sell this kind of geographic detail. These data sets will provide not only detail but also essentially a reverse mapping between profile characterization and the places in which people live, and we will explore this in greater detail in the next chapter.

Behavior Analysis

The behavior analysis is a combination of the collection of actions or transactions, an analysis looking for patterns that have business relevance, and the use of the identified patterns for process improvement or for predictive purposes. We can illustrate this using an example of web-based transactions.

Practically every web-based transaction carries relevant information and practically every transaction is seen by a multitude of parties. Every web page you request is seen by your Internet service provider, other Internet service providers, content servers, numerous ad servers, numerous social networking sites, owners of the network infrastructure, as well as the owner of the web site you are requesting. There is a tremendous amount of data generated and generally accumulated as web statistics, page visit logs, time stamps, and so on.

Each of these parties can track the movements of visitors to various web sites. This provides an opportunity to analyze the correlation between content and individuals (based on their IP addresses) and potentially use that behavior information themselves or package it for the purposes of developing marketing strategies through the combination of traditional demographic and psychographic information with

online behavioral data. This information can be used in the creation of rich customer profiles, and the mining of these profiles for useful behavioral patterns and the application of the knowledge inherent in those patterns can help solve numerous business problems. Particularly exciting is the potential to convert Web visitors from browsers to purchasers. Profiling customers in the context of an e-business intelligence strategy can assist in providing micro-segmentation for targeting value-added products and services for cross-sell and up-sell opportunities.

Tracking Web Activities

Tracking user behavior involves more than just collecting server log files. Instead of relying on the traditional server log data, we can incorporate a more meaningful characterization of user activity. First, it is necessary to specify the kinds of actions that a user may perform while browsing at your site. This is more than just page views; rather, we want to superimpose business meaning on top of appropriate page views and to ignore meaningless ones. Behavior modeling then becomes a process of analyzing the sequence of actions that users perform, within what context those actions are performed, and whether any particular behaviors can be generalized for later predictive purposes.

Although each e-business's list of user actions may vary, here is a short list of some user actions that are interesting to log:

- Content request, asking for a specific page to be served
- Content impression, when a Web page containing specific content is served
- Content read, when served content is read
- Hyperlink click-through
- Advertisement impression, when an advertisement is served
- Advertisement click-through
- Social network click-through (such as a ''like'' button)
- Social network login
- Social network posting
- Social network repost (such as the ''retweet'' feature on Twitter)
- Comment posting (such as a comment on an interactive page)
- Comment read (such as clicking on a ''more'' button to expose the full text of a comment)
- Syndicated posting (such as content pushed through an RSS feed or forwarded from a page to a social network site)
- Initial registration, when a user registers
- Subsequent registration, when a user reregisters
- User login
- User logout

- Password change
- Password request, when a user forgets a password
- Input of new profile information (any time a user voluntarily enters new profile information)
- Forced data input accepted (when a user is asked to input new information and that request is followed)
- Forced data input rejected (when a user is asked to input new information and does not follow through)
- Information query (the user searches for information)
- Select product for purchase, such as when using a shopping basket and a product is selected for purchase
- Purchase sequence initiated, when purchase information is requested
- Purchase sequence completed, when enough information has been collected to complete a purchase
- Purchase sequence aborted, when a user does not complete the purchase sequence (“abandoned cart”)

A specific data mart can be constructed to capture this kind of activity for later analysis.

Customer Behavior Patterns

Now, after having captured customer actions for a period of time, the information in the user activity data set will represent a collection of time series of the ways that all the web visitors act when they are visiting the site. We can analyze this time series data to look for particular user behavior patterns. If these patterns represent desired business activity, we have discovered actionable knowledge that can be used in influencing visitor behavior.

One analysis framework presumes a desired outcome, and it looks for patterns that lead to that outcome. For example, we might want to explore how viewing a specific content item correlates to making an online sale. Or we may want to see how well the placement of advertising affects click-through rates associated with browsing sequences. This is actionable knowledge, because it gives us information about how well our expectations are converting into good business practices.

A different analysis framework looks for behaviors that are not known a priori. For example, we might extract activity sequences that lead to a completed purchase and then look for patterns in those sequences. We might discover that a purchase-completed action takes place 25% of the time that a specific order sequence of content views take place. This becomes actionable knowledge, because it suggests different ways to configure the browsing experience to accelerate a customer’s purchase.