Pitney Bowes
Business Insight

Core Data Services: Basic Components for Establishing Business Value

WHITE PAPER:

DATA QUALITY

David Loshin

Core Data Services: Basic Components for Establishing Business Value

INTRODUCTION

EFFECTIVE USE OF BUSINESS INFORMATION EXCEEDS THE DEMANDS OF STRUCTURAL DATA VALIDATION OR VALUE DOMAIN MEMBERSHIP TESTS. AS MORE DECISIONS ARE DISTRIBUTED ACROSS THE ORGANIZATIONAL HIERARCHY AND AS MORE INFORMATION IS PROCESSED AND EXPOSED TO END-CONSUMERS, THERE IS A GROWING NEED FOR VISUAL REPRESENTATIONS OF MANY ASPECTS OF BUSINESS TRANSACTIONS, SUCH AS TYPE, TIME, DURATION, AND CRITICALLY, BUSINESS TRANSACTION LOCATIONS. ENHANCING THE VALUE AND UTILITY OF BUSINESS INFORMATION FOR BOTH OPERATIONAL AND ANALYTICAL APPLICATIONS REQUIRES A COMBINATION OF SOUND DATA MANAGEMENT PRACTICES, SUCH AS DATA GOVERNANCE, DATA QUALITY ASSURANCE, DATA ENHANCEMENT, AND INCREASINGLY, LOCATION INTELLIGENCE CAPABILITIES SUCH AS GEOCODING, MAPPING, AND ROUTING.

STANDARDIZING SOUND DATA MANAGEMENT PRACTICES WHEN DEVELOPING OR REENGINEERING THE ENTERPRISE ARCHITECTURE LOWERS RISK AND INCREASES CONSISTENCY; ORGANIZATIONAL TRUST IN THE DATA SIMPLIFIES APPLICATION DEVELOPMENT WHILE INCREASING DATA RELIABILITY. WHILE MUCH HAS BEEN SPECULATED REGARDING CONCEPTS SUCH AS THE INTEGRATION OF MASTER DATA AS AN OPPORTUNITY FOR IMPROVED DATA QUALITY, THIS PAPER EXPLORES HOW A CONSOLIDATED SET OF VALUE-ADDED DATA SERVICES THAT SPAN THE SCOPE OF DATA MANAGEMENT BEST PRACTICES CAN REDUCE OPERATIONAL COSTS, INCREASE EFFICIENCY, ALL WHILE IMPROVING THE QUALITY AND UTILITY OF ENTERPRISE DATA. A SINGLE DATA SERVICES LAYER WILL REDUCE THE COMPLEXITY OF THE DATA MANAGEMENT FRAMEWORK; STANDARDIZING THE WAY DATA TOOLS ARE USED AS WELL AS THE WAYS THAT RULES ARE APPLIED WILL ENSURE PREDICTABLE INFORMATION RELIABILITY AND VALUE.

AS MORE ORGANIZATIONS TRANSITION TO A SERVICES ORIENTED ARCHITECTURE, SEAMLESS DATA INTEGRATION BECOMES A FUNDAMENTAL BUILDING BLOCK FOR ENTERPRISE APPLICATIONS

Business Value Drivers and Information Reliability

Despite senior management agreement about general expectation for high value information, there are certain events that dictate an immediate requirement for improving the value and utility of data. Corporate mergers and acquisitions, changes in regulatory compliance requirements, increased customer attrition, noticeable increase in call center activity, system migrations, improvements to customer-facing business processes, introduction or CRM or ERP, or committing to MDM are all examples of business events that depend on organizational adjustments in data management practices in order to succeed.

Each of these events demonstrates a need for instituting solid business processes and the corresponding information management practices to improve the customer experience while simultaneously maintaining economies of scale. As opposed to blindly throwing technology, perhaps multiple times, at a presumed problem and hoping that something will stick, a more thoughtful approach seeks a return on the technology investment that is achieved when the necessary information management practices are supported with core data services implemented once but deployed multiple times in a consistent way.

This approach combines fundamental information management practices with those services that enhance the value and utility of business information, such as data quality assurance, data enhancement, and location intelligence. Enabling these core data services requires

- Shoring up the information production processes that support multiple data consumers,
- Standardizing the tools and techniques across the enterprise,
- Standardizing the definition and subsequent use of business rules, and
- Standardizing spatial analysis/location intelligence services associated with operations and transactions.

Envisioning and standardizing the catalog of value-adding data services will eliminate the situation where the same or competing products are purchased multiple times or where different groups are implementing the same techniques without being aware of each other's plans. However, this approach will ultimately solidify and reuse investment in data management tools to get multiple returns on investment.

Core Data Services

Although the criteria that characterize a capability as a "core" service are not completely cut and dried, they can be summarized in terms of how they support enterprise information management best practices. For example, one might expect that any services that enhance business process interoperability, reduce application development complexity, standardize techniques, unify the definition and application of business rules, or improve quality and utility of customer information might be considered as "core services."

But in order to guarantee a predictable observance of good information management practices, a reasonable list should include at least these data integration, data quality and location intelligence services:

- Data access and integration
- Data profiling
- Parsing
- Standardization
- Identity resolution
- Record Linkage and merging
- Data cleansing
- Data enhancement
- Geocoding

Core Data Services: Basic Components for Establishing Business Value

- Mapping & Spatial Analysis
- Routing
- Data inspection and monitoring

Data Access and Integration

The techniques employed in data warehouse development have become ubiquitous, especially as the need for system interoperability has grown. As more people want to reengineer their environments to allow sharing, or even consolidation of siloed data sets, there is a growing need for the underlying data integration services to provide the bridging ability to seamlessly access data from many different sources and deliver that data to numerous targets. The ability for remote servers to access data from their original sources supports the efficient implementation and deployment of data management best practices that can be implemented using the core data services described in this paper. Integration services rely on the ability to connect to commonly-used infrastructures and data environments coupled with the parsing and standardization services for transforming extracted data into formats that are suitable for data delivery.

In the modern enterprise, data sharing, federation, and synchronization is rapidly becoming the standard. As more organizations transition to a services oriented architecture, seamless data integration becomes a fundamental building block for enterprise applications.

Data Profiling

An objective data quality assessment is a process of analysis and discovery, and in turn this is based on an objective review of the data values that populate the targeted data set. A statistical quantitative analysis will help data analysts pinpoint potentially flawed data such as outliers, missing values, data elements that do not conform to defined patterns or value domains, or are inconsistent with respect to other data elements.

Data profiling is a set of algorithms for statistical analysis and assessment of the quality of data values within a data set, as well as exploring relationships that exist between data elements or across data sets. Data profiling provides the ability to identify data flaws as well as a means to communicate these instances with subject matter experts whose business knowledge can confirm the existences of data problems. A data profiling tool will scan all the values in a column and provide a frequency distribution of that column's values, suggesting the type and potential uses of each column. Cross-column analysis can expose embedded value dependencies, while inter-table analysis explores overlapping values sets to identify foreign key relationships between entities. Data profiling can help discover business rules embedded within data sets, which can be used for ongoing inspection and monitoring.

To paraphrase Lord Kelvin, one cannot improve that which you cannot measure. Data profiling provides a concrete mechanism to analyze, evaluate, quantify, and measure the quality of data across different dimensions such as completeness, validity, and reasonability. Data profiling is now recognized as the first step in most data migration, data integration, and master data management projects to understand what potential issues exist within the source data set and to help design transformations that will cleanse the data as it is moved to its target.

Parsing

In many instances, data values are expected to conform to predefined sizes and formats. However, slight variations or ambiguity in data value representation may confuse individuals as well as automated applications. For example, consider these different data values: {California, CA, Calif., US-CA, Cal, 06}. Some of these values use characters, another uses digits, and some use punctuation or special characters. For the most part, a human would read these and recognize that they all represent the same conceptual value: the state of California. Yet automating the process of determining

whether these values are accurate or to investigate whether duplicate records exist, the values must be parsed into their component segments and then transformed into a standard format.

Describing the different component and format patterns used to represent a data object (person's name, product description, etc.) support the core service of parsing, which is used to determine whether a value conforms to a recognizable pattern as part of the assessment, matching and cleansing process. Pattern-based parsing then enables the automatic recognition and subsequent standardization of meaningful value components, such as the area code and exchange in a telephone number, or the different parts of a person's name. This type of service incorporates, by necessity, the ability to define sets of patterns, lists of data values belonging to defined value domains, or regular expressions and grammars.

When the same information is represented multiple times in different ways it becomes difficult to determine that different records (either in the same data set or across different data sets) refer to the same entity. But even before more complex record linkage algorithms can be applied, it is necessary to parse the text is the critical data fields to identify and subsequently segregate and reorder the key tokens carrying identifying information. In concert with standardization, parsing will help reduce variance across multiple records and simplify matching and linking.

Standardization

Parsing uses defined patterns, regular expressions, or grammars managed within a rules engine along with table lookups to distinguish between valid and invalid data values. When patterns are recognized, other rules and actions can be triggered to transform the input data into a form that can be more effectively used, either to standardize the representation (presuming a valid representation) or to correct the values (should known errors be identified).

To continue our example, each of our values for the state of California {California, CA, Calif., US-CA, Cal, 06} can be standardized to the United States Postal Service standard 2-character abbreviation of CA.

The standardization service builds on the parsing service, which can be combined with a library of data domains to split data values into multiple components and rearrange the components into a normalized format. Standardization can also change full words to abbreviations, or abbreviations to full words, transform nicknames into a standard name form, translate across languages (e.g., Spanish to English), as well as correct common misspellings.

Standardization provides two different kinds of value. First, by mapping variant representations to well-defined patterns, the standardization process can help normalize different representations and effectively cleanse flawed data. Second, mapping a value to a standardized representation (even without specifically cleansing it) reduces the variance in value structure to improve record linkage for purposes of deduplication and master data consolidation. Standardizing nicknames (such as "Bob," "Rob," or "Bobby" to "Robert") will increase the possibility that two records representing the same entity will be linked together, and that significantly improves the unified resolution necessary for a successful data quality and master data management program.

Identity Resolution

Because operational systems have grown organically into a suite of enterprise applications, it is not unusual that multiple data instances in different systems will have different ways to refer to the same real-world entity. Alternately, the desire to consolidate and link data about the same business concepts with a high level of confidence might convince someone that a record might not already exist for a real-world entity when in fact it really does. Both of these problems ultimately represent the same core challenge: being able to compare identifying data within a pair of records to

Core Data Services: Basic Components for Establishing Business Value

determine similarity between that pair or to distinguish the entities represented in those records.

Both of these issues are addressed through a process called identity resolution, in which the degree of similarity between any two records is scored, most often based on weighted approximate matching between a set of attribute values between the two records. If the score is above a specific threshold, the two records are deemed to be a match, and are presented to the end client as most likely to represent the same entity. Identity resolution is used to recognize when only slight variations suggest that different records are connected and where values may be cleansed, or where enough differences between the data suggest that the two records truly represent distinct entities. Since comparing every record against every other record is computationally intensive, many data services use advanced algorithms for blocking records that are most likely to contain matches into smaller sets in order to reduce computation time.

Identity resolution is a critical component of most data quality, master data management, and business intelligence applications. The desire for a "360-degree view of the customer" or a comprehensive product catalog is predicated on the ability to find all records that carry information about each unique entity and resolve them into a unified view. Parsing and standardization are preliminary steps, but the key technique is identity resolution.

Linkage, Merging, and Consolidation of Duplicate Records

Identity resolution provides the foundation of a more sophisticated core data service: duplicate record analysis and elimination. Identifying similar records within the same data set probably means that the records are duplicated, and may be subjected to cleansing and/or elimination. Identifying similar records in different sets may indicate a link across the data sets, which helps facilitate merging or similar records for the purposes of data cleansing as well as supporting a master data management or customer data integration initiative.

Identity resolution leads into the merging and consolidation process to establish the single view of the customer. While some business applications may be able to proceed by looking at a set of records containing information, automating the merging process will select the best data values from those records and allow for the creation of a single "best copy" of the data that can support both operational processes as well as analytics to improve ways of doing business.

Data Cleansing

After determining that a data value does not conform to end-user expectations, the data analyst can take different approaches to remediation. One approach is to transform the data into a form that meets the level of business user acceptability. This data cleansing service builds on the parsing, standardization, and enhancement services, as well as identity resolution and record linkage. By parsing the values and triggering off of known error patterns, the data cleansing service can apply transformation rules to impute data values, correct names or addresses, eliminate extraneous and/or meaningless data, and even merge duplicate records.

While a best practice in data quality is to determine where the source of errors is and to eliminate the root causes of poor data quality, frequently the data quality team does not have the administrative control to effect changes to existing production applications. Therefore, an alternative that may be necessary to assure high quality data is to cleanse the data directly, and at least provide data that meets a level of suitability for all downstream data consumers, and the data cleansing service makes this possible.

Data Enhancement

A different approach to data value improvement involves enhancement, in which additional information is appended

DATA PROFILING IS NOW RECOGNIZED AS THE FIRST STEP IN MOST DATA MIGRATION, DATA INTEGRATION, AND MASTER DATA MANAGEMENT PROJECTS

to existing records. Data enhancement, which builds on parsing, standardization, and record linkage is a data improvement process that adds information from third-party data sets (such as name standardization, demographic data imports, psychographic data imports, and household list appends). The enhancement service is significantly simplified when the provider has partnered with data aggregators, whose data can be used both as a system of record against which data instances are matched, as well as a resource for appending. A typical data enhancement activity, address standardization and cleansing, relies on parsing, standardization, and the availability of third-party address information.

There are few industries that do not benefit from the use of additional data sources to append and enhance enterprise information. Whether it means address standardization and correction to reduce delivery costs and improve direct response rates, geodemographic appends that provide insight into customer profiling and segmentation, credit data used to assess and reduce risk, historical information (such as automobile service histories or medical events) to provide greater insight into diagnostic and preventative care, among others, data enhancement is a process that optimizes business decisions within operational and analytical contexts.

Geographic Location and Geocoding

Geocoding is a location intelligence process for translating a conceptual location to specific coordinates on a map of the earth's surface. A geocoded is a pair of coordinates for the latitude and longitude of the location, and the geographic location (or "geolocation") service supplies a "latlong" coordinate pair when provided with an address. When combined with the data enhancement service, values can be determined (geocoded) from a place name, street address, postal code, or from the intersection of two streets. This service also works in reverse if you provide it with Latitude/Longitude values.

Many industries depend on geographic location to support risk assessment business processes, especially in the finance and insurance sectors. Retail business of all shapes and forms are constantly reviewing locations to determine the optimal places for new stores, branches, etc. And the combination of increased mobility and a growing capability to deliver information "anytime, anywhere" suggests that business applications will have a growing need to be "location-aware."

Mapping and Spatial Analysis

Applications that track location of business interactions often will are expected to note the positions of spatial information (e.g., location, a point of interest, a route, or an address) on a map. The mapping service enables editing, spatial analysis and display of spatial information on a map. In turn, the end user may want to consider additional locations or points of interest in a relative proximity to identified positions, and a directory service matches spatial analysis with provided reference data to perform proximity searches along defined categories (such as a restaurant guide or a list of public parks) to provide a list of the closest matches.

Spatial analysis, providing a statistical, quantitative review of phenomena and activities across space, enables the derivation of knowledge about the geographic context from existing information. Knowing where an individual lives or where an event takes place provides one vector of data; spatial analysis allows one to answer more interesting questions such as:

- How many similar customers are within a specified radius?
- What are the mortality rates by disease in different parts of the country?
- Which retail locations best service a particular region?
- How many foreclosures have occurred within a particular neighbourhood?
- How many clients are serviced within each ZIP code?

Core Data Services: Basic Components for Establishing Business Value

- How has a region's median income changed over a specified time period?

Every business event or transaction happens somewhere; spatial analysis allows you to seek out location-based patterns and use the results to improve business performance.

Routing

Routing is a location intelligence service to plan routes between two points, determine the optimal path between many points and create drive time matrices and polygons. The routing service can calculate the path based on different criteria, such as distance and time, and can provide results based on optimizing for the minimum distance travelled or the minimum amount of time spent when driving. With the routing service, organisations can maintain high service levels and cost effectiveness of service dispatch, identify catchment areas and improve decision-making.

From an operational standpoint, the routing service can help optimize logistics and delivery and improve customer service. Finding the most optimal routes will reduce energy costs and may enable better scheduling to increase delivery throughput. At the same time, decreasing delivery times will improve service to customers and increase satisfaction.

Inspection and Monitoring

Because data profiling exposes potential business rules used within and across different business processes, the data analyst can document the rules and confirm their criticality in coordination with the business subject matter experts. These rules essentially describe the end-user data expectations, and can be used to measure and monitor the validity of data. Auditing and monitoring of defined data quality rules provides a proactive assessment of compliance with expectations, and the results of these audits can feed data quality metrics populating management dashboards and scorecards.

The auditing and monitoring service builds on the parsing and data profiling services to proactively validate data against a set of defined (or discovered) business rules. This data inspection can notify analysts when data instances do not conform to defined data quality expectations, as well as baselining measurements against which ongoing data quality inspections can be compared.

We have already seen that data profiling enables the analysis and discovery of potential data issues that may impact business success. But by supporting data inspection and monitoring, data profiling fulfills another critical business need for continuous data quality assurance. The data profiling service drives operational data governance and feeds information to data stewards about data failures that can be reviewed and remediated before any negative business impact can happen.

Layering Business Information Services

More and more, organizations are seeking opportunities for reducing costs for design, development, and maintenance of both internally developed and externally acquired technology. Combining the core data services with aggregated information products into standardized business information services will support different applications across the enterprise. These services typically address needs that are common across an array of different business applications, and providing standardized versions will help in reducing costs while introducing operational efficiencies. Here are some common examples:

- **Data quality validation** – When commonly used data is acquired or created via more than one channel (for example, customer data entered via a web site or transcribed by a call center representative), there are opportunities for variations in spelling, name order, or absence of titles and honorifics. Inconsistently-applied rules for validation may lead to duplicate entry of data, or will prevent records from being located in existing data sets.

PITNEY BOWES BUSINESS INSIGHT PROVIDES AN ENTERPRISE SOLUTION THAT MAKES IT POSSIBLE TO CENTRALIZE CONTROL AND IMPROVE PERFORMANCE

To address the potential inconsistency, a data quality validation service can be composed using parsing, standardization, cleansing, and profiling services. This data quality validation service can apply common sets of data standardization and validation rules to multiple data acquisition channels across the organization. Standardizing the data quality constraints and transformations across all points of data capture and exchange will significantly reduce the risks of duplication or mismatching, and will lead to data consistency across the enterprise.

- **Property location risk assessment** – Any industry that depends on assessing the risks relating to locations and properties would benefit from a standardized location risk assessment service. For example, property and casualty insurance carriers assess the risks associated with underwriting a property insurance policy, including risks of natural catastrophes such as hurricanes or earthquakes, area crime, and availability and effectiveness of emergency services. Mortgage underwriters are interested in property classification (e.g. single family vs. condominium), owner occupancy, comparable property values, and neighborhood defaults and foreclosures. Yet again, though, different business applications within the same organization may apply their risk assessment rules in different ways, using different techniques and tools.

  A standardized property location risk assessment service would incorporate address parsing and data enhancement to provide a standardized address. That standardized address can be submitted to the geographic coding and mapping services coupled with aggregated data sets containing property risk, natural catastrophe zone, and other attributes related to the geocoded location.

- **Operational data stewardship** – At many organizations there is a growing interest in operationalizing data governance policies and procedures by introducing the role of a data steward who is responsible for ensuring that information meets the needs of downstream information workers. Data quality expectations are defined by the ultimate end users of information, but as data sets created in one context are used by multiple business processes, the collected data quality requirements need to be imposed across the numerous processing stages across multiple end-to-end processes.

  Operational data stewardship services can be used to institute consistently-applied data quality inspections and monitoring within ongoing business processes. Data quality rules can be defined once and deployed using data profiling services to validate individual records or sets of records, either in place or as they are handed off between activities. A common data stewardship service will automatically inspect data for validity, and notify the data steward when the data quality expectations are not met. In addition, cumulative statistics can be collected and provided via a data governance scorecard or dashboard.

- **Demographic profiling** – One aspect of targeted marketing involves a combination of trustworthy individual data and geographic demographics. As "birds of a feather flock together," understanding the general characteristics associated with a prospect's location informs the marketing approaches for targeting.

  Given a prospect's name and home address, a demographic profiling service can parse and standardize the name and employ identity resolution to match it against known customers or to identify related parties within a known customer data set. Next, the address can be parsed and standardized. In turn, aggregated data associated with the mapped location can be used to append valuable demographic data relating to the region in which the location is mapped. Combining the demographic data with the prospect's record will provide the sales team targeting recommendations based on the materialized profile.

Core Data Services: Basic Components for Establishing Business Value

- **Jurisdictional tax management –** Another aspect of compliance and risk management involves the determination of jurisdictional tax rates. While this type of service is important for property-based risk analysis, it is also necessary for organizations with corporate tax liabilities associated with different locations.

  This type of service combines address standardization and cleansing services, geocoding and location services, and aggregated provider data to provide the correct state, county, township, municipal, and special tax district information associated with an address to track with constantly changing tax jurisdictions and municipal and jurisdictional mandates.

Considerations

These examples are just a few examples of business services that can be engineered using the core data services. Interestingly, they tend to exhibit these three characteristics as a byproduct of combining core data services:

- Parsing, standardization, and data inspection and monitoring services to normalize and cleanse data,
- The combination of cleansed data and location intelligence to prepare for data enhancement, and
- Enhancement and value improvement via appending information from aggregated data sets.

If there is agreement that sound data management practices such as data governance, data quality assurance, and data enhancement can improve the bottom line, then it is worth considering approaches to deploying standardized business information services using a core data services approach. A business information needs assessment will highlight the business services that, if standardized across the enterprise would improve quality and consistency while decreasing operational costs.

A final thought – there may be many ways to design and deploy the core data services, ranging from internally built, to acquired product suites, to completely outsourced. But if it is clear that critical business information services rely on a combination of data quality, location intelligence, and value-added data sets, one should incorporate the quality of those aspects into any buy or build decision. In other words, the solution evaluation criteria must incorporate the degree to which the products support core data quality services, location intelligence services, and integration with value-added data sets.

PITNEY BOWES BUSINESS INSIGHT: YOUR SOURCE FOR ANSWERS AND SOLUTIONS

Pitney Bowes
Business Insight

UNITED STATES

One Global View
Troy, NY 12180-8399
main: 518.285.6000
1.800.327.8627
fax: 518.285.6070
www.pbinsight.com

CANADA

26 Wellington Street East
Suite 500
Toronto, Ontario
M5E 1S2
main: 416.594.5200
fax: 416.594.5201
www.pbinsight.com

92318 AM 908