Data Profiling

Data Profiling Best Practices

Overview

Objective

This document provides an overview of best practice with data profiling activities, in particular focusing on two areas:

- **Why use data profiling technologies?** – Looks at the benefits of data profiling solutions and the circumstances in which is best deployed
- **Data profiling process** – looks at the best for deploying and using Business Data Quality's (BDQ) data practices for using data profiling technologies

Why use data profiling technologies?

Until the last few years, the area of data analysis – understanding the quality and structure of data assets within an application – was a relatively ill-defined area within an organization's IT strategy.

Traditional approaches to data analysis are usually dependent upon on a combination of inputs – documentation, individual knowledge and ad hoc data base query tools – which are used to selected aspects of a data source. Such approaches are often time-consuming and incomplete, as analysis tends to be concentrated in known areas of the data.

Data profiling tool sets, like BDQ Analysis, allow organizations to accurately and efficiently analyze and diagnose the quality of their data. By completing a process of analyzing complete data sources as one process, organizations capture a complete understanding of their data assets.

Deployment of data profiling technologies

BDQ Analysis can be deployed into any project, or activity requiring data analysis. These activities fall into one of two main core categories:

- Data quality management
- Data integration

Data Quality Management

Data quality management initiatives are focused on the process of ensuring that an organization's data assets are of sufficient quality to meet its needs. There are three main areas of interest:

- **Completeness** — Does the organization have data assets that are incomplete or missing? For example, do all customers have associated addresses?
- **Accuracy** — Is the organization's data assets sufficiently accurate to meet internal (e.g. business processes, decision making, etc) and / or external (regulatory, third parties) requirements.
- **Integrity** — Are the organization's data assets consistent across the enterprise? (For example, does the list of suppliers in a companies ERP system match those in the finance application?) Do the relationships between different data assets make sense?

The most practical approach in any data quality management initiative is to be goal oriented. Analyzing all data assets will not be an efficient use of time. Instead, effort should be focused on the information assets the organization believes to be of the greatest importance to their business. In addition, the metrics by which these data assets are to be assessed (i.e. the quality criteria, quality tolerance levels, etc.) must also be defined. Typically, one would expect these to be determined by the key users / owners of the data, typically business or operational units.

Having established which assets are to be assessed and defined the quality criteria, BDQ Analysis will assist the analysis process in several ways:

- Its scope function allows users to determine what attributes they need to examine up front and make the analysis process more targeted.
- The above processes are greatly enhanced by the ability to rapidly and automatically analyze data. Traditional manual, approaches are time consuming, error prone, and expensive. BDQ Analysis enables users to rapidly analyze data, and compare it to the business requirements.
- It allows the centralized management of data issues and documentation.
- It gives visibility as to where certain issues are occurring.

Data Integration

Data integration initiatives focus the consolidation of one or many sources of data into a new technology application. These activities are usually the result of a new business initiative, or system implementation, such as Customer Relationship Management, or management information initiative.

By their very nature, data integration projects require a solution that allows an organization to physically move data from application to another. However, differences in application design and the usage of information often mean that the data must be manipulated (be it transformations, or translations) as part of the process.

Mapping specifications are developed to achieve this; stating how the source data is to be transformed so as to load it into the target application. The accuracy of these specifications is therefore critical to the integration process; especially as the cost of resolving inaccurate data-mappings grows exponentially the closer a project gets to completion.

One key cause for this issue is an initially poor understanding of the source data to be moved. Data integration projects have a number of commonly occurring risks:

A number of factors often lead to such a situation occurring, these include:

- Definition of the target application is not finalized.
- Understanding of the source data is based original documentation that is out of date.
- Insufficient analysis of the source data as traditional data analysis techniques are costly and take time.
- The staff integrating the data is not necessarily users of the data.
- Scope of source data to be integrated increases as understanding of the source data improves and / or target system requirements change.

BDQ Analysis addresses many of these issues by providing a thorough understanding of the source data to be moved.

Data profiling process

The methodology contained herein takes a centrist approach to the problem and will need adjusted based on the goals and desires of the specific project at hand and the features and functions of the tools used. For example, if the tool supports loading samples within its functionality, the extract programs

will not need to concern themselves with this aspect. On the other hand, if the data profiling tool does not support sampling this functionality (if required) will need to be included in the extract programs requirements.
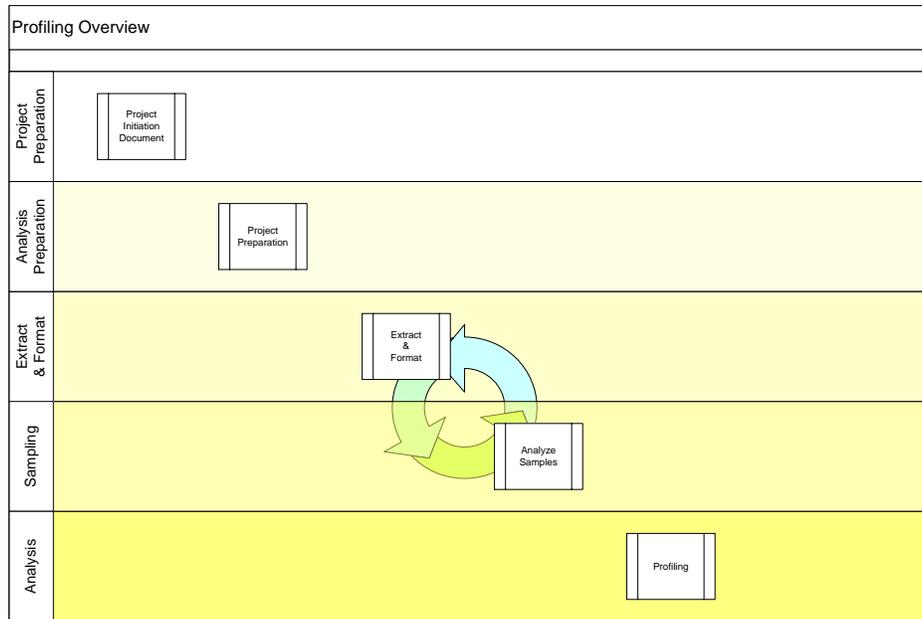
(Diagram 1)

The approach consists of five steps or Phases:

1.   Prepare for the Project

2.   Prepare for the Analysis

3.   Extract and Format the Data

4.   Sampling

      a.   Load a Sample of the Data

      b.   Analysis of the Sample

      c.   Adjust the Extracts and Formats of the Data

      d.   Produce deliverables

      e.   Delete the Samples

5.   Analysis

      a.   Load the Data

      b.   Perform the Analysis

      c.   Produce deliverables

Prepare for the Project

This Data Profiling methodology is not intended to take the place of the Project Plan Methodology used to create the Project Initiation Document, which is covered under a separate publication. However, the major components are listed here to maintain continuity and reference for the "profiling" aspects of the project. Remember, the Project Initiation Document allows you to manage expectations.

The Project Initiation Document serves four major purposes:

- The data profiling project works within the boundaries established by the Project Initiation Document.
- The data profiling project produces the deliverables as outlined in the Project Initiation Document.
- The data profiling project communicates via the policies and procedures as identified in the Project Initiation Document. This identifies the communications to those closely involved in the project as well as those of general interest in the project.

  Assigned work load. Typically one or more Entities would be assigned to a single analyst. In the case of very wide entities or short time frames, attributes of one entity may be assigned to multiple people. Be sure the analyst are clear as to the boundaries so as not to miss entities/attributes or have redundant work occurring

The Project Plan or Project Initiation Document contains the following sections.

      I.   Background

      II.   Scope

      III.   Deliverables

      IV.   Project Team

      IV.   Communications Management

      V.   Change Management Plan

      VI.   Risk Management

      VII.   Cost Management

      IX.   Task List

X.   Gantt Chart

XI.   Network Chart

Analysis Preparation

Gather, train, and otherwise familiarize the analysis team with the materials available that may assist in the analysis, the tools being used, and setup the Data Profiling tool for use.

Exactly where you start depends on the initiative being undertaken. For most conversion projects the target is known and it is most effective to drive towards that goal. For a data quality initiative there is no target and the focus needs to be the correct data in the correct column and the consistency of that data.

Review Project Initiation Document

The Project Initiation Document will contain the deliverables required to be produced, the time frames expected, and the boundaries of the project. The team must be familiar with this critical document; else excessive time may be spent on irrelevant matters.

Current Documentation

Gather and gain familiarity with all pertinent documentation.

- Gather current system documentation,
- Target systems documentation
- Logon's and passwords to existing metadata repositories,
- Other Information Sources

This review is intended to know what documentation is available, not become fully cognizant with all the details.

Team Training

The team must not only be trained in the use of the tool being used, but in the deliverables expected and the process and policies being used on this project as well as any other tools employed in moving the project to a successful completion.

Pay particular attention to functionality that will require additional effort in the extract and format process. I. E. If the profiling tool does not allow you to dynamically parse and reanalyze data this will have to be done in the extract and reformat process.

Internal Setup/Decisions

A few decisions need to be made on how the data is going to grouped internally. BDQ Analysis supports virtual grouping of entities using its "System" construct. You can choose to group data by functional area or by physical system.

The team should decide on what the scope of the activity is to be – i.e. choose the appropriate systems, entities and attribute to examine. This decision is usually driven by external requirements, for example:

- Particular data within a company must meet compliance regulations
- The information requirements of a target system within a migration project.

BDQ Analysis allows the user to set a Scope flag on all attributes, which is available from all attribute lists. The team can select all attributes that are within scope at the beginning of the exercise. During the analysis process, users can view this information, and ignore attributes that are not in scope. This can help to prevent "paralysis by analysis".

Users can see attribute scope at any time by using the attribute viewer windows, at either project, system or entity level. The filtering or grouping functionality available in these windows can assist with seeing what is in and out of scope.

Activity Workflow

There are three main roles in analysis projects. An individual may cover more than one of these roles.

- The project manager: interested in issues, scope, progress
- Data analysts: staff who are actually analyzing the data, and looking for candidate anomalies
- Business users/analysts: staff who have particularly detailed knowledge about the data content, and who can identify whether data is correct or not

It is important to ensure that all data that is in Scope is analyzed, and that no analysis work "falls through the cracks". There are two perspectives here:

- Each user needs to know which Entity has been assigned to him
- The team as a whole needs to view what Entities have been assigned to whom, and what if any, have been left unassigned.

Each user can view all the items assigned to them from the "Users->Username" node within the tree. Filtering by Entity will show all Entities.

In order to view the entities assigned to particular people, this can be performed from the Project Entities or System Entities windows. Filtering on Assigned To = "Blanks" will reveal all those entities which have not been assigned to staff.

During a data profiling exercise, there are two main types of activity that must be monitored to ensure that the project stays on track:

1. How much of the data that was in scope has been analyzed

2. How issues that have been found are being progressed. On occasions, candidate issues may be found by a Data analyst, but have to be referred to a Business analyst who has deeper knowledge of the specific data.

Ultimately, the goal is to finish analyzing all data that is in scope, and to have closed off all open issues.

To support (1), when analysts have finished analyzing data to their satisfaction, they can mark the attribute as analyzed. This can be done from the profile view, or the attribute list views available at Entity, System or Project level.

It is then simple for team members to view what attributes have been analyzed by using a filter or group on any of the attribute list views. This can be combined with a filter/group on the Scope flag, so that it is possible to see all attributes that have been analyzed or not, within the scope, and thus track progress.

For example, you can open a System view, set the filter on "In scope" to "Y", and then group by "Analysed". You will then have two groups – attributes that have and have not been analysed, all within the given scope.

To support (2), BDQ Analysis allows the creation and assignation of Notes, with reporting functionality to assist team management. Analysts can create notes, and then assign them to specific individuals for resolution, along with priority and status information. Individuals can log in to BDQ Analysis to see what notes are assigned to them, or alternatively, notes can be exported to HTML or Excel and sent to the appropriate users.

To see all the notes assigned to a given user, the user can open the "Items Assigned To" list from the Users node on the tree (for a given user), and filter on Type="Notes". This information can also be seen from the Project Notes window.

When an issue has been resolved, the note can be marked as closed.

Managers can track the progress of note status, and analysis status by using the Notes windows. By using filters, it is possible to see what issues are still open, their priority, and whom they have been assigned to. For example, a manager could filter on Status='Open', and group by "Assigned to". This would provide a grouped view by name of all open notes, whereby it is simple to chase up those notes that are not getting resolved.

To facilitate this process, when creating users it is helpful to use the user description field so that analysts can easily choose the most likely candidate user to request research from.

Extract and Format the Data

This activity consists of creating the extracts and any required format definitions required by BDQ Analysis.

This activity is for data files that:

- Cannot be accessed directly via ODBC connections,
- Those data files where suitable production file formats do not exist,

- Where data is not stored at the granular level required to support the analysis (when known), and

Create the Extract Program(s)

Creation of the extract program(s) should be treated as with any other program development, that is, requirements, code walk-thrus, test plans, and all the other items that create a well behaved functioning program. An error here is liable to make the analysis of this data inaccurate.

Load Preparation

Prepare the data, file definitions, and in general, prepare the systems to support the analysis to be performed.

Prepare the data to be analyzed

- Create csv,
- Create flat file and flat file definition,
- Create appropriate ODBC connection as required

Csv	Each field if separated by a comma and text fields are enclosed within quoted. Watch for fields that contain double quotes representing inches and such. Generally this type of file allows the first row to contain the name of the column.
csv File Definition	Some product require or allow you to create of definition for csv files. This can be handy to add or change column names to the file or in some cases add descriptions to the attributes.
Flat File Definition	This varies based on the data profiling product chosen. It varies from a flattened copybook or equivalent for the language used to pre-defined formats specific to the tool itself.
ODBC Connection	Open DataBase Connectivity, a standard database access method developed by Microsoft Corporation. The goal of ODBC is to make it possible to access any data from any application, regardless of which database management System (DBMS) is handling the data.

Sampling

Preparation for this Phase is somewhat on the condition of the documentation, the knowledge of the analyst or the experts available to the analyst, and the volume of data to be loaded.

This sample load is quite important as it allows you to:

- Determine any fields that are incorrectly defined in the associated file definition.
- Determine fields that may have been totally misunderstood and should be defined differently (I. E. a field that was defined as a date that contains other information of interest).
- Identify fields that should be separated into component parts for analysis (I. E. a concatenated key field, of even a telephone number field to separate analysis on area codes or exchange codes from each other).
- Identify fields that are out of scope.
- Identify tables of values that may need to be created and loaded into the system.
- Allow determination of the correct level of granularity for the fields being analyzed.

Running the sample step can allow significant savings on large files that may take hours to load and then reloaded once a basic understanding has been achieved.

Load a Sample of the Data.

Many types of sampling exist. Try to choose one that will provide the best results in the shortest period of time. Where possible work within what's available in the data profiling tool. The first 100 or so generally provide enough information and save processing the entire input file for gathering the sample.

Analysis of the Sample

Data analysts can analyze the data by Attribute. All the attribute profile information is available in the Entity Profile viewer. Assuming that the scope flag has been set, only those attributes that are "In Scope" should be analysed. Of course if an analyst sees something in another attribute that he feels may be important, he can raise a note.

When the Data Analyst has finished analyzing the attribute, he can set the "Analyzed?" flag to 'Y' to indicate that analysis is complete. If he sees issues within the data, he can create a note by right clicking on the Attribute and selecting "Raise Note". This note will then be linked to that Attribute to aid subsequent reporting. If the analyst feels that other attributes should be involved with the note, he can drag those in to the "Attached to.." tab.

Alternatively, the notes tab on the Entity itself can be used to report anomalies that the Analyst considers to be Entity specific.

Large free form text fields are generally not benefited by Data Profiling. However, it is not uncommon for these large fields to contain some cryptic codes in the beginning of the field. Consider splitting the first twenty of so positions into a separate field for analysis and bypassing the remainder.

In some cases it is desirable to split or combine fields prior to loading. Concatenated Keys, telephone numbers, address, etc. If known these fields can be adjusted prior to the load or data can be exported form the tool and re-imported. It is only worthy to note what the impact this may be on your overall goal(s).

Adjust the Extracts and Formats of the Data

Based on the results of the analysis make any adjustments required to the data extract program(s)

- Misunderstood Data (incorrect documentation or rusty expert knowledge)
- Finer Granularity Desired – data is not at the atomic level.
- Bypassing of out of scope fields

The intent of the sample analysis is to identify the gross discrepancies that can be aided by refinements in reformatting – do not get wrapped up and get into the detail analysis that is best identified in the analysis section.

BDQ Analysis supports multiple profiles for an Entity, so analysis can be concentrated around data sets with particular sampling characteristics.

Produce deliverables

In/Out of Scope

Delete the Samples

Once data has been analysed, the data profiles can be deleted. This will not remove issues or metadata, but will free up space on the Analysis server. The ability to drilldown to records and patterns etc. will not exist, however.

Analysis

The actual analysis can very based on ones knowledge of the business, the system, and of course the results desired. However, the approach below, while not all encompassing sets the core that, where appropriate, needs to be done.

The analysis results can be found by examining the Single Pane Analysis windows for each Entity Profile.

Analysis Assistant

BDQ Analysis auto generates analysis results for the user. These results can be used as a sanity check, or to prompt further investigation. The results can be found in the Assistant tab, within the Entity Profile window. The list is as follows:

- Outliers: this indicates that there may be anomalous data for a given attribute. Further investigation of the frequency value pairs should be performed to see if anomalous values exist.
- Primary key: this will inform the user as to whether the attribute is a candidate primary key, or whether is potentially corrupted.
- Code: this will indicate that the field may be a code field. If it contains Outliers, these may well be values that are not valid. Checking them with a business user is a sensible precaution.
- Indicator: this will indicate that the field is used as an indicator field.
- Constant: this indicates that the field contains only one value
- Pattern outlier: this indicates that there may be outliers within the patterns.

Assessment

Blanks/Nulls/Low Values/High Values

Blanks, nulls, low-value, and high values give a strong indication of the current condition. If the field is a key field this generally represent an erroneous condition. While all null usually represent a field not be required by the business and remnants (screens, reports, etc.) should be removed.

Obviously, if this field is required by a new system a way of populating it (pre or post conversion must be identified)

Low Value	High Value
000-00-0000	999-99-9999
NULL	

If this field were required it would be known that these records are in error and would require an approach to find the correct values.

Minimums/Maximums

Minimum and maximum values can help to quickly show that the data needs additional research.

System	Minimum	Maximum
System 1	000-00-00001	

We know that 0's in the first three, second two, or third four positions would be an error (everybody knows that). It would also seem unlikely to be correct if we found a range of numbers, say from 1 to a hundred.

Low values / high values are generally indications of problems if the data.

Ranges can sometimes also be identified by the low and high values.

Patterns

Patterns are one of the best indicators. If I need to standardize onto a format, for whatever reason, patterns will immediately give me the systems and records in violation.

System	Values	Pattern
System 1	123-45-6789	9(3)-9(2)-(4)
System 1	12-3456789	9(2)-(7)
System 2	123456789	9(9)

The above example shows several things. Assuming that there were no other variations in the formats:

I can determine that both the SSN's and the TIN's are stored in a formatted fashion in system 1. If storage of this attribute was intended not to carry the format I would need a business rule to remove the special characters.

If both systems carried individual and corporation tax identifiers I would need to find a code or some other means of determining individuals from corporations in system 2.

Duplicates / Inconsistencies

Some fields should not contain duplicates. Finding the number of occurrences quickly identifies the items that need to be researched and corrected.

System	Values
System 1	123-45-6789
System 1	123-45-6789

Finding the number of occurrences to be more than one within the same system (in this example) might indicate an error condition. Further research would be required.

Invalid Codes

For "flags" and "codes" such as state codes, country codes, color codes, and the like create a table of these codes and join to the content field. This will allow missing, misspelled, or unexpected values to be identified

Identify Keys

Where keys exist or are required a unique field is required.

System	Values
System 1	123-45-6789
System 1	123-45-6789

If this field were intended to be the key the above occurrence would require further research/correction using the Key testing functionality described below.

Key testing

You may wish to determine whether duplicate values exist for collections of attributes. For example, you may be testing for composite keys. If there are duplicate pairs of values, you can drill down to the underlying records for further examination.

Join testing

This allows you to test whether attributes from different table join together. This is typically used when data is being duplicated across systems, and orphan records begin to creep in. Essentially, you may choose to test a join between two attributes, and BDQ Analysis will give you detailed information about the join, and any integrity issues that may be present. You can then drill down to the actual records to see the problems in detail, and raise a note if necessary.

Outputs

Data Quality Specification

- Complete analysis of data in scope
- Summary of analysis activities (using BDQ Analysis' notes functionality)

Data Integration specifications

- Input into mapping specification

Business Data Quality Ltd (BDQ) is an independent software company specialising in next- generation solutions aimed at helping organisations realise the true potential of data as a strategic asset through the development of practical data analysis and management products.

BDQ Suite allows you to analyse, track and report on the quality of data stored within IT systems. Combining data profiling and monitoring technologies with a multi-user repository and issue management framework, BDQ Suite delivers a detailed and accurate view of corporate information from a data quality perspective. Aside from issue identification, it also addresses the second core ingredient of an effective data strategy – quality management. Administrative data activities and remedial cleansing tasks can be documented, assigned and tracked through BDQ's integrated issue management platform, giving visibility to improvements in quality and, thereby, raising confidence across the business.

For further information about BDQ products and services, please contact us at:

Business Data Quality Ltd	Tel: +44 (0) 844 826 5236
2nd floor, Walsingham House	Fax: +44 (0) 208 211 9794
1331 – 1337, High Road	
Whetstone	
London	
N20 9HR	enquiries@businessdataquality.com
United Kingdom	www.businessdataquality.com