THREE STEPS TO A FLEXIBLE DATA MART:

How to extend your business intelligence tools

AN ARBUTUS SOFTWARE SOLUTION PAPER

EXECUTIVE SUMMARY

The demands for timely, accurate, and critical analysis and reporting in a complex and disparate data environment are greater than ever. To successfully tackle the problem, both business and IT professionals are turning to flexible technology solutions.

An effective, modern, flexible data mart solution should ensure:

- Rapid, cost-effective deployment providing secure access to all necessary data across the organization and within its disparate systems
- Quick adaptation to changing data and evolving needs
- On-demand generation of accurate business intelligence by the users who actually need it, when they need it

Flexible data mart solutions from Arbutus address each of these issues, allowing organizations to use every piece of corporate data, no matter how disparate, as if it were already in a data mart. These flexible data marts can then be used on-demand to generate the analysis and reporting required.

The following is an example of the reporting challenges faced by one organization and the steps they took to achieve a flexible data mart.

A DEPARTMENT'S REPORTING CHALLENGE

In order to create departmental management reports, analysts at a large municipal government spent days each month repurposing mainframe data from standard reports produced by their in-house business intelligence (BI) application. The BI application had been programmed years earlier to access specific data in their IMS database, in order to create the standard periodic reports.

Over time, city programs evolved or were added, meaning that reporting requirements were also modified. As with most city governments, responding to these changes with a programmed solution was a slow and potentially costly challenge, particularly because of the specialized – and scarce – mainframe expertise required for any changes.

Not surprisingly, this rigid solution failed to meet many of the new reporting requirements, while the 'must-have' reports required too many days of staff time each month.

From a technical perspective, the department faced three major problems:

1) Adding new or changed program data from fields and tables in their existing IMS database that were not part of the original requirements for the static data reports

2) Dealing with data that was improperly formatted or of questionable quality

3) Providing cleansed mainframe data directly into Microsoft Excel for use with established report templates that presented data in the format required by management

Many governments and organizations face similar challenges as they try to reconcile existing BI solutions and IT environments with new or changing reporting needs. This solution paper addresses the need for increased agility and focuses on the following capabilities:

- Providing access to some or all data that is not currently in a data warehouse
- Supporting dynamic decision-making, even in cases where the path to the outcome was unclear ahead of time
- Allowing disparate data sources to be easily combined
- Enabling changes to occur quickly and with minimal resources

3 STEPS TO FLEXIBILITY

The following three steps were taken by the city to achieve the reporting and analysis flexibility they needed with a flexible data mart:

1 – Identify the Required Data

2 – Test and Resolve Data Quality

3 – Select and Set-up Tables and Fields

STEP 1 – IDENTIFY THE REQUIRED DATA

A key component of a flexible solution is that all data sources – whether they are part of a core database or part of a legacy system – can be included for use. You start by identifying the required data. This would commonly involve working with IT or other subject matter experts.

Next, you should ensure that the identified data sources contain all of the required data elements. Arbutus allows you to view any selected data sources immediately, so that you can confirm that the data in each field is what was expected. This ability to view the actual source data trumps any metadata information that might traditionally be relied upon.

After the initial installation and set-up of Arbutus Analyzer and the Mainframe server (a half-day task) the city was ready to proceed with Step 1. Using existing COBOL copybooks, analysts were able to use Arbutus to define the various IMS segments. However, within 15 minutes, the analysts realized that one of the required data elements was not in the IMS database, as they had originally been told. A quick inquiry confirmed that it was, in fact, contained in a VSAM file. Again, a COBOL copybook was used to quickly access and view this new data source and verify that it contained the data needed.

Arbutus in Action: Step 1

- *In order to verify some of the data, Analyzer's Smart Search was used to enter sample customer information that should have been present*
- *A brief report was printed containing real data from the IMS and VSAM data to visually confirm the data*

STEP 2 – TEST AND RESOLVE DATA QUALITY

Regardless of the source system, data quality (DQ) is almost always an issue. This may be because of poor edit checks or where the usage of columns has been repurposed over time. It is important that you don't make assumptions about data quality, and a flexible solution shouldn't require you to do so.

There are two distinct aspects to data quality: identification and cleansing. A flexible solution should, at a minimum, allow you to identify all relevant DQ issues. Cleansing is a more complex issue, because it often entails correcting the source data. A data mart will not normally cleanse source data, but it should provide flexible capabilities to correct most issues in the data mart. Ideally, this is accomplished dynamically by the end-users, as issues are discovered along the way. This information can also be used as part of a feedback loop to correct the source data.

STEP 2 – TEST AND RESOLVE DATA QUALITY *(CONT'D)*

As part of step 2, the city found that the purchase order system was allowing the letter 'O' to be accepted as a zero. It also found that the system allowed both lower case and upper case letters to be entered.

Utilizing the purpose-built data analytic and transformation functionality in Arbutus Analyzer, virtual fields were easily and permanently established to test for and correct the poorly formatted or faulty data. The resulting data mart thus contained only normalized and cleansed data. A number of other data issues were similarly – and quickly – corrected.

Arbutus in Action: Step 2

- *Conditional computed fields were created to test and correct data dynamically*
- *The Verify command identified missing dates and invalid characters in certain fields*
- *The format of the purchase order should have been 99-9999XX. Verifying with the FORMAT function quickly found not only the correct format, but also 99-X999XX, 99-9999xx, and XX-9999XX*
- *The Classify command was used to identify every unique data format in the file, including those noted above*

STEP 3 – SELECT AND SET-UP TABLES AND FIELDS

Once you have identified and tested the data in your data mart, you should have flexible deployment options. You should also be able to easily add or delete columns or tables at any time. Because the creation of a data mart often arises out of a need to access certain data sources that are not already in a data warehouse, the flexibility to choose all types of data – especially difficult legacy data sources – is a key part of flexibility.

Similarly, based on the technical considerations, you should have the flexibility to choose the hosting platform for the data. With a truly flexible data mart, you should not be required to implement complex or expensive hosting environments, particularly when the value of some newly implemented data has yet to be determined. Being able to implement a "virtual" data mart is also a key component of agility. Once all factors have been considered, there should be no barriers regarding hosting, while ETL should be kept to a minimum.

STEP 3 – SELECT AND SET-UP TABLES AND FIELDS *(CONT'D)*

For the utmost in flexibility, the city chose to both stage their IMS and VSAM data on a Windows Server, and set up direct access to some IMS tables. They also required the staged data to be refreshed on a daily basis. Using Arbutus, the ETL process – which consisted of moving and transforming the data from the z/OS platform to the Windows server – was created in about two hours and scheduled to run every weeknight.

End-users began to use their various reporting and analysis tools the same day the process started. The analysts continued to use Excel on their local workstations, but with Arbutus, they were able to directly access the city's critical mainframe data, as it was now fed directly into Excel reporting templates. Furthermore, the Excel reporting templates were set up to automatically refresh the source mainframe data upon opening, so end-users generated accurate and reliable business reports on demand.

Arbutus in Action: Step 3

- *Powerful functions and conditions were used to create harmonized 'virtual' key fields needed to facilitate joining (physical) and relating (logical) various tables together*
- *With Arbutus LegacyLink™ deployed on each analyst's workstation, staged mainframe IMS and VSAM data can be easily accessed with any application that is ODBC- or JDBC-compliant*
- *The Vendor ID field between the IMS data and VSAM file was structured differently: A computed field was created with the VSAM table layout that removed a dash ('–') and also lengthened the field by three characters, allowing a straightforward join to be created as part of the ETL process to the Windows Server*

ACHIEVING AGILITY WITH ARBUTUS

Most industrial-strength BI and ETL solutions available fulfill a good portion of their reporting and analysis tools requirements. Yet, by their very nature, they are rarely considered flexible.

This is where Arbutus comes in.

Data marts from Arbutus are fast and flexible solutions. The agility inherent in the Arbutus solution presents a number of valuable opportunities for municipal governments and organizations alike.

ACHIEVING FLEXIBILITY WITH ARBUTUS *(CONT'D)*

Data marts from Arbutus can be used to:

- Implement short-term or temporary solution requirements
- Prototype a data mart before implementation using existing methodologies
- Gain a better understanding of the source data before implementing a data mart
- Provide a more cost-effective solution to accommodate limited budgets or when the business case does not justify large expenditures

The Arbutus data mart solution provides:

- The ability to rapidly deploy a flexible solution. This can dynamically expose and blend necessary data from disparate sources using technology optimized to process large amounts of corporate data.
- The flexibility to immediately adapt the solution to changing data needs as well as evolving or emerging business intelligence requirements. Exposing new data, cleansing data, defining new relationships between disparate data sources and modifying existing data definitions can all be done interactively within minutes.
- The option to immediately expose the mapped data directly to end-users using the Arbutus Analyzer analytical tool, or to tools of your choice (e.g., Excel, Access or SQL Server), via Arbutus LegacyLink™.

No matter what its intended purpose, an Arbutus data mart is a pragmatic, cost-effective solution compared to the time-consuming and inefficient alternatives some organizations currently use.

TECHNICAL OVERVIEW - AN ARBUTUS FLEXIBLE DATA MART

Arbutus Analyzer – designed to handle large volumes of corporate data and quickly identify issues, Analyzer has a wide range of unique analytic capabilities that will have you producing results in minutes. A key capability of Analyzer, as demonstrated in Steps 1 & 2, above, is that there is no need to prepare or stage the data prior to adding to the data mart and verifying or testing data quality.

Arbutus LegacyLink™ – an ODBC driver that allows any of the data in the data mart to be accessed by your existing tools, this driver is also fully compatible with the other common data access protocols, including .NET, JDBC, and OLE/DB.

TECHNICAL OVERVIEW - AN ARBUTUS FLEXIBLE DATA MART *(CONT'D)*

Choose and Verify a Wide Range of Data Sources

Mainframe sources include:

- VSAM, ISAM, and QSAM, generation data sets, data stored in PDS files
- Data formats include fixed length, variable length, and multiple record types
- Data sources include tape, cataloged data sets, DB2, IMS, and ADABAS
- Data types include ASCII, EBCDIC, Packed, Zoned, Binary, Float, as well as proprietary data types from a variety of other platforms. Through the use of virtual fields, even bit fields and other complex data can be implemented

Arbutus data marts can also include data from Windows, UNIX, and AS/400 platforms, as well as data downloaded from most other legacy environments. Data can also be automatically integrated from web-based sources.

Test Data

Capabilities such as:

- Verify – automatically identify every corrupted data field in a table, regardless of the data type
- Classify – easily identify every unique value in a coded field
- Gaps, Duplicates, and Sequence – verify the correctness of data keys
- Filters – instantly identify selected data, whether for analysis or data quality testing

Virtual Fields

Virtual fields can encapsulate any level of complexity, from simple expressions to complex conditional logic, and are commonly used to:

- Apply data cleansing rules
- Test for or apply business rules
- Augment the source data as required

Procedures

Procedures in Arbutus can be used to efficiently implement ETL or complex analyses. Procedures can be scheduled, triggered or run ad-hoc. Triggered or ad-hoc procedures can optionally include run-time user input or parameters.

Combine Data from Multiple Files

Join data based on shared keys, or by defining star-schema or snowflake-schema relationships between the tables. When two or more tables share a key that is formatted differently, it is easy to harmonize the keys through the use of virtual fields.

TECHNICAL OVERVIEW - AN ARBUTUS FLEXIBLE DATA MART *(CONT'D)*

Procedures

Procedures in Arbutus can be used to efficiently implement ETL or complex analyses. Procedures can be scheduled, triggered, or run ad-hoc. Triggered or ad-hoc procedures can optionally include run-time user input or parameters.

Customize Data Set-up

Dynamically add or delete tables, columns, or relationships. In addition, selectively hide or show existing tables or columns at any time, and create custom interfaces to different users or groups.

- Visit www.ArbutusSoftware.com/form-eval.html to request a 30-day free trial of Arbutus Analyzer.
- Request a 20-minute web demo to see a flexible data mart in action. Email us today or call toll-free.

ARBUTUS

#270 - 6450 Roberts Street
Burnaby, British Columbia
Canada V5G 4E1

Toll-free: 877.333.6336
Direct: 604.456.6336
Fax: 604.437.7872
info@ArbutusSoftware.com
www.ArbutusSoftware.com

Based on 25 years of innovation excellence, Arbutus delivers the very best in purpose-built audit analytics technology to meet the exacting demands of today's business environment. Auditors, business analysts, and fraud investigators rely on Arbutus to enhance their testing, analysis, and compliance capabilities.

The data universe is wide and varied. One of our core strengths as a technology firm is the ability to easily work with all types of data, both legacy and non-legacy. Arbutus solutions allow auditors, IT, and business professionals to overcome many of their current constraints in areas such as data migration, data quality, fraud detection, and data analysis.

Arbutus Audit Analytics, our flagship product suite, is a proven solution used by auditors, business professionals, IT, and management all over the world. With outstanding customer service, strong product support, and flexible licensing, Arbutus Software offers the best value for advanced data conversion, migration, and analysis solutions.

Arbutus Software Inc. is a privately held company based in Greater Vancouver. For more information about our company or products, please contact us.