ARBUTUS

With hundreds of unique and advanced capabilities, Arbutus is fast, effective, and efficient – ideal for today's audit work environment.

Based on 25 years of software innovation excellence, Arbutus will help you achieve the highest level of business assurance within your organization.

Already using data analysis software? Learn how Arbutus is compatible with your existing audit analysis tools. Switching is fast and easy.

Contact us today and get started with a free 30-day license of Arbutus!

Data Warehousing:

Legacy Data Access Challenge at Public Utility

THE PROBLEM

A large public utility wanted to create a data mart populated with data from their customer master file, which contained information on every account and activity in the region. It was a massive, complex file that had evolved over decades to meet the needs of the corporation. The file was so large, that if flattened, it would be *terabytes* in size.

This file was their primary legacy system. (For complete details on this massive file, see the Appendix on page 2.)

Access to it in the past had been provided by special purpose extraction routines, a costly and inflexible process. In addition, the data and reports produced by the extraction routines did not reconcile well with the production reporting processes. As such, the uses to which the data could be applied were very limited.

A number of large ETL (Extract, Transform, Load) vendors submitted proposals, but each involved a similarly complex PL/1 program, so the utility looked at other options.

THE SOLUTION

Arbutus Analyzer was selected by the utility because it could directly read their complex master file, and at a fraction of the cost of other proposals.

THE PROCESS

Working with Arbutus technology specialists, the team began the Discovery Phase, defining and profiling the data. This step turned out to be the largest and most significant, as they soon discovered that the file was rife with undocumented transaction types and unforeseen exceptions.

Nevertheless, a small team tackled the project by using Arbutus Analyzer to create virtual columns and data models that mirrored what they expected in the file. Then they iteratively addressed the largest differences, refining their understanding of the actual business processes and reducing the differences with their production reports. In the end, they were able to reconcile the major systems to within 1%. This was by no means ideal, but was still an order of magnitude better than their previous best efforts.

RESULTS

Armed with a more complete understanding of their data, they were able to quickly create the appropriate transformations to match the model of the data mart. The entire process took three months, of which virtually all of the time was spent in the discovery phase of the project. When complete, they declared that this was their first successful data warehousing project.

ARBUTUS

Data Warehousing: Legacy Data Access Challenge at Public Utility

APPENDIX: THE COMPLEX DATA FILE

Once you see the complexity of this data file it will be easy to understand why other products were unable to read it. Luckily, most files are not this complex, but it often doesn't take much to make a file inaccessible to most tools.

Facts about the Public Utility's Customer Master File

- Physically, the file is IBM variable length, 7GB in size. It is updated nightly in batch by a very complex legacy PL/1 application.
- The file is blocked at a fixed 8,000 bytes, with records being spanned across blocks.
- There is one record for each account, with each record containing a variety of different segments, each describing specific activities or facts.
- Each record contains a 50 byte header, of which 14 bytes is reserved for a bit array identifying which of 112 possible segments are included in this record.
- The segments themselves cannot be identified by their content. The only means of identification is to rely on the fact that they are stored sequentially in the record, and that the next data relates to the next bit in the array that is set.
- The original designer likely thought that 112 would be more segment types than they would ever need, but of course this assumption was wrong. At some point in the past, they had used all 112 segment types and needed more. They addressed this by making one of the latter segments an additional 10 byte array of bits, which logically extended the first and identified the presence of an additional 80 possible segment types. Needless to say, this is only present if one of the 80 new segment types is included in the record.
- The lengths of segments themselves followed no particular pattern. Many are of a fixed length specific to that segment type, but the length isn't included in the file. This has to be discerned from the PL/1 copybooks.
- Other segment types are variable length. At least for these, a standard is followed. The first two bytes are a binary segment length. Most variable length segments are actually instances of a repeating block of data of the same type, where the number of occurrences must be inferred from the total segment length divided by the length of each element (again, only available from the copybooks).
- Taken as a whole, across all the segment types in use, there are over 9000 fields. Each is stored in the densest format practical. For example, all dates are stored as two byte binary values, counting the number of days since April 1, 1940.
- Many of the individual segment types include codes to specify the nature of the particular transaction(s). For example, a segment holding cash adjustments would include a code indicating the nature of the adjustment. Of course, the sign of the adjustment amount is specific to each adjustment code, with the "normal" transaction being positive.
- *It is estimated that if the source file were flattened using conventional techniques, the resulting flattened file size would be in the terabytes.*

To request a free 30-day evaluation of Arbutus Analyzer, please contact us.

"Arbutus has outstanding customer service and technical support. Its delivery is above and beyond any other data analysis software I have used."
Margie Reinhart,
CEO, Reinhart Forensic Consulting LLC

"Support by the Arbutus team has been second to none. Often there has been a response time of less than 10 minutes, and they don't mind 'simple' enquiries either. Things are explained in a non-technical manner, which is very helpful."
Dick Price
S2Mprofits.co.uk

ARBUTUS SOFTWARE INC.
#270-6450 Roberts Street,
Burnaby, BC V5G 4E1 Canada

Toll Free: 1.877.333.6336
T: 604.437.7873 | **F:** 604.437.7872

General Inquiries:
info@ArbutusSoftware.com
Technical Support:
support@ArbutusSoftware.com

ArbutusSoftware.com