

Data to Information - Ab Initio Capabilities

At BitWise we understand that a company's profitability hinges on its ability to act swiftly and make sound business decisions, based on a complete and accurate single view of customers, suppliers, and operations.

However, the critical information required to gain this 360° view and make these key decisions is scattered throughout the enterprise, across multiple applications, departments and divisions. And while each source system contains pieces to the puzzle of enterprise profitability, each is an island of raw data with a gulf of incompatibility separating them.

This paper describes some of the ETL processes BitWise has implemented using Ab Initio to bridge these islands of raw source data to create Information using approaches that eliminate proliferating, distinguish when applying source-specific and non source-specific business rules and phased transformations.



Executive Summary

BitWise has a dedicated Center of Excellence (CoE) for Ab Initio which has helped BitWise to create optimized solutions built on Ab Initio making the Data Warehousing initiatives efficient thereby maximizing the ROI.

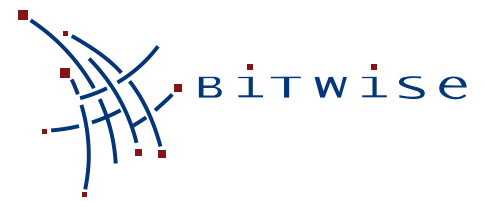
BitWise's project experience in Data Warehousing space using Ab Initio includes:

- » Development and Maintenance of ETL applications
- » Development of non ETL core applications
- » Migration of legacy ETL Applications to Ab Initio
- » Development of List Pulls
- » Adhoc requests
- » Production support of ETL & non ETL applications

All of these projects have been developed using the Best Practices derived from our CoE and recommendations by Ab Initio.

This paper describes the Ab Initio capabilities developed by our Ab Initio CoE using the vast experience, gained while working with Ab Initio on both ETL & non ETL applications.





Capability Framework

The development of Ab Initio applications is based on the framework created and adopted by BitWise. This framework is extensively researched and built to ensure it covers the following important aspects of the development

- » **Innovative Approach**
- » **Coding & Documentation standards**
- » **Parameterization**
- » **Scripting**
- » **Performance Tuning & Reliability**
- » **Reusability**
- » **Data Quality Assurance**
- » **EME for Source Code Management**

Innovative Approach

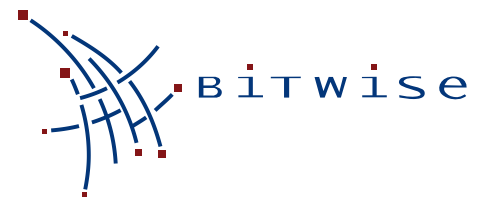
More often than not each project has to face some unique challenges of its own. These challenges are the opportunities to innovate ideas for the benefit of the project. At the end of the project these innovative ideas become milestones achieved and benefit other projects in the organization. These innovative approaches have transformed into reusable components, some of which are described below:

SFTP Custom Component

There was a need to transfer sensitive data file over the network. There was no feature in Ab Initio available that time to achieve this. Ab Initio CoE team at BitWise stepped up to develop a custom component that will first establish a secure layer connection with the peer and then transfer the file over the secure network. This component was later used by the client across all Ab Initio projects for secure data file transfer.

Write Block Compressed Lookup

The challenge was to scan the daily transaction file having around 3 million records for identifying the fraud accounts. The unique values for fraud account were in couple of thousands. There was a major performance gap due to this. The WRITE BLOCK COMPRESSED LOOKUP component was created precisely for scanning huge files for specific values. It used the concept of Data Block, Compression and Indexing in the backend to optimize the performance. The result - 90% performance gain.



Coding & Documentation Standards

BitWise CoE for Ab Initio has put standardization in place for better traceability and predictability of the deliverables. Standardized naming conventions for graphs and jobs make it easier to associate the graphs/jobs with the main stream application pools. This is further supported by strict format and naming standards followed while developing the XFRs, DMLs, field names and component names.

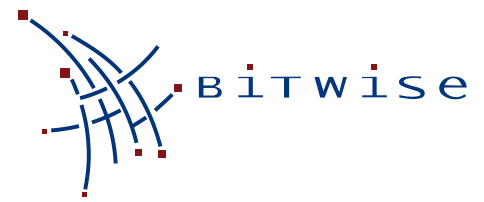
Explanatory comments are written in the graphs which help new developers in understanding the process. All important flows are labeled. Other information such as Name of the Author and Change History is also maintained in the graphs.

Each application is supported by a System Handbook which describes in detail about the application design and key the graph components.

Parameterization

One of the main advantages of using EME is 'Parameterization'. The ABI environment uses parameters to organize project and environment structure. Parameters are used to control the behavior of graphs and projects in a controlled and uniform way. BitWise development team strictly sticks to the standardized naming format for each of the parameters. The parameters are very efficiently divided as Team Level parameters, Project level parameters and Graph level parameters. The parameters that are common to all the process under one project are defined under Project level parameters. Thus it avoids re-defining of same parameter in the process. Adopting such kind of hierarchy makes the applications more maintainable. The standards followed also helps making graphs portable across multiple platforms and environment. The Application can be deployed with minimum coding changes to different operating system or any environment (Development / Test / production).

In the 2.14 and higher versions of the product, BitWise has embraced and immersed into the Concept and practical uses of psets (parameter sets). Elimination of code redundancy and re-use of the code is the prime pro of a PSET. PSET's are used within the EM>E for dependency analysis and data lineage, within job deployment to give a job name, and within the Standard>Environment to support the environmental features of graphs.



Scripting

Ab Initio and Unix scripting are tightly coupled. Scripting complements the Ab Initio environment by performing any pre-requisite tasks for the graph. These tasks could include:

- » Setting up the environment variables with appropriate values Effective use of environment variables for File Assignments, Date/Time Assignments, Script Based Assignment and Constant Value Assignment.
- » Providing information about the execution region (Development/Staging/Production) so variables can be populated and used accordingly.
- » Performing clean-up and catalog activities using wrapper scripts.
- » Dynamic Script Generation is the latest in Ab Initio world and one of its finest features, it has enabled the use of Ab Initio PDL (Parameter Definition Language) and Component Folding.

Performance Tuning and Reliability

Since the ETL applications typically handle large volumes of data, special attention is paid to performance tuning.

Some of the techniques used include:

- » Row Partitioning for parallel processing of the data
- » Sort-Dedup, Filtering and Column partitioning to process only required data
- » Checkpoints to provide re-run ability and reliability
- » Phases to ensure optimum memory and CPU available for job execution.

Reusability

Reusability is an essence of all application design and architecture. Ab Initio provides reusability in form of sub-graphs and custom components to carry out common tasks across the job stream.

XFRs in Ab Initio provide a kind of extension to reusability by containing common transformation logic at one place to be used by other graphs wherever required.

Code reusability was introduced and accomplished through Generic Load graph by Bitwise offshore Team. This graph loads source file to Teradata.

The advantages of the generic graphs developed are:

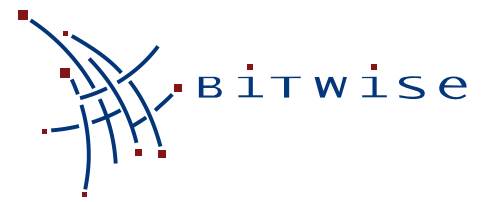
- » Reduction in development time for load process.
- » Reduction in the number of objects maintained in the repository.

Data Quality Assurance

Data quality is of extreme importance in ETL and BI applications as strategic planning and decision making is fully dependent on the summary of data processed. Any missing or malformed values can deteriorate the quality of the reports. Hence, Data Quality Assurance is of extreme importance and is covered extensively at Bitwise by providing Ab Initio based processes to ensure DATA COMPLETENESS & CORRECTNESS.

This includes graphs and application that compare the source and target data once the tables are loaded.

Data to Information - Ab Initio Capabilities



Reports are generated to track any variance between source and target data.

In an effort to automate, Bitwise offshore team has developed Generic Comparison Graph to perform data validation between the source and target.

This graph generates two reports:

- » A field level comparison report of the two datasets specified giving the match percentage and whether each field has passed the tolerance percentage specified
- » A report of all the distinct values present in the datasets for each field specified for comparison.

Quality Assurance Process

The primary objective of BitWise is to deliver a zero defect delivery to its customers. To meet this objective the Ab Initio CoE has devised a comprehensive Validation checklist which helps the development team to validate the Ab Initio graphs or any start/end scripts making the development process efficient.

This checklist in itself is quite exhaustive, containing checkpoints for Environment settings, graph standards, parameters, and Mainframe and Table I/O components, transform components and various scripts present in a sandbox. The Ab Initio CoE has also developed a component to Automate code review process. This component validates the syntax in the graphs and scripts adding more efficiencies to the already efficient development process.

EME for Source Code Management

The most influential feature of EME (Enterprise Metadata Environment) i.e. “Dependency Analysis” is followed for In-Depth analysis of time critical applications. The storage of graph related Metadata allows for data impact analysis to occur, giving the user a visual sense of how the data is changing in a graph, and the impacts those changes have on another graph.

Rudimentary EME features such as Version Control, Statistical Analysis, Code Sharing, Code Re-Usability, Data Lineage, Metadata Management and Code Promotion are productively trailed.

For all applications, the standard directory structure is followed. Thus, the locations for storing the layouts, transformations, SQLs, etc. are fixed. This makes the applications more maintainable and it becomes easier for new developers to understand the application.

BitWise has developed strict coding conventions and guidelines based on its experience working with Ab Initio and these are followed throughout all the projects undertaken. Three important Metadata goals namely DEFINE, BUILD & ADMINISTER and NAVIGATE are primarily achieved throughout all the projects undertaken.



CONTACT INFORMATION

BitWise Inc.
1515 Woodfield Rd.
Suite 930
Schaumburg, IL 60173

Phone : 847-969-1544
Fax : 847-969-1500
Email : info@bitwiseusa.com

BitWise Australia Pty Ltd.
Level 39, 2 Park Street
Sydney,
NSW 2000

Phone : 61 2 9004 7887
Fax : 1300 790 860
Email : info@bitwiseaustralia.com

BitWise Solutions Pvt. Ltd.
BitWise World
Off Int'l Convention Centre
Senapati Bapat Road
Pune - 411016 - INDIA

Phone : 91 20 40102000
Fax : 91 20 40102010
Email : info@bitwiseindia.com