Automated Application Cartography for Data Lineage



Mapador International Inc.

As enterprise portfolios are becoming more and more complex, understanding the underlying applications, and especially how these applications interact with each other, has become extremely difficult.

To create and maintain a good level of understanding, some applications can be analyzed using existing tools. However, many of these tools have limited support or none at all. Moreover, the knowledge about inter-application dependencies is almost always captured manually - an error-prone solution with results that become outdated quickly.

The lack of reliable understanding represents a real risk to everyday maintenance and significantly adds to the costs of IT initiatives such as modernizations, portfolio consolidations and others.

Regulatory requirements place an increased demand for a solution to this problem, and potentially at even more granular levels. Auditors may require proof of understanding where the value of a particular, critical business data element originates from, and how its value is derived through potentially dozens and dozens of filters and transformations.

To develop such an understanding is a daunting task when dealing with legacy based systems, spanning across different, even home-grown technologies and architecture implementations. As a result, such regulatory compliance extremely difficult, presents an potentially very costly and time consuming challenge.

The lack of reliable understanding of today's enterprise applications represents a real risk to everyday maintenance and significantly adds to the costs of future IT initiatives.





Data Lineage

The trace output of a specific item from the original systems of records through all transformations, aggregations and calculations to its end result (usually displayed on a screen or report) is referred to as the data lineage for that item. In today's systems environments, such data linage is rarely simple. Especially for critical business data, such as data elements representing complex financial information, data lineage is usually a large dependency tree with many originating sub-paths, multilevel filters and conditions.

3" party be

0

(D)



Automated Application Cartography (AAC)

Automated Application Cartography uses sophisticated parsing engines to automatically analyze individual programming languages, accompanying scripts and operational components based on their syntax, regardless of implementing technologies. It then outputs all results into a common database and connects the various elements regardless of what technology type they originated from.

Using available relationship and impact views, as well as the accompanying graphical maps, IT personnel can quickly and reliably understand applications and portfolios. Impacts can be correctly understood in seconds, instead of potentially requiring days of manual analysis.

Just as a traditional street map helps you navigate from point A to point B, AAC allows you to understand impacts within and across your applications. A component within the application (e.g. a program, paragraph, database table, etc.) is connected to other components via the programming statements and constructs used. This is similar to how a given location of a traditional map is connected to another location via streets and avenues. And just as a traditional map provides higher level contexts (e.g. city, province, country, world), AAC organizes lower level components into increasingly higher level groups (e.g. subsystems, systems, portfolios).

Best of all, once implemented, the underlying repository can be continuously kept up to date by synchronizing its contents with production updates, thus providing an ongoing, correct reflection of the entire application portfolio.



Mapador's Automated Application Cartography has proven to provide **25-30%** of savings of maintenance budgets, year over year.



Deep Parsing

Automated Application Cartography provides the necessary base for tracing a given data element across multiple systems with the highest degree of confidence and agility. It records the transformations a data element is exposed to, as well as understands the flow and order of all processes where such transformations occur.

A Deep Parsing engine is used to accomplish all such tracing. This engine allows for specific targets to be defined for tracing and queries Mapador's own repository to find the given data lineage.

The 'Target Definition' can define components directly (i.e. a specific field name), a generic name (i.e. fields with certain characters in their name), a generic type (i.e. fields with a certain definition) or a combination of these.

However, by relying on a language's actual syntax for parsing, Mapador does not depend on naming conventions or any given variable name. These are only used as a starting point for impact analysis. For example, one could define to find all impacts of a field named 'Customer-Number'. This would be used as input to the analysis only – after that, regardless of how the fields are named in the impact chain, Mapador will find all impacts.





As described above, Deep Parsing can also be initiated by defining a certain field definition as the target. In this case variable naming is completely irrelevant.

Deep Parsing outputs a Target Impact List, comprising of all components matching the Target Definition as well as all their affected components, organized by technology type. This information is saved in the Repository, for further processing.

Deep Parsing is a recursive process. From the target list, components are automatically followed through within the technology and all impacts are output into the repository. Impacts connected to other technologies are then fed back into the parsing process to identify impacts in those other technologies. For example, if a field is followed through within COBOL all the way to an output data store (such as a file or database table) the recursive process then investigates any other technologies that may input the affected data store, such as sort or other utilities, database loads or other system components written in other languages.

This is done by a feedback-loop into the affected other technology's deep parsing process, with the new target as input. The process continues to execute until all impacts in all technology types are exhausted.



Data Lineage Output

Data lineage output should caters to the various types of audience - including reports organized for Business Analysts, Technical Architects or digital output in XML format for further input to other tools (e.g., data governance suites such as IBM's Infosphere) or other custom formats.



Benefits of Automated Application Cartography

The Code Never Lies

Automatic Application Cartography with Deep Parsing reports Data Lineage directly from production code. As such, it mines processes and transformations as they actually occur in real operational environments.

Automated and Repeatable

As enterprises endeavor to address data lineage issues, one possible approach is to manually document data flow through the systems. For a relatively smaller set of applications, this can potentially be accomplished in the allocated timeframes, even if expensively.

However, manual analysis is error prone and not cost-effectively repeatable. In fact, this is exactly why Data Flow Diagrams developed sometimes decades ago were never really kept up to date.

Without keeping the results up to date, even high quality manual analysis quickly becomes outdated. As a result, confidence in the correctness of manually obtained data lineages cannot be as high as desired.

Data Lineage obtained from Automated Application Cartography and Deep Parsing directly overcomes these problems.

Ready for Growth

Data governance is already formally being addressed by many organizations. Automated Application Cartography bridges data governance gaps by enriching the data collected into data governance frameworks through automation.

Implementing Individual data lineage requirements will also likely grow over the next few years due to increasing regulatory requirements. Implementing Mapador's repeatable process allows for the inclusion of such future growth in a decreasingly costly manner.



About Mapador:

Mapador's technology stems from its success in Y2K, providing turn-key solutions and covering a wide variety of technologies with its unique business model of combining superior technology with human intelligence and expertise.

Mapador's technology has been installed in a wide variety of industries, including the Canadian federal and provincial governments, the finance, insurance, mining, educational verticals and others.

Mapador is based out of Toronto, Canada with offices in the United States and Latin America.

www.mapador.com