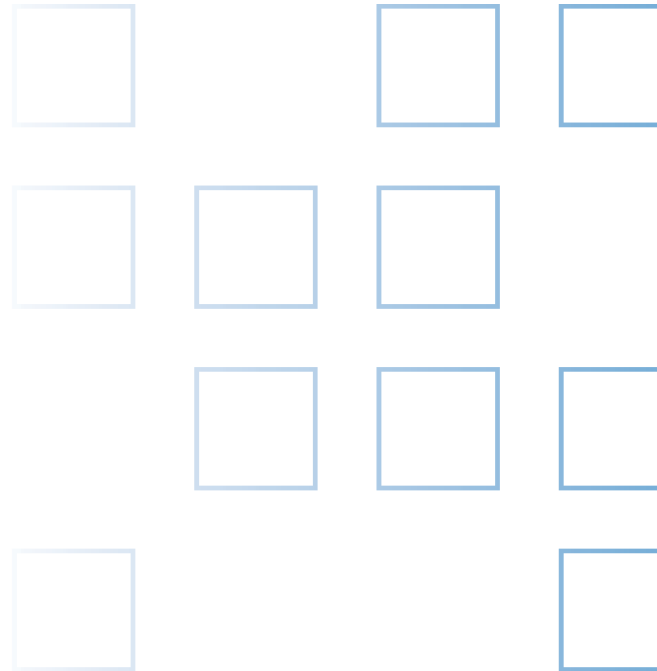




Profiler Plus: Architecture Choices





Introduction

This white paper reviews the benefits and trade-offs of the Profiler Plus architecture, and alternatives.

The Task of Analyzing Data

Generally speaking, data profiling is the task of examining a set of data to identify characteristics, and empowering an analyst to work with a much greater degree of efficiency. When starting to analyze a set of data, it is common that people “don’t know what they don’t know”. Analysis tends to be iterative and exploratory – a data issue is discovered, which points to another data issue, and so forth.

In this environment, we want analysts to be as productive as possible, and to have to think about using the tool as little as possible. We want them to focus on the data itself and to navigate freely through the data. Some users may be business oriented, so it is desirable to minimize the technical skills necessary, such as how to write queries. This becomes particularly important with a larger number of users – support becomes a greater issue.

Given this, we take the approach of generating a full set of results up front to ensure that users have immediate access to all the information that they need. Adjusting profiling parameters up front reduces workflow efficiency if users need to go back and re-profile items that were not fully processed in the original profile. Partial profiling forces users to slip into iterative cycles, which is what we are trying to avoid.

Technology Approaches

There are two broad approaches for data profiling architectures:

- Extraction – The data is extracted from an existing source, be it database or flat file, and processed separately.
- In situ – The data is left in the source database or file and processed on the fly. Sets of results are stored in a repository.

It is worth noting is that these architectures are not mutually exclusive – it is relatively easy to add “in situ” profiling as extra functionality to an extraction architecture. This involves using the existing data access infrastructure to perform data profiling processing against source systems. The results of this profiling are stored within the metadata repository. Drilling down to values require access to the source systems.

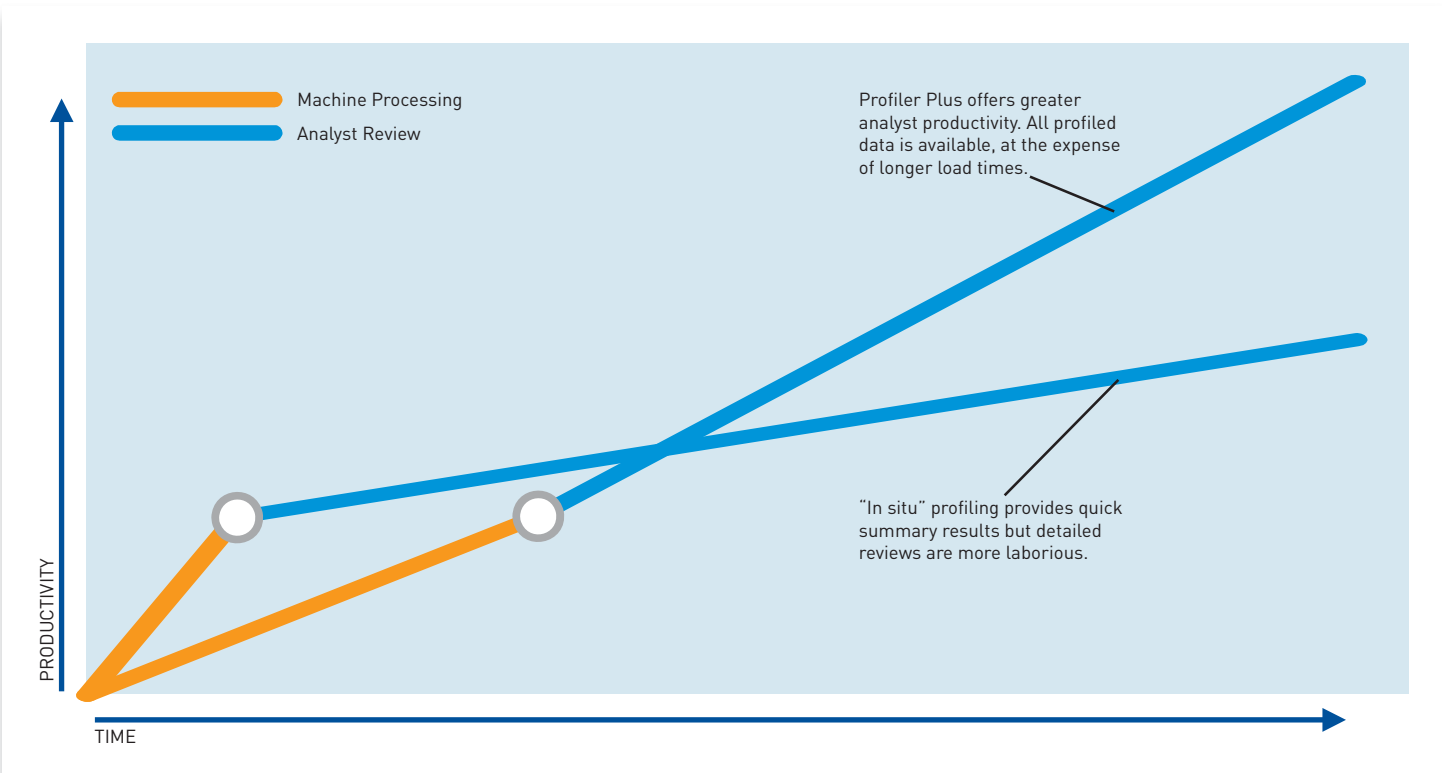
The Profiler Plus uses an ‘extraction’ architecture. Our experience shows that this architecture generally gives much better ongoing interactive performance to the user, at the cost of an initially longer load time. The task of data analysis is, by nature, a resource intensive activity with regards to staff time. The majority of effort is spent reviewing the results of data profiles, looking for data issues.

Maximizing the interactive performance of the profiling tool is important for both overall productivity and user experience. If users perform arbitrary drilldowns on large datasets, it is very frustrating to have to wait a long time for each set of records to be returned. In turn, this discourages users from analyzing the data as thoroughly as they may otherwise have done. In addition, it is not necessary for users to know how to write queries – they simply have to point and click to view the records.

Extraction Architecture

The extraction architecture has benefits and trade offs. The data is loaded once (or several times, if different samples are required). It is brought to a central location where it is processed and made available for analysis. The benefits are as follows:

- >> Data drill down always has fast and predictable performance, and is not determined by external factors such as indices or load on an external source. Our previous experience shows that ad-hoc drill downs into generic databases have poor performance. This causes inefficiencies, particularly with larger groups of users, as they have to wait for queries to respond, or alternatively, ask (DBAs) to tune databases for particular analysis queries.
- >> Referential integrity (join) checking is more performant, with rich drill down capabilities, showing the records where problems exist. These checks are performed between any data sources, e.g. between a flat file and an Oracle table.



- >> The load on external databases are better managed. Data is extracted from the source databases as a one off operation, and external systems are not required for further processing, such as data drill downs or joins.
- >> Additionally, Profiler Plus always generates a complete set of profiling results for a data source. While this increases the load time, the benefits are that analysts always get a complete and consistent set of profiling results. This has particular benefits when multiple users are looking at the same profile – there need be no concerns about what options were activated to generate the profile.
- >> Having all the results for a given profile in place simplifies external reporting interfaces. It is guaranteed that all results are available and consistent. Custom queries are performed on frequency values or pattern sets, if required, along with drilldowns into the actual

data. If data profiling is performed in situ, the results, including both summarized value frequency information and the data itself, is distributed across source systems and the profiling repository, increasing the complexity of writing reports.

“In situ” Architecture

“In situ” architecture leaves the data in place, either in its relational database or flat file. Subsets of frequency value pairs may be retrieved from the database.

- >> Easier to implement, by utilizing external technology.
- >> Good for rapid summarization of external relational sources – the profile engine fires off queries to get back summary metadata. No data extraction is required.



- >> External databases are not usually optimized for the random queries a data analyst performs when drilling down to records.
- >> Data drill downs into large files take a particularly long time to accomplish, reducing analyst productivity.
- >> Testing referential integrity between heterogeneous data sources is more difficult. Drilling back to data within those sources, based upon join results, is difficult to achieve.
- >> Puts additional load on existing database systems, as drill downs require ad hoc queries to be performed.

For more information about our products and services, please log onto our website at www.g1.com or call us today at 888-413-6763.