Functional Practice Statements - Data Profiling

Level 1: Initial

1.1 Basic profiling is performed for a data store(s).

Basic profiling includes such things as analyzing the types or number of distinct values in a column, number or percent of zero, blank or null values, string length, date ranges, patterns, as opposed to the more advanced analysis such as cardinality, frequency distributions, or key integrity

Example Work Products
- Data profiling reports
- List of data profiling checks

Level 2: Managed

2.1A data profiling methodology is established and followed.

The methodology adopted or created by the organization describes the approach to data profiling. The methodology will typically address planning and scoping the effort, profiling techniques, report templates, and presentation formats for summary results. In addition, profiling processes should be reusable and leveraged across multiple data stores and shared data repositories.

2.2 Data profiling plans are established for projects

Components typically included in the plan:

- Selection of the data store(s) to examine
- Identification of the data set(s) to profile
- List of stakeholders and definition of their involvement
- Objectives of the profiling activity
- Data quality criteria based on the objectives, which includes referential integrity (parent and child) of the data, consistency of the data with respect to its documented metadata, consistency with established rules and patterns, and standard data quality dimensions
- Rules to be applied during the profiling activity
- Method(s) and tool(s) for data profiling
- Template(s) for documenting results
- Schedule of activities, including resources

2.3 Plans for profiling a data store are shared with relevant stakeholders and data governance.

Data profiling activities should not be planned or executed in a vacuum.

Stakeholders and data governance authorities may have specific needs that should be taken into consideration. In addition, recognizing that the expense and effort to conduct profiling activities is not trivial, it is important for the profiling activities to be aligned with business needs. Sharing plans for profiling can help to ensure agreements and continued alignment.

2.4 Data profiling activities are conducted according to the plan, and efforts are adjusted when significant deviations from plan are detected.

Because a profiling effort often produces some unexpected results, the organization needs to be flexible enough to determine, during the profiling initiative, if initial results justify additional time and effort to expand the scope.

2.5 Data profiling results and recommendations are reported to the stakeholders.

Results should be used as input to data quality assessment and data cleansing efforts, as well as to inform the data quality strategy. For example, an organization may determine that its product data has an unacceptable percentage of errors. The stakeholders need to weigh in on the impact of these errors and have a role in determining the remediation alternatives and approach.

Example Work Products
- Data profiling methodology documentation
- Approved data profiling plan and schedule
- Data profiling findings reports and metrics
- Defined skill set and training plan for staff with data quality responsibilities

Level 3: Defined

3.1 Data profiling methodologies, processes, practices, tools, and resultstemplates have been defined and standardized.

Standard profiling tools should be identified and consistently used across the organization to gain efficiencies. An organizational standard method for analyzing and presenting business and technical impacts of data profiling on remediation activities should also be defined and followed. Report templates and metrics are standardized, centrally stored, and published to ensure consistent application across the organization.

3.2 All techniques identified to meet the profiling objectives are performed.

While data profiling can be considered primarily a discovery activity, it is typically initiated to meet specified objectives. The profiling techniques and tools employed must support achieving those objectives. Tailoring of techniques and templates may be required. Often a profiling effort has several phases; for example, out of the box profiling checks (e.g., value ranges, ID uniqueness, etc.); standardization analysis (e.g., addresses, syntax); tests for selected business rules; and known issues (which may require complex queries).

The data profiling team should be fully conversant in all techniques and corresponding tool capabilities selected for the profiling activities.

3.3 Traceability between data requirements, documented metadata, the physical data, and data quality rules is captured and maintained.

Data profiling activities should be executed by data profiling experts aware of data requirements, data quality rules, data content, and data structures.

This is achieved through establishing traceability between data requirements, physical data, and metadata.

3.4 Data governance is engaged to identify core shared data sets and the corresponding data stores that should be regularly profiled and monitored.

The organization should have defined rules for when various data sets are profiled (e.g., when data is acquired; or prior to being consolidated, migrated, exported, analyzed, reported for compliance purposes, or structurally transformed).

3.5 Profiling processes are reusable and deployed across multiple data stores and shared data repositories.

Sharing and mentoring among peers to build data profiling best practices should occur across the organization, because data quality activities are often highly valuable but resource-intensive. Therefore, it is desirable to leverage efficiencies and avoid rework. Some organizations find that the most effective approach is to operate with a single data profiling team with skilled

3.6 The SDLC includes data profiling tasks with tailoring criteria, guidance, and governance.

Most data store development efforts (for example, creation of a new data warehouse) should include data profiling activities as a planned part of the project. Institutionalization of data profiling practices requires that the SDLC include reference and guidelines for these activities, and that tailoring criteria are defined and followed.

Example Work Products
- Data profiling standards, including criteria for processes, standards, best practice criteria, tailoring, and reporting formats
- Data profiling methodologies tailored from the organizational standard
- Report showing traceability of data requirements with the data content and characteristics revealed through profiling results
- Documented tailoring of data-related decisions and rationale
- Documentation that practitioners have required profiling skills
- Data profiling metrics
- Recommendations reports from data profiling efforts
- Business and technical impact analysis results template
- Standard data profiling report requirements
- Approved standard data profiling tool(s)
- Data profile baselines

Level 4: Measured

4.1 Performance of data profiling processes is measured and used to manage activities across the organization.

- Plans and schedules for data profiling should be managed according to the feedback provided by data quality measurements. Measurement should indicate how well the output of this activity addresses and aligns with the business need and priorities. Decisions on what, when, and how to profile data should be driven by indications of quality and criticality, which may vary by business application. Highly shared data and data sets deemed vital to key business processes should be regularly profiled, as data quality is critical and needs to be frequently monitored.
- Data quality measures should also indicate how well the staff performed the data profiling activities. The evaluation of plans and execution (actual vs. estimates) should consider such things as use of techniques, impact of results and decisions, compliance with methods and standards, quality of output, and level of effort.
- Data profiling process performance baselines can be created and used to inform the planning and execution of data profiling activities and results.

4.2 Data profiling efforts include evaluation of the conformity of data content with its approved metadata and standards.

Approved standard business terms, meanings, values, and ranges are used as a benchmark for profiling the data content in a data store. Additional documentation is typically found in corporate data dictionaries, data models, system requirements documentation, etc. It is a best practice to update this documentation as needed.

4.3 During a data profiling activity, actual issues are compared to the statistically predicted issues based on historical profiling results.

Results should be systematically compared to corresponding historical profiles to evaluate impact of profiling activities on corrective actions and quality improvements.

4.4 Results are centrally stored, systematically monitored, and analyzed with respect to statistics and metrics to provide insight to data quality improvements over time.

A consistent impact analysis method can be applied to evaluate business, technical, and cost impacts of remediation. Summary profiling results are provided to data governance bodies and senior management. Results are used to inform data governance and data architecture decisions, especially for highly shared data.

Example Work Products
- Documented profiling methodology, best practices, and standards
- Project reports showing application of profiling results to data quality governance
- Dashboards, scorecards, or other decision support tools for data quality, showing the results of data profiling efforts
- Data quality portal displaying data quality models and results to be used for performance baselines

Level 5: Optimized

5.1 The organization addresses root causes of defects and other issues based on an understanding of the meaning, technical characteristics, and behavior of the data over time.

5.2 Data profiling processes and other activities are analyzed to identify defects and make improvements based on the quantified expected benefits, estimated costs, and business objectives.

Data profiling results performed on the same data over time can be statistically analyzed periodically to measure the performance of profiling activities.

5.3 Real-time or near-real-time automated profiling reports are created for all critical data feeds and repositories.

Automating the performance and scheduling of profiling improves the data quality program's efficiency and responsiveness to planned and unplanned events.

Example Work Products
- Log of stakeholders' usage of profiling results
- Control charts demonstrating that the processes used across data stores have stabilized (data stores have been sufficiently profiled)
- Data profiling process objectives for improvement included in standard data management strategies, programs, and reports
- Real-time data profiling reports generated on schedule
- Conclusions drawn from data profiling process analyses and recommendations for improvement