

A Medallion-Based Lakehouse Architecture and ISO/IEC 25012 Evaluation Framework for Government Open Data¹

วิวัรรณ์ ทองโชติฉัตร (Vivat Thongchotchat)²

วันที่รับบทความ: 1 มิถุนายน 2529

วันแก้ไขบทความ: 25 มิถุนายน 2529

วันตอบรับบทความ: 28 มิถุนายน 2529

Abstract

This study is positioned as Design Science Research (DSR) and a pre-production design study that produces two artifacts for government open data: a medallion-based lakehouse architecture and an ISO/IEC 25012-based evaluation framework. The population statistics dataset from the Department of Provincial Administration (DOPA), Ministry of Interior (MOI), Thailand, is used as a case dataset. The architecture organizes data into Landing, Bronze, Silver, Gold, and Audit layers to separate raw file preservation, source-faithful structuring, standardization, analytical modeling, and audit evidence. Open-source components, including MinIO, Apache Iceberg, Hive Metastore, PostgreSQL, Trino, Airflow, and OpenMetadata, are used to specify a reproducible lakehouse environment. The results present planned storage objects, table schemas, metadata, and quality measurement structures for transforming a headerless source file into structured and traceable lakehouse tables. ISO/IEC 25012 is used to define measurable indicators for completeness, consistency, credibility, currentness, traceability, understandability, accessibility, and compliance across lakehouse layers. Rather than reporting production deployment or empirical performance results, the study contributes an architectural and methodological foundation for future implementation and validation of transparent government open-data lakehouse pipelines.

Keyword: Medallion-Based Lakehouse Architecture, ISO/IEC 25012 Evaluation Framework, Government Open Data

Introduction

Open government is a governance approach that promotes transparency, accountability, public participation, and responsiveness in the public sector. This agenda is advanced by the Open Government Partnership (OGP), a global initiative that encourages collaboration between governments and civil society to strengthen open and participatory governance. The Open Government Partnership (n.d.) emphasizes commitments to increasing access to information about government activities, publishing data in formats that can be easily located, understood, used, and reused by the public, supporting civic participation, strengthening public integrity, and applying technology to promote openness and accountability. Thailand's open-data development reflects a similar policy direction. The Open Data Institute (2024) reported that the Government of Thailand and the Digital Government Development Agency (DGA) are committed to realizing the benefits of open data to empower Thailand's digital government and promote open data adoption and use across government and wider society.

¹บทความจากฝ่ายบริการวิเคราะห์ข้อมูล สถาบันข้อมูลขนาดใหญ่ (องค์การมหาชน) ; Article from the Data Analytics Services Division, Big Data Institute (Public Organization)

²นักวิทยาศาสตร์ข้อมูลอาวุโส ฝ่ายบริการวิเคราะห์ข้อมูล สถาบันข้อมูลขนาดใหญ่ (องค์การมหาชน) ; Senior Data Scientist, Data Analytics Services Division, Big Data Institute (Public Organization) E-mail: napha17@uru.ac.th

Thailand's participation in OGP further demonstrates this commitment, as the Open Government Partnership (2026) identifies Thailand as a member country, while the Office of the Prime Minister (2025) states the government's intention to promote open, transparent, and accountable public administration, encourage inclusive participation across sectors, and develop a National Action Plan under the OGP framework. As part of this open-data ecosystem, the Digital Government Development Agency (2021) provides Open Government Data services, including Data.go.th, as a central platform for publishing government datasets and supporting dataset metadata guidance.

Although open government data is expected to strengthen transparency and public value, prior research shows that data publication alone is insufficient to ensure meaningful reuse. Users of open government data may encounter practical barriers related to data quality, metadata quality, accessibility, usability, documentation, and the ability to interpret and transform datasets for specific analytical purposes. In a systematic literature review of empirical open government data research, Wirtz et al. (2022) found that governments and organizations continue to face challenges in realizing the value of open government data despite its increasing availability. Similar concerns are reflected in research on Thailand Open Government Data, where Kattiyapornpong et al. (2017) proposed a quality assessment framework based on characteristics such as traceability, currentness, completeness, compliance, understandability, accuracy, and consistency, indicating that data quality is a fundamental requirement for useful open-data utilization. The Open Data Institute (2024) also reported that although Thailand has made progress in promoting open data through policy commitment, national infrastructure, guidance, and awareness-building activities, challenges remain in expanding the publication and use of open data across government and wider society. These studies suggest that open government data must be prepared not only for publication, but also for practical consumption. Datasets with difficult formats, unclear schemas, weak metadata, or poor data quality may remain publicly available while still being difficult for users to transform into reliable analytical outputs.

To address these limitations, this study proposes a medallion-based lakehouse architecture supported by metadata governance design considerations. A medallion architecture organizes data into progressive layers, commonly referred to as bronze, silver, and gold, where raw data are preserved before being incrementally validated, standardized, enriched, and curated for analytical use. This design pattern improves data structure and quality across layers, with the gold layer representing curated data optimized for analytics and consumption (Databricks, 2026; Microsoft, 2026). Recent lakehouse research also emphasizes that the lakehouse has emerged as an architectural approach that combines data lake flexibility with data warehousing capabilities. For example, Mazumdar et al. (2023) described the lakehouse as an architecture that supports data warehousing practices on open data platforms and provides capabilities beyond conventional warehouse-centered approaches. However, their discussion focuses on general lakehouse capabilities rather than government open-data pipelines, ISO/IEC 25012-based quality interpretation, or audit evidence. This gap indicates the need for a design that connects lakehouse architecture with explicit data quality assessment and governance requirements in government open-data environments.

The effectiveness of a lakehouse pipeline, however, depends not only on transformation processes but also on the management of metadata, schema, lineage, and data quality throughout the data lifecycle. Prior studies on data lakes emphasize that metadata management is essential because data lakes store heterogeneous data with limited predefined structure; without effective metadata systems, users may face difficulties in discovering, understanding, and reusing data (Sawadogo et al., 2019; Sawadogo & Darmont, 2021). In this study, metadata governance is positioned as a design component for improving transparency, traceability, and reliability by defining how data definitions, schema information, quality indicators, and data movement from raw sources to analytical outputs should be documented and monitored.

This study is positioned as Design Science Research (DSR) and a pre-production design study that produces two related artifacts: a medallion-based lakehouse architecture and an ISO/IEC 25012-based evaluation framework for government open data. Rather than reporting production deployment or experimental performance results, the study establishes the architectural foundation and assessment criteria for future full-scale development. ISO/IEC 25012 defines a data quality model consisting of fifteen characteristics from inherent and system-dependent perspectives and can be used to define and evaluate data quality requirements in data acquisition, integration, assessment, and improvement processes (International Organization for Standardization, 2008). The proposed architecture and evaluation basis therefore clarify how raw government open data can be structured, governed, traced, and measured across lakehouse layers, providing a methodological foundation for transparent, traceable, and analytically reliable implementation.

Objectives

1. To design a medallion-based lakehouse architecture for transforming government open data into analysis-ready datasets.
2. To develop evaluation metrics for assessing the quality and effectiveness of a medallion-based lakehouse architecture and data pipeline.

Conceptual Framework

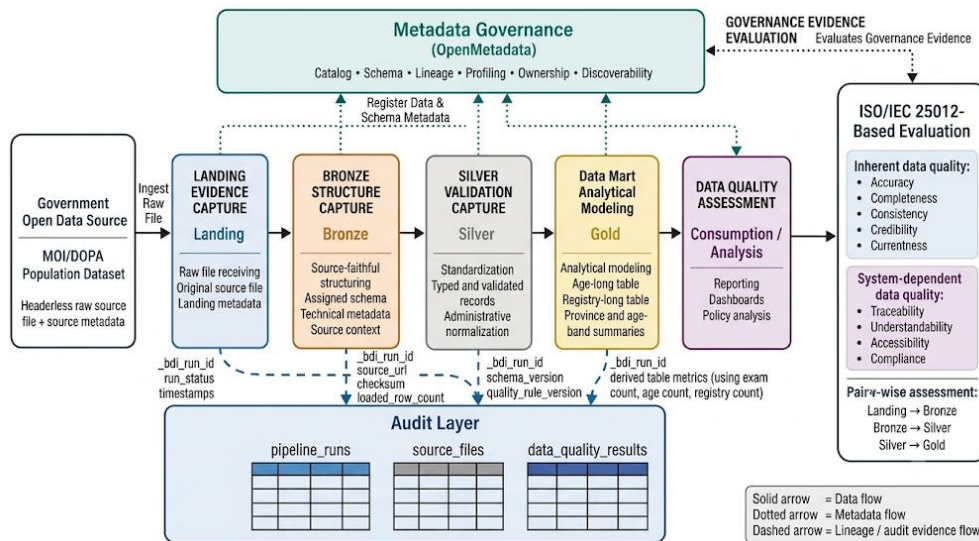


Figure 1 Proposed Integrated data, metadata, and lineage flow in the proposed medallion-based lakehouse architecture

Methodology

This study was conducted as a pre-production design study for developing a medallion-based lakehouse architecture for government open data. The methodological focus is on proposing the architectural structure, defining the role of each lakehouse component, and designing the pipeline foundation required before full-scale implementation and operational deployment. The study establishes a technical and methodological basis for future development by configuring a reproducible lakehouse environment, applying it to a selected government dataset, and identifying how the architecture can support ingestion, storage, metadata governance, auditability, and analytical preparation.

1. Case Dataset and Government Data Context

The case dataset was obtained from the statistical download service of the Department of Provincial Administration (DOPA), Ministry of Interior (MOI), Thailand. It provides province-level population statistics for B.E. 2568, separated by single-year age and sex, together with registry-related population categories. This dataset was selected because population statistics are fundamental government data that can support demographic analysis, service planning, budget allocation, welfare design, health and education policy, and ratio-based indicators used by government agencies.

The dataset is also appropriate for demonstrating the proposed lakehouse design because it has high public value but low analytical readiness. The source file is a pipe-delimited text file without a header row, as illustrated in Figure 2. The first field represents the administrative area description, while the remaining numerical fields require assigned column names, including male and female population counts from age 0 to age 101, registry-related fields, and total population fields. Reporting year, month, province code, province name, source period, and source-file identity must also be derived from pipeline parameters, file naming, reference data, and ingestion metadata.

These characteristics make the dataset difficult to consume directly. Users must reconstruct schema meaning, interpret administrative context, validate reporting information, and reshape age and sex columns before analysis. Therefore, the dataset is suitable for showing how Landing preserves the original file, Bronze assigns a source-faithful schema and metadata, Silver standardizes records, and Gold restructures the data into consumption-ready tables, functioning like data marts for age, registry, province, and age-band analysis.

```
กรุงเทพมหานคร|11640|10838|14323|13274|15588|14635|15814|14567|17733|16699|19301|18173|20543|19576|21443|20518|22556|21836|23165|22067|24015|23203|25698|24507|
|26114|24831|28228|27405|27006|26115|26447|25571|27320|26680|27955|27265|29191|28372|29170|28288|29325|29029|33610|29951|33444|29477|30955|29701|30107|30037| | | | | | | |
|33420|33217|31371|31822|33470|33724|38112|39203|39328|40743|40360|42174|38366|40214|38140|40280|38073|40981|37749|40649|37176|40617|35551|38814|35707|39567|
|33557|37417|34984|39456|36669|41829|37192|42556|38201|44399|38943|45947|40039|46494|40993|47878|41111|47591|38893|46393|39468|46697|39220|46574|37538|44796|
|37109|44615|35947|43468|37168|45023|37910|46265|37003|45174|35859|44445|36648|45623|35386|43656|33672|42939|34340|43957|33967|44027|32189|42569|31046|41434|
|29564|39875|29527|40656|27988|38755|26012|35907|24683|34740|23992|33589|21848|31682|21598|31612|19915|29085|18576|27586|17442|26538|16364|24496|15163|23153|
|13123|20250|11517|17682|9967|15718|8900|14191|8216|13294|7503|12180|6503|11021|6219|10905|5028|8436|5161|9086|4527|8000|3969|6941|3453|6183|2792|5129|2382|
|4293|1897|3434|1619|3060|1295|2344|1028|1812|814|1458|671|1148|582|912|510|719|377|482|2332|2355|0|0|0|41512|32052|73564|37804|28545|66349|1333|1150|2483|
2526272|2896296|5422568|
```

Figure 2 Example first line of the headerless DOPA population source file

2. Medallion-Based Lakehouse Architecture Design

The lakehouse architecture was designed using a medallion-based structure because it separates raw data preservation, standardization, analytical modeling, and audit evidence. The proposed architecture consists of Landing, Bronze, Silver, Gold, and Audit areas.

The Landing area receives the original source file during each pipeline run and preserves it in its raw format for validation and ingestion. The Bronze layer is the first durable lakehouse layer. It stores source-faithful data with minimal technical changes by assigning column names, attaching source context, and adding technical metadata. The Silver layer is designed as the standardized layer, where data type handling, validation, standardization, and quality preparation are applied while preserving the source-aligned grain. The Gold layer is designed as the analytical consumption layer, where schema redesign, reshaping, derived attributes, aggregation, and dashboard-ready tables can be produced. The Audit layer stores operational and measurement evidence, including pipeline runs, source-file metadata, checksums, row counts, schema evidence, and quality measurement results.

The logical schema structure of the lakehouse was designed as follows:

- lakehouse.bronze
- lakehouse.silver
- lakehouse.gold
- lakehouse.audit

For the DOPA population dataset, the design includes a Bronze raw table, a Silver standardized table, Gold analytical tables, and Audit tables for pipeline runs, source files, and quality measurement evidence.

3. Pre-Production Lakehouse Infrastructure Design and Configuration

The proposed lakehouse was configured as a reproducible Docker-based infrastructure using open-source components. Docker Compose was used to define and run the main services required for storage, catalog management, query access, orchestration, and governance.

MinIO was used as the S3-compatible object storage layer. Apache Iceberg was selected as the lakehouse table format, with Parquet as the underlying columnar file format. Hive Metastore was used as the catalog backend for Iceberg, with PostgreSQL serving as the catalog metadata database. Trino was used as the SQL query engine for creating schemas, writing data, and querying Iceberg tables. Apache Airflow was used for workflow orchestration. OpenMetadata was adopted as the metadata governance and discovery layer.

The infrastructure separates responsibilities across components. MinIO stores the physical lakehouse files. Iceberg manages table structure, schema evolution, snapshots, and versioning. Hive Metastore and PostgreSQL manage catalog metadata. Trino provides SQL access to the lakehouse. Airflow executes ingestion, validation, loading, and audit tasks. OpenMetadata provides metadata discovery, lineage visibility, profiling, ownership, and governance representation.

A shared workspace was mounted into the Airflow services to store DAGs, logs, plugins, landing files, temporary files, and reference data. For the MOI/DOPA population use case, the reference directory stores province code and province name data, allowing the pipeline to preserve government administrative context. The

landing directory was organized using a partition-style path based on dataset domain, source agency, year, month, and province code.

4. Landing-to-Bronze Pipeline Design

The Landing-to-Bronze pipeline was designed as the first stage of the lakehouse flow for the MOI/DOPA population dataset. The pipeline downloads the source file, validates its structure, records landing metadata, assigns source context, and loads the data into a durable Bronze Iceberg table.

The Landing phase constructs the run context, including reporting year, reporting month, province code, source URL, source file name, dataset identity, and agency context. It then loads the province reference file, prepares the landing directory, downloads the source file, validates the file structure, and generates landing metadata. Since the source file does not provide complete metadata, contextual information is derived from the source URL, file name, pipeline parameters, and reference data.

The Bronze phase creates the target Iceberg table if it does not already exist:

```
lakehouse.bronze.moi_dopa_population_raw
```

The pipeline reads the landing metadata and raw file, assigns source column names, attaches source context, adds technical metadata columns, writes temporary Parquet data, inserts the records into Iceberg through Trino, validates the Bronze load, and records audit evidence. The Bronze table contains technical metadata columns, source context columns, and assigned source data columns. This allows the data to remain source-faithful while becoming queryable, interpretable, and traceable.

5. Bronze-to-Silver Pipeline Design

The Bronze-to-Silver pipeline is designed to transform source-faithful Bronze data into a standardized Silver dataset. The proposed Silver table is:

```
lakehouse.silver.moi_dopa_population_standardized
```

This stage preserves the source-aligned grain while improving structure, consistency, and interpretability. The pipeline verifies that Bronze records are associated with valid run identifiers, source checksums, reporting periods, and province reference information. It then applies schema standardization, data type handling, administrative identifier normalization, and validation support.

The Silver layer is designed to check for missing values, invalid codes, duplicated records, inconsistent row counts, unexpected schema changes, and values outside expected ranges. Quality issues are recorded as audit evidence rather than silently modifying the source meaning. Each Silver record remains traceable to the original source file and pipeline run through inherited technical metadata.

6. Silver-to-Gold Pipeline Design

The Silver-to-Gold pipeline is designed to transform standardized population data into curated analytical datasets. The Gold layer supports reporting, aggregation, dashboarding, statistical analysis, and public-sector data products. Unlike the Bronze and Silver layers, the Gold layer may redesign the schema to improve analytical usability.

For the MOI/DOPA population use case, the proposed Gold tables are:

- lakehouse.gold.moi_dopa_population_age_long
- lakehouse.gold.moi_dopa_population_registry_long
- lakehouse.gold.moi_dopa_population_by_province
- lakehouse.gold.moi_dopa_population_by_age_band

The age-long table converts age-related columns into long format. The registry-long table normalizes registry-related population categories. The province table aggregates population values by province and reporting period. The age-band table groups detailed age values into analytically meaningful age groups. Each Gold table preserves sufficient metadata to maintain lineage to the Silver and Bronze layers.

7. Auditability and Traceability Design

Auditability and traceability are designed as core functions of the lakehouse architecture. The Audit schema stores structured evidence generated throughout the pipeline, including source-file information, pipeline execution records, validation evidence, and measurement results.

The proposed Audit tables are:

- lakehouse.audit.pipeline_runs
- lakehouse.audit.source_files
- lakehouse.audit.data_quality_results

The pipeline_runs table records execution-level information such as run identifier, DAG identifier, dataset identity, pipeline version, schema version, quality rule version, execution status, and timestamps. The source_files table records source-level evidence such as source URL, file name, reporting period, province code, landing path, file size, checksum, validation status, target Bronze table, and loaded row count. The data_quality_results table is designed to store numerical measurement results by pipeline run, layer, table, column, metric, and measurement time.

Traceability is maintained through technical metadata columns and audit records. Bronze records contain run identifiers, source URLs, source file names, checksums, ingestion timestamps, and schema versions. These metadata elements are inherited or referenced by the Silver and Gold layers, allowing analytical outputs to be traced back to their source file and processing history. Iceberg snapshots and Airflow execution logs further support auditability, while the Audit schema stores pipeline-specific evidence in a structured and queryable form.

8. Metadata Governance Application

OpenMetadata was adopted as the governance and discovery layer of the proposed lakehouse architecture. The study does not develop a new metadata governance platform; instead, it applies OpenMetadata to provide visibility over lakehouse assets. OpenMetadata connects to the lakehouse through Trino and can ingest metadata about schemas, tables, columns, lineage, ownership, profiling, and quality-related evidence.

In the proposed architecture, OpenMetadata serves as the external governance interface, while the Audit schema stores reproducible evidence inside the lakehouse. This separation allows the lakehouse to preserve durable audit records while enabling users and data stewards to discover, understand, and govern the data assets.

The methodology defines a complete medallion-based lakehouse architecture for government open data, covering ingestion, durable storage, standardization, analytical modeling, auditability, and metadata governance. By preserving source context, transformation evidence, and audit records across the data lifecycle, the architecture establishes a basis for applying ISO/IEC 25012 quality dimensions to assess how raw government data can be progressively improved into analysis-ready datasets.

Result

The result of this Design Science Research study consists of two interdependent design artifacts: the proposed medallion-based lakehouse architecture and the ISO/IEC 25012-based evaluation structure embedded in the Audit layer. The architecture defines the Landing, Bronze, Silver, Gold, and Audit layers, while the evaluation structure defines how measurable quality evidence can be stored and compared across layer transitions. These artifacts are presented together because the lakehouse architecture was designed to support ISO/IEC 25012-based assessment through metadata, lineage, and audit evidence.

1. Landing Storage Structure

The Landing layer stores the original source file before it is converted into a lakehouse table. For the DOPA population dataset, the source file is stored in a partition-style path that preserves the data context:

landing/moi/dopa/population/year=2025/month=12/province_code=10/6812cc10.txt

The source file is a pipe-delimited text file without a header row. The first field contains the administrative area description, followed by numerical population and registry values. Column naming is not performed in Landing; it is assigned later in the Bronze layer using the predefined DOPA source schema.

Table 1 Stored objects

Stored object	Format	Description
6812cc10.txt	Pipe-delimited text file	Original headerless DOPA source file containing administrative area descriptions and population values.
_landing_metadata.json	JSON	Metadata describing the source, context, validation result, file path, file size, checksum, and pipeline run.

Table 2 Example Landing metadata fields

Metadata group	Description
Pipeline metadata	_bdi_run_id, _bdi_dag_id, pipeline_version, schema_version, quality_rule_version
Government context	ministry, department, dataset
Reporting context	snapshot_year, snapshot_month, source_year_buddhist, source_period
Administrative context	province_code, province_name_th, province_name_en
Source evidence	source_file, source_url, landing_path, file_size_bytes, file_checksum_sha256
Validation evidence	validation_status, validation_checks

The Landing layer therefore preserves the raw source file and records the evidence needed to reproduce and verify ingestion.

2. Bronze Table Foundation

The Bronze layer converts the headerless source file into the first durable lakehouse table while preserving the source-aligned structure. The proposed Bronze table is:

lakehouse.bronze.moi_dopa_population_raw

The Bronze schema is constructed from the expected DOPA source layout: one administrative area field, one initial registry field, paired male and female age fields from age 0 to 101, and the remaining registry and total population fields. The Bronze table also adds technical metadata and source context before the assigned source data columns.

Table 3 Bronze table foundation

Column group	Main columns	Description
Technical metadata	_bdi_run_id, _bdi_dag_id, _bdi_ingested_at_utc, _bdi_source_url, _bdi_source_file, _bdi_source_file_checksum, _bdi_schema_version	Records pipeline, ingestion, source-file, and schema evidence.
Source context	ministry, department, dataset, snapshot_year, snapshot_month, province_code, province_name_th, province_name_en, source_year_buddhist, source_period	Preserves government agency, reporting, and administrative context.
Administrative source data	desc_ccaatt	Stores the administrative area

Column group	Main columns	Description
		description from the source file.
Age population fields	male_age_0, female_age_0, ..., male_age_101, female_age_101	Stores paired male and female population values by single-year age.
Registry and total fields	male_tdob, female_tdob, total_tdob, male_xhouse, female_xhouse, total_xhouse, male_other_nat, female_other_nat, total_other_nat, male_move, female_move, total_move, male_tot, female_tot, total_tot	Stores registry-related and total population values.

This table makes the raw source data queryable and traceable without changing its source meaning.

3. Silver Table Schema

The Silver layer is designed to store standardized and quality-prepared records derived from Bronze. The proposed Silver table is:

lakehouse.silver.moi_dopa_population_standardized

Table 4 Silver table schema

Column group	Main columns	Description
Traceability metadata	_bdi_run_id, _bdi_dag_id, _bdi_source_file_checksum, _bdi_schema_version, _bdi_standardized_at_utc, _bdi_bronze_table	Maintains linkage to the Bronze table, source file, and pipeline run.
Government and reporting context	ministry, department, dataset, snapshot_year, snapshot_month, source_year_buddhist, source_period	Preserves source agency and reporting-period context.
Administrative geography	province_code, province_name_th, province_name_en, area_name_th, area_level	Standardizes geographic and administrative interpretation.
Standardized population fields	male_age_0, female_age_0, ..., male_age_101, female_age_101, registry fields, total population fields	Stores typed and standardized population values.
Validation support	row_validation_status, row_validation_notes	Records row-level validation or standardization notes.

The Silver table keeps the source-aligned grain while improving consistency, typing, naming clarity, and administrative interpretation.

4. Gold Table Schemas

The Gold layer is designed to store curated analytical tables derived from the Silver table. Unlike Bronze and Silver, Gold may reshape or aggregate data to support reporting, dashboards, demographic analysis, and policy-oriented use.

Table 5 Gold table schema

Gold table	Main columns	Analytical purpose
lakehouse.gold.moi_dopa_population_age_long	_bdi_run_id, snapshot_year, snapshot_month, source_period, province_code, province_name_th, province_name_en, area_name_th, area_level, age, sex, population	Converts age columns into long format for flexible age- and sex-based analysis.
lakehouse.gold.moi_dopa_population_registry_long	_bdi_run_id, snapshot_year, snapshot_month, source_period, province_code, province_name_th, province_name_en, area_name_th, area_level, registry_category, sex, population	Normalizes registry-related population categories for comparison across area and period.
lakehouse.gold.moi_dopa_population_by_province	_bdi_run_id, snapshot_year, snapshot_month, source_period, province_code, province_name_th, province_name_en, total_male_population, total_female_population, total_population	Provides province-level population summaries for reporting and dashboards.
lakehouse.gold.moi_dopa_population_by_age_band	_bdi_run_id, snapshot_year, snapshot_month, source_period,	Groups population by analytical age

Gold table	Main columns	Analytical purpose
	province_code, province_name_th, province_name_en, area_name_th, area_level, age_band, sex, population	bands for demographic analysis.

The Gold schemas show how standardized Silver data can be reshaped into analytical structures while preserving lineage to the source pipeline.

5. Audit Table Schemas

The Audit layer stores structured evidence generated throughout the pipeline. It supports operational monitoring, source traceability, transformation traceability, and measurement-based assessment. The proposed Audit tables are:

- lakehouse.audit.pipeline_runs
- lakehouse.audit.source_files
- lakehouse.audit.data_quality_results

Table 6 Audit table schemas

Audit table	Main columns	Description
lakehouse.audit.pipeline_runs	_bdi_run_id, _bdi_dag_id, _bdi_logical_date, ministry, department, dataset, pipeline_version, schema_version, quality_rule_version, run_status, started_at_utc, completed_at_utc, error_message	Records execution-level evidence for each pipeline run.
lakehouse.audit.source_files	_bdi_run_id, source_url, source_file, source_period, snapshot_year, snapshot_month, province_code, province_name_th, province_name_en, landing_path, file_size_bytes, file_checksum_sha256,	Records source-file evidence and Bronze loading evidence.

Audit table	Main columns	Description
	downloaded_at_utc, validation_status, bronze_table_name, bronze_loaded_row_count	
lakehouse.audit.data_quality_results	_bdi_run_id, layer_name, layer_order, schema_name, table_name, column_name, iso_characteristic, measurement_name, measurement_value, measurement_unit, numerator_value, denominator_value, expected_direction, result_status, rule_version, measured_at_utc, details_json	Stores numerical measurement evidence for later quality assessment across layers.

Table 7 Example ISO/IEC 25012-Based Data Quality Metrics and Formulas

ISO/IEC 25012 characteristic	Metric name	Formula	Expected direction	Applicable layer comparison
Completeness	data_cell_missing_ratio	Number of missing cells / Total number of cells	Decrease	Bronze–Silver
Consistency	duplicate_record_ratio	Number of duplicate records / Total records	Decrease	Bronze–Silver
Currentness	source_period_presence_flag	1 if source period is available; otherwise 0	Increase	Landing–Bronze

The Audit layer provides the evidence needed to trace data movement and measurement across the lakehouse without embedding interpretation directly into stored records.

Landing preserves the original file and metadata, Bronze assigns a durable source-faithful schema, Silver standardizes the data for reuse, Gold restructures the data for analytical consumption, and Audit stores evidence required to trace the data lifecycle.

Discussion

The proposed lakehouse architecture and table schemas demonstrate how government open data can be organized from a raw source file into structured Landing, Bronze, Silver, Gold, and Audit layers. However, the architecture itself does not directly indicate how effectively data quality improves across the pipeline. A structured evaluation concept is therefore required to assess whether each transition, from Landing to Bronze, Bronze to Silver, and Silver to Gold, improves data quality, usability, traceability, and analytical readiness.

ISO/IEC 25012 is adopted as the conceptual foundation for this assessment because it provides a data quality model for defining and evaluating data quality requirements in data acquisition, integration, assessment, and improvement processes (International Organization for Standardization, 2008). This makes it suitable for the proposed lakehouse architecture, which does not only store government open data, but also restructures, standardizes, documents, and prepares data for analytical use.

1. ISO/IEC 25012 and Layer-Based Data Quality Interpretation

ISO/IEC 25012 defines data quality through fifteen characteristics grouped into two main perspectives: inherent data quality and system-dependent data quality. Inherent data quality refers to characteristics of the data itself, while system-dependent data quality refers to characteristics influenced by the system or technical environment in which the data is stored, accessed, maintained, and used.:

Table 8 ISO/IEC 25012 data quality perspectives

Perspective	Main meaning	Characteristics
Inherent data quality	Quality characteristics related to the data itself	Accuracy, completeness, consistency, credibility, currentness
System-dependent data quality	Quality characteristics influenced by the data system	Accessibility, compliance, confidentiality, efficiency, precision, traceability, understandability, availability, portability, recoverability

This distinction is important because not every quality characteristic can be improved directly by the lakehouse pipeline. For example, factual accuracy may require external reference data, while traceability, understandability, accessibility, and compliance can be improved through architectural design. Therefore, ISO/IEC 25012 is used not as a checklist, but as a framework for identifying which quality characteristics can be measured, improved, or documented within the proposed medallion pipeline.

The medallion architecture supports layer-based interpretation because each layer has a distinct role in the data lifecycle. Landing preserves the original file, Bronze converts the source into a governed lakehouse table, Silver standardizes and prepares the data for reuse, and Gold restructures the data for analytical consumption.

Table 9 Data quality interpretation by lakehouse transition

Layer transition	Main architectural change	Expected quality improvement
Landing to Bronze	Converts a headerless source file into a structured lakehouse table with assigned column names, source context, and technical metadata	Improves understandability, accessibility, traceability, credibility, and compliance
Bronze to Silver	Applies standardization, validation support, data type handling, and administrative context normalization	Improves consistency, completeness, compliance, and readiness for reuse

Layer transition	Main architectural change	Expected quality improvement
Silver to Gold	Reshapes and aggregates standardized data into analytical tables	Improves analytical accessibility, understandability, and usability for reporting

This interpretation is relevant for government open data because datasets may be publicly available but still difficult to analyze when they lack headers, metadata, schema definitions, or query access. The proposed lakehouse architecture addresses this issue by creating intermediate layers that preserve source evidence while progressively improving structure and usability.

2. Operationalizing ISO/IEC 25012 through Measurable Metrics

ISO/IEC 25012 provides high-level quality characteristics, but these characteristics must be operationalized into measurable indicators before they can be applied to a data pipeline. For example, completeness should be measured through values such as missing cell count, missing cell ratio, required attribute missing ratio, or metadata completeness ratio.

Numerical metrics are necessary because they provide reproducible evidence, support comparison across layers, allow both counts and ratios to be stored, and separate measurement from interpretation. The Audit layer stores measured values, while quality improvement is interpreted later by comparing those values according to the expected direction of each metric.

Table 10 Examples of numerical interpretation across layers

Metric	Landing	Bronze	Silver	Interpretation
assigned_attribute_name_ratio	0.00	1.00	1.00	Bronze improves understandability by assigning names to unnamed source fields.
data_cell_missing_ratio	0.04	0.04	0.01	Bronze preserves source missingness, while Silver improves completeness through treatment.
query_success_flag	0	1	1	Bronze improves accessibility by converting the file into a queryable table.
record_traceability_ratio	0.00	1.00	1.00	Bronze improves row-level traceability through technical metadata.

The expected direction of each metric is essential. Some metrics improve when values increase, such as assigned attribute name ratio or traceability ratio. Others improve when values decrease, such as missing value ratio or duplicate record ratio. This enables consistent comparison without storing subjective conclusions in the audit record.

3. Selected Characteristics and Universal Metrics

The proposed framework applies the five inherent ISO/IEC 25012 characteristics as a baseline: accuracy, completeness, consistency, credibility, and currentness. However, the extent to which each characteristic can be improved depends on the available metadata and validation evidence. For example, the pipeline can record evidence related to accuracy, such as declared type coverage or reference rule availability, but it cannot prove factual correctness without an external authoritative source.

For system-dependent characteristics, the framework focuses on traceability, understandability, accessibility, and compliance because these are directly affected by the medallion process. Traceability improves when records are linked to source files, checksums, and pipeline runs. Understandability improves when headerless fields receive readable names and contextual metadata. Accessibility improves when raw files become registered and queryable Iceberg tables. Compliance improves when tables conform to internal naming standards, schema contracts, and metadata requirements.

Table 11 Selected ISO/IEC 25012 characteristics and proposed metrics

Characteristic	Perspective	Example metrics	Expected direction
Accuracy	Inherent	declared_type_coverage_ratio; reference_value_coverage_ratio; inferred_type_ambiguity_ratio	Coverage increases; ambiguity decreases
Completeness	Inherent	data_cell_missing_ratio; required_attribute_missing_ratio; technical_metadata_missing_ratio; context_metadata_missing_ratio	Missing ratios decrease
Consistency	Inherent	record_width_conformance_ratio; schema_attribute_match_ratio; duplicate_record_ratio; context_value_consistency_ratio	Conformance increases; error ratios decrease
Credibility	Inherent	source_uri_present_flag; source_checksum_present_flag; source_metadata_record_present_flag	Evidence flags become true
Currentness	Inherent	source_period_present_flag; ingestion_timestamp_present_flag; ingestion_age_days; source_period_age_days	Presence flags become true; age interpreted by context
Traceability	System- dependent	record_traceability_ratio; source_checksum_traceability_ratio; run_identifier_traceability_ratio	Ratios increase
Understandability	System- dependent	assigned_attribute_name_ratio; human_readable_attribute_name_ratio; context_attribute_presence_ratio	Ratios increase

Characteristic	Perspective	Example metrics	Expected direction
Accessibility	System-dependent	table_exists_flag; query_success_flag; catalog_registration_flag	Flags become true
Compliance	System-dependent	naming_standard_conformance_ratio; required_metadata_attribute_presence_ratio; schema_contract_conformance_ratio	Ratios increase

The framework prioritizes universal metrics that can be applied to government open data generally, rather than only to the DOPA population dataset. This is important because public datasets from different agencies may share similar issues, such as missing headers, weak metadata, inconsistent formats, limited documentation, and lack of query access. Dataset-specific checks, such as population total reconciliation or negative population validation, may be added later as extensions, but they are separated from the universal assessment framework.

4. Audit-Based Measurement and Governance Visibility

The lakehouse.audit.data_quality_results table is designed to store numerical measurement evidence across layers. Each record can include fields such as iso_characteristic, measurement_name, measurement_value, measurement_unit, numerator_value, denominator_value, expected_direction, measurement_scope, layer_name, layer_order, measured_at_utc, and details_json. This structure allows the same metric to be calculated across layers and compared according to its expected direction.

Cross-layer comparison must consider measurement scope. Some metrics can be compared directly between Landing, Bronze, and Silver, but Gold tables may change the data grain through reshaping or aggregation. For example, an age-long table converts age columns into row-level age values, which changes row and column counts. Therefore, metrics for Gold should be interpreted with supporting context such as measurement scope, data grain, and calculation notes.

OpenMetadata complements the Audit layer by providing governance visibility. It can help users discover tables, inspect schemas, view lineage, review profiling information, and expose quality-related metadata. However, the Audit schema remains the durable source of measurement evidence inside the lakehouse. This separation allows the architecture to support both reproducible quality assessment and practical data governance.

Conclusion

This study is positioned as Design Science Research (DSR) and a pre-production design study that proposes two related artifacts for government open data: a medallion-based lakehouse architecture and an ISO/IEC 25012-based evaluation framework. Using the population statistics dataset from the Department of Provincial Administration (DOPA), Ministry of Interior (MOI), Thailand, as a design case, the study shows how a headerless government source file can be organized into Landing, Bronze, Silver, Gold, and Audit layers. The proposed architecture separates raw file preservation, source-faithful structuring, standardization, analytical modeling, and audit evidence. The Landing layer preserves the original file and metadata; the Bronze layer assigns a durable schema and source context; the Silver layer prepares standardized and quality-ready records; the Gold layer restructures data into consumption-ready analytical tables; and the Audit layer stores evidence for traceability, lineage, and measurement.

The lakehouse architecture and ISO/IEC 25012 evaluation framework are presented together because they were designed as interdependent artifacts. The architectural layers define where data are preserved, structured, standardized, and reshaped, while ISO/IEC 25012 provides the conceptual basis for interpreting and measuring data quality across those layer transitions. Therefore, quality dimensions such as completeness, consistency, credibility, currentness, traceability, understandability, accessibility, and compliance are translated into measurable indicators that can be stored in the Audit layer and used for future assessment.

As a pre-production design study, this research does not report prototype implementation, before-and-after empirical measurement, or system performance benchmarking. These activities require operational deployment and are beyond the current scope. Instead, the contribution of this study is architectural and methodological: it provides a reusable design foundation for developing transparent, traceable, and analytically ready government open-data lakehouse pipelines. Future work should implement the proposed architecture, automate quality measurement in the Audit layer, publish lineage and profiling evidence through metadata governance tools, and evaluate the framework across multiple government datasets and operational environments.

Reference

- Databricks. (2026). *What is the medallion lakehouse architecture?*
<https://docs.databricks.com/aws/en/lakehouse/medallion>
- Digital Government Development Agency. (2021, January 14). *Open government data*.
<https://www.dga.or.th/en/our-services/digital-platform-services/open-government-data/> (In Thai)
- International Organization for Standardization. (2008). *ISO/IEC 25012:2008: Software engineering—Software product Quality Requirements and Evaluation (SQuaRE)—Data quality model*.
<https://www.iso.org/standard/35736.html>
- Kattiyapornpong, W., Phunchongharn, P., & Suteerawattananonda, N. (2017). Open data quality measurement for Thailand Open Government Data. *Engineering Journal of Siam University*, 18(35), 26–37. <https://e-library.siam.edu/e-journal/wp-content/uploads/2018/11/Engineering-Journal-of-Siam-University-2017-vol18-no35-03-compressed.pdf> (In Thai)
- Mazumdar, D., Hughes, J., & Onofre, J. B. (2023). *The data lakehouse: Data warehousing and more*. arXiv.
<https://arxiv.org/abs/2310.08697>
- Microsoft. (2026). *Implement medallion lakehouse architecture in Microsoft Fabric*.
<https://learn.microsoft.com/en-us/fabric/onelake/onelake-medallion-lakehouse-architecture>
- Office of the Prime Minister. (2025, December 30). *Thailand: Letter of intent to join OGP*. Open Government Partnership. https://www.opengovpartnership.org/wp-content/uploads/2026/02/Thailand-Letter_of_Intent_to_join_OGP.pdf
- Open Data Institute. (2024). *Empowering Thailand's digital government with open data*.
<https://theodi.org/insights/reports/empowering-thailands-digital-government-with-open-data/>
- Open Government Partnership. (n.d.). *Open government declaration*. <https://www.opengovpartnership.org/how-we-work/joining-ogp/open-government-declaration/>
- Open Government Partnership. (2026). *Thailand*. <https://www.opengovpartnership.org/members/thailand/>
- Sawadogo, P., & Darmont, J. (2021). On data lake architectures and metadata management. *Journal of Intelligent Information Systems*, 56, 97–120. <https://doi.org/10.1007/s10844-020-00608-7>
- Sawadogo, P., Scholly, E., Favre, C., Ferey, E., Loudcher, S., & Darmont, J. (2019). Metadata systems for data lakes: Models and features. *Proceedings of the 28th International Conference on Information Systems Development*. <https://arxiv.org/abs/1909.09377>
- Wirtz, B. W., Weyerer, J. C., Becker, M., & Müller, W. M. (2022). Open government data: A systematic literature review of empirical research. *Electronic Markets*, 32, 2381–2404.
<https://doi.org/10.1007/s12525-022-00582-8>