Data Curation

1. Introduction

Data curation is the process of collecting, organizing, and preserving data for future use. It is essential for ensuring the quality and usability of data, and it is becoming increasingly important as the volume and complexity of data continues to grow.

In the context of a data ocean, data curation is even more critical. A data ocean is a vast repository of data that is collected from a variety of sources. This data can be structured, unstructured, or semi-structured, and it can be of varying quality.

The goal of data curation in a data ocean is to ensure that the data is:

- **Accurate:** The data must be free of errors and omissions.
- **Complete:** The data must be comprehensive and cover all aspects of the domain of interest.
- **Consistent:** The data must be formatted and organized in a consistent manner.
- **Reliable:** The data must be trustworthy and reliable.
- **Usable:** The data must be easy to find, access, and understand.

Data curation in a data ocean can be challenging, but it is essential for making the data valuable and accessible to users.

2. Data Normalization

2.1. Definition

Data Normalization involves transforming data into a common format to enable seamless integration and analysis.

Data normalization is the process of organizing data in a consistent manner. This involves standardizing the data format, removing duplicate data, and identifying and correcting errors.

2.2. Importance

When data from various sources is aggregated, there's often a mismatch in formats, units, or encoding. Normalization resolves these disparities, ensuring consistency and reducing redundancy, making data integration and analytics more efficient, ensuring a single version of truth exists within the Data Ocean.

Data normalization can improve the efficiency of data processing and analysis, and it can also help to improve the quality of data.

2.3. Typical Rules and Actions

1. Capitalization: Uniformly capitalize textual data.

2. Date Formatting: Standardize date formats to YYYY-MM-DD UDT.
3. Currency Conversion: Convert all currency to a standard unit.
4. Measurement Unit Standardization: Convert all measurements to a standard unit (e.g., kilometers, USD).

2.4. Metrics and KPIs

Some relevant metrics to implement in a monitoring Data Quality Dashboard:

1. Data Consistency Ratio: the level of uniformity in the dataset after normalization procedures have been applied
   - $$\frac{\text{Number of Consistent Records}}{\text{Total Number of Records}} \times 100$$
2. Efficiency Gained Post-Normalization: measures the improvement in data processing and management tasks after normalization has been implemented.
   - Typically, it might involve measuring the time saved in data processing, the reduction in errors due to standardization, or the improvement in speed of data retrieval.
3. Data Redundancy Factor: Measure of duplicate data before and after normalization.
4. Normalization Time: Time required to normalize a dataset.
5. Normality Score: A composite score representing how well the data conforms to normalization rules.

2.5. Data Ocean Enforced Rules

Data normalization is currently carried out via the ETL (Extract, Transform, Load) tool, tailored individually to the requirements of each case.

The specific normalization procedures are outlined within the mapping rules established for every Business Entity pertinent to a particular Domain (see Data Mapping Rules in each Domain).

Standard data normalization practices currently in operation include:

- Cast or Data Type Conversion: Essential for normalizing Codes and IDs across varying source systems to ensure uniformity.
- Date Format Normalization:
  - Dates should be consistently formatted to UTC, adhering to the international standard of YYYY-MM-DD UDT.
  - If necessary, maintain an additional column for the date in the original source system format.
- Text Case Standardization: By default, convert text to lowercase with the initial letter capitalized, unless specific business requirements dictate otherwise.
- Whitespace Trimming: Remove leading and trailing spaces from all string data not utilized as Primary or Foreign Keys (PK/FK).
- Surrogate Key (SGK) Generation: Each table will be equipped with a nonsensical, unique technical key to ensure a consistent method of record identification (see Table creation and definition).
- "Ghost" Record Insertion: All tables serving as Dimensions will include "Ghost" records to accommodate exceptions and guarantee Referential Integrity (as detailed in the Dimension and SCD definitions).
- Derived Column Creation: Implement derived columns as necessary for enhanced data analysis and reporting.
- Handling NULL Values: Substitute NULLs with a default value for all columns acting as Primary or Foreign Keys (PK/FK) (as outlined in the Table definition and Default Value documentation).

A comprehensive list of over 20 potential rules is identified, with their implementation definitions outlined; ready to be used.

3. Data Validation

3.1. Definition

Data Validation (DV) is the process that ensures the data complies with the defined formats, rules, standards and business-specific constraints. It is the process of checking data for errors and omissions, of ensuring that the data is accurate, complete, and consistent.

This process is more concerned with validating data against specific criteria, such as format checks, value constraints, and relationships.

Data Validation can be achieved following several approaches:

1. **Data Profiling:** by profiling the incoming data to understand its structure, patterns, and anomalies. This includes examining data types, values, and ranges.
   - Analyzing incoming customer data to understand its structure. For instance, identifying find fields like 'Name,' 'Email,' 'Address,' and 'Phone.'
   - Identifying patterns, such as email addresses should contain "@" and have a valid domain.
2. **Rule-Based Validation:** Defining and implement validation rules that data should adhere to. These rules can include format checks, value constraints, and referential integrity.
   - Defining validation rules, e.g., 'Email' must follow a valid email format, 'Phone' should consist of only numbers, and 'Customer ID' must be unique.
   - For example, ensuring that dates are in the correct format or that numeric values fall within specific ranges.
3. **Statistical Analysis:** Utilizing statistical methods to identify outliers and unusual data patterns. This can help in detecting potential issues.
   - Using statistical methods to detect anomalies. For example, you detecting an unusually high number of customers with the same 'Phone' number.

4. **Data Schema Validation:** Ensuring that the incoming data aligns with the predefined schema and metadata. Any variances should be flagged.
    - Ensuring that the incoming data aligns with the predefined schema. If a new field, like 'Birthday,' is introduced, ensuring that the schema is updated.
5. **Automated Testing:** Implementing automated testing processes to continuously validate data as it enters the DW. Automated tests can run regularly to detect issues promptly.
    - Implementing automated tests that run upon data arrival. If any data violates the predefined rules, generate alerts or logs for further investigation.

3.2. Importance

It's crucial for building trust and reliability in data.

Unverified or incorrect data can lead to erroneous conclusions, and misleading insights, which in turn can have a significant adverse impact on business decisions.

It involves validating the data against quality standards and identifying any errors or inconsistencies.

This can be done manually or automatically using a variety of tools and techniques.

3.3. Typical Rules and Actions

1. Type Checks: Validate the data type (text, integer, float, etc.).
2. Format Checks and Validation:
    - The data must match a specific format.
    - Validate text patterns like email, phone numbers, and dates.
3. Range Checks: Verify that numerical data lies within defined ranges.
    - The data must be within a specified range.
    - Boundary Values Validation
4. Completeness Checks: Ensure all mandatory fields are filled.
5. Uniqueness Check: Verify that primary keys or unique identifiers do not have duplicates.
6. Consistency Check
    - The data must be consistent with other data.
    - Cross-Field Checks: It verifies the relationships between different data fields. For example, ensuring that an order's shipping date is not earlier than the order date.
    - Data Integrity: Validating that data relationships and constraints are maintained. This includes checking that primary keys and foreign keys in a database are correctly linked.
7. Domain Checks: Ensure data belongs to a defined set of permissible values.

Rules and metrics that can be used for data validation include:

- **Completeness**: Ensure that all required data fields are present and contain valid values.
- **Consistency**: Ensure that the data is consistent across all sources and that there are no conflicting values.
- **Accuracy**: Ensure that the data is accurate and reflects the real-world values it represents.
- **Timeliness**: Ensure that the data is up-to-date and reflects the latest information.

Data Validation practices in terms of Data Management:

- **Detection**: DV focuses on detecting data anomalies, errors, and issues.
    - The goal of DV rules is to detect errors, anomalies, and inconsistencies in the data.
    - Detected issues are typically related to non-compliance with specific data standards and rules.
- **Compliance**: It ensures that data adheres to defined rules and constraints, such as referential integrity checks.
    - DV rules are primarily concerned with ensuring the correctness and integrity of the data.
    - DV focus on validating data against predefined criteria and constraints, often related to data structure and integrity.
    - DV rules include checks like format validation (e.g., email format), uniqueness validation (e.g., unique IDs), and structure validation (e.g., address format).
- **Immediate Feedback**: When issues are detected, the primary action is to raise alerts or notifications and possibly reject or flag the non-compliant data.
    - When DV rules detect violations, the primary action is to provide immediate feedback, such as alerts or data rejection.
    - The validation results should also be recorded in some table for later analysis. This table can have columns to capture information such as the rule name, record details, and the date and time of the validation.
    - Alternatively, the ETL tool log can be used to log validation rules messages
- **Data Cleansing**: DV may involve basic data cleansing steps to make the data conform to standards.
    - DV rules are often applied during data ingestion and initial processing phases to prevent incorrect data from entering the system.

3.4. Metrics and KPIs

Some relevant metrics to implement in a monitoring Data Quality Dashboard:

1. Data Validation Success Rate or Validation Accuracy: The percentage of records that have been validated correctly (that meet all validation rules) out of the total records processed.
2. Data Rejection Rate: The percentage of records that were rejected during validation due to errors or not meeting predefined criteria.
3. Time Taken for Validation: The total duration required to complete the validation process for a batch of data or a single record.
4. Number of Manual Interventions Required: The count of instances where human input or correction was necessary during the data validation process.
5. Field-Level Compliance Rate: The proportion of individual data fields across all records that pass validation checks.
6. Failed Validation Alerts: The total number of automated notifications generated when data does not pass the validation process.

3.5. Data Ocean Enforced Rules

Data validation is an important part of data curation, as it helps to ensure that the data is accurate and complete.

Presently, it primarily relies on the ETL (Extract, Transform, Load) tool for real-time execution. In this approach, data validation checks are seamlessly integrated into the ETL pipeline. This ensures that data quality issues are promptly detected and addressed during data ingestion and transformation. Real-time data validation enables immediate feedback and corrective actions, mitigating the impact of poor-quality data on downstream processes.

Furthermore, the approach is tailored to the specific needs of each case. Detailed validation procedures are delineated within the mapping rules established for each Business Entity associated with a specific Domain. For more information on validation within a particular Domain (refer to the corresponding Data Mapping Rules Document).

Standard data validation practices currently in operation include:

- Data Profiling
  - To understand data structure, patterns, and anomalies.
  - This procedure is also being used to drive the Data Model implementation
  - Two methods are available: ETL feature and Python script
- Schema validation for files with a control and logging mechanism is in place
  - Needs to be enhanced with the implementation of an automatic alert mechanism to notify senders.
- Schema validation and error prevention for tables are directly facilitated by followed approach, which mandates the explicit identification of source columns.
  - This approach ensures that the process will not be halted unless an existing column is intentionally removed
- Completeness Checks for mandatory fields and default values generation.
- Uniqueness Check with primary keys verification.
- Data Integrity and Referential Integrity checks
  - Performing direct ETL lookups with the identification of exceptions.
  - Needs to be enhanced with the implementation of an automatic "ghost" record insertion

This process must be optimized, promoted and shared among peers. An extensive list of more than 20 potential rules is available, along with clear implementation definitions, that are ready to be applied to various data patterns, including emails, phone numbers, dates, and more.

3.5.1. New Developments

Data validation within the Data Ocean framework is still an ongoing area of research.

An extra approach under consideration involves the use of a Batch Data Validation (Scheduled Moment in Time), where data undergoes scheduled checks outside the ETL process, typically on a weekly, or monthly basis, depending on the organization's needs. Rather than presenting an alternative, this approach is viewed as an additional layer of validation, providing a double-check to ensure thorough data validation. This process is likely to be integrated into the Data Quality process.

It's worth noting that a comprehensive implementation proposal has already been developed, which includes design and architectural considerations.

In addition to this implementation proposal, another avenue being explored is the utilization of Open-Source tools like "Great Expectations" and "Google Data Validation Tool (DVT)". While these tools offer robust capabilities, they do require a certain level of effort for learning and implementation. Nevertheless, their potential to significantly enhance data validation processes is acknowledged

A comprehensive list of over 20 potential rules is identified, with their implementation definitions outlined; ready to be used.

4. Data Quality

4.1. Definition

Data Quality is a broader evaluation of the overall health and fitness of the data, measuring the fitness of data for its intended uses in operations, decision-making, and analytics.

This process considers the quality of data in a more comprehensive manner, looking at not just specific rules but the impact of relationships on the data's usefulness and reliability.

Data Validation (DV) is primarily about the detection of issues, ensuring that data meets predefined criteria and standards, while Data Quality (DQ) extends beyond detection to taking actions to maintain and improve the overall quality of the data:

- **Action and Improvement**: DQ goes beyond detection to take actions for maintaining and enhancing data quality.
- **Reporting**: DQ often includes reporting to stakeholders about data quality issues and trends.
- **Feedback Loop**: It establishes a feedback loop with data stewards, source system owners, and other relevant parties to resolve issues.

- **Corrective Actions**: In cases where data quality issues are significant, DQ may involve corrective actions such as issuing ghost records or other measures to maintain DW consistency and coherence.

Context

- **Operational Systems**: Implement corrective actions that rectify data directly.
- **Analytical Systems**: The focus should remain on identifying, monitoring and alerting anomalies without altering the data. Changing data at this level will not correct the source, making the exercise futile.

For more information on the subject, please refer to the link Data Quality.

4.2. Importance

Data quality is a measure of the accuracy, completeness, and consistency of data.

High-quality data is essential for making informed decisions, and it is also important for ensuring the reliability of data-driven systems.

There are a number of factors that can affect data quality, including:

- The quality of the data collection process
- The quality of the data storage and processing systems
- The quality of the data governance processes

Data quality can be improved through a variety of measures, including:

- Implementing data validation and normalization procedures
- Enforcing data quality policies and standards
- Educating users about data quality
- Using data quality tools and techniques

4.3. Rules and Actions

1. **Accuracy Checks**: Verify that data accurately represents the real-world values it is expected to model.
2. **Completeness Checks**: Ensure all essential data is present and that missing values are handled appropriately.
3. **Consistency Checks**: Confirm that data across different systems or datasets remains consistent in format and value.
4. **Timeliness Checks**: Check that data is updated regularly and is available when needed.
5. **Uniqueness Checks**: Validate that each data entry is unique as required and there are no unintended duplicates.
6. **Conformity Checks**: Verify data is in the expected format and range and conforms to data domain constraints.
7. **Integrity Checks**: Ensure that relationships among data entities and attributes are maintained correctly.

4.4. Metrics and KPIs for Data Quality:

1. **Accuracy Rate**: The percentage of data entries that pass accuracy checks.
2. **Data Completeness**: The proportion of complete records in the dataset (no missing values where required).
3. **Consistency Index**: A measure of the uniformity of data across sources and over time.
4. **Timeliness Score**: The degree to which data is updated and provided within the expected time frame.
5. **Uniqueness Ratio**: The ratio of unique records against potential duplicates found.
6. **Conformity Level**: The percentage of data entries that adhere to defined format and range constraints.
7. **Integrity Compliance Rate**: The percentage of data relationships (foreign keys, parent-child records) that are correctly represented.

A comprehensive list of over 20 potential rules is identified, with their implementation definitions outlined; however, they have not been put into practice yet.

4.5. Data Ocean Enforced Rules

4.5.1. Existing Initiatives

A current production-level product already covers Data Quality in some of the company's operational domains. This product includes KPIs, visualized through a dashboard, and triggers corrective actions.

For more in-depth information on existing solution, please refer to the link Data Quality dashboard.

4.5.2. Importance of Integration with Data Ocean

At this point, Data Quality validation is not an immediate focus. It's important to highlight that a dedicated project is already in production, addressing this specific aspect. However, the plan is to eventually incorporate these improvements into the Data Curation layer in upcoming phases.

While the current initiatives are beneficial, their integration into the Data Ocean will establish a centralized control system for data quality, offering significant value for overall data governance and analytics.

5. Future Actions

Some proposed actions.

5.1. Implement Data Quality within the Data Ocean ecosystem

Taking a cue from the Data Quality KPI Dashboard, a potential step forward to enhance data curation within the Data Ocean context is the introduction of a centralized data quality initiative. This initiative would have the responsibility of overseeing data quality throughout the entire Data Ocean ecosystem. Its primary role would involve identifying and promptly alerting stakeholders about any data quality concerns.

5.1.1. Strategy

An iterative and continuous approach is essential for Data Validation and Data Quality. These processes should be regularly assessed, monitored, and improved, and involve collaboration with data stewards to ensure data quality is maintained.

1. **Data Quality Assessment:**
    a. Develop a framework to assess the overall data quality.
    b. This can involve metrics such as completeness, accuracy, consistency, and timeliness.
2. **Data Profiling for Quality:**
    a. Extend data profiling to identify specific data quality issues.
    b. This can include missing values, duplicates, and inconsistent data.
3. **Data Quality Rules:**
    a. Define quality rules to evaluate the data quality dimensions.
    b. For instance, a completeness rule might ensure that critical fields are never empty.
4. **Data Quality Dashboards:**
    a. Create visual dashboards to provide real-time insights into data quality.
    b. This helps both data stewards and source systems to monitor the quality of their data.
5. **Data Quality Reporting:**
    a. Develop automated reports that summarize data quality issues and distribute them to data stewards and source system owners.
6. **Collaboration with Data Stewards:**
    a. Continue to promote existing processes
    b. Establish communication channels with data stewards or source system owners, not yet included.
    c. When data quality issues are detected, inform and collaborate with them for resolution.
7. **Data Quality Feedback Loop:**
    a. Implement a feedback loop where data quality issues detected in the DW are reported back to the source systems. This allows them to take corrective actions.
8. **Documentation:**
    a. Document data quality rules, issues, and their resolutions.
    b. This knowledge repository helps in maintaining data quality over time.

5.1.2. Data Model

Introduction:

To effectively monitor, assess, and enhance data quality, a well-structured data model is essential. This chapter delves into the concept of a data quality assessment data model, offering insights into its design, advantages, and considerations.

Design of the Data Model:

The core of the data model revolves around a pivot-style structure that incorporates dimensions, metrics definition, and values. This flexible design allows for dynamic expansion and adaptability to changing data quality assessment needs.

Breaking down the components of this model:

Dimensions:

1. **GBU and or Site:** This dimension categorizes data quality assessments based on the business unit or department responsible for the data.
2. **Domain:** Data quality can vary between different subject areas or domains. This dimension helps classify assessments by subject area.
3. **Sub-Domain or Business Object:** Further granularity is achieved by categorizing data quality assessments based on sub-areas or specific business objects.
4. **KPI Name:** Key Performance Indicators (KPIs) are essential for evaluating data quality. This dimension captures the specific KPIs being measured.
5. **Source System:** It is important to trace data quality issues back to their source. The source system dimension helps identify the origin of data quality challenges.

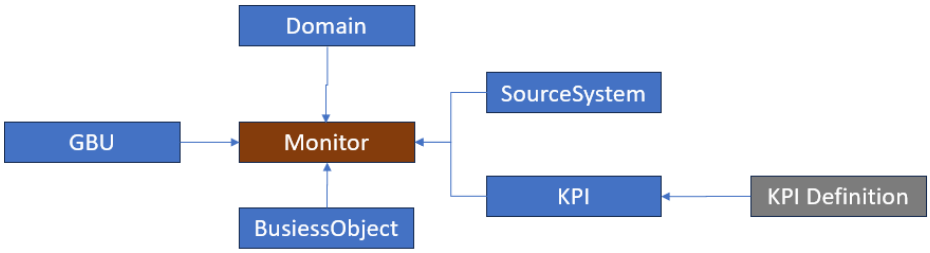
Fact Table:

The fact table links the dimension data to the assessment metrics, and it stores values in different formats:

1. **Dimension Foreign Key:** This connects each data quality assessment to the relevant dimensions, providing context for the assessment.
2. **Value as String:** Captures data quality assessment values as strings, facilitating the storage of textual information.
3. **Value as Number:** Data quality assessments can often involve numeric measurements. This field stores assessment values as numbers.
4. **Value as Percentage:** Percentage-based measurements, crucial for some KPIs, are recorded in this field.

Supporting Components:

1. **KPI Definition Table:** This table stores SQL instructions that define each Key Performance Indicator (KPI) implementation. It provides the flexibility to define new KPIs by simply adding a new SQL query entry.
2. **ETL Process:** An Extract, Transform, Load (ETL) process is responsible for executing each entry in the KPI Definition Table. It processes the SQL queries and feeds the resulting output into the Fact Table.
3. **Automated Maintenance:** The Fact Table and Dimensions are automatically maintained, ensuring data integrity and consistency in the data quality assessment process.



Advantages of the Data Model:

1. **Flexibility:** The model dynamically adapts to new DQ checks, dimensions, and values, accommodating evolving requirements.
2. **Scalability:** It easily grows to meet expanding DQ assessment needs, providing a comprehensive view of data quality.
3. **Aggregation:** Capturing DQ metrics in multiple formats allows for diverse analysis and aggregation options.
4. **Historical Tracking:** Historical records enable trend analysis and continuous improvement of data quality.
5. **Granular Detail:** The model captures fine-grained information, aiding in the diagnosis of specific data quality issues.
6. **Ease of Integration:** The data model integrates seamlessly with data quality tools, dashboards, and reporting systems, simplifying DQ reporting and analysis.

Considerations:

1. **Data Volume:** As the model grows, ensure sufficient storage capacity to handle increased data volume.
2. **Performance:** Optimize database systems and queries for efficient data retrieval and analysis to maintain good performance.
3. **Data Security:** Implement stringent access controls and security measures to protect sensitive DQ data.
4. **Data Retention:** Define data retention policies to manage historical data effectively.
5. **Documentation:** Maintain clear and up-to-date documentation to ensure consistency and understanding among users.

Conclusion:

A well-structured data quality assessment data model serves as the foundation for effective DQ monitoring, reporting, and continuous improvement. Its adaptability and scalability make it a valuable asset in maintaining and enhancing data quality throughout the data lifecycle. By considering the advantages and considerations outlined in this chapter, organizations can build a robust framework for data quality excellence.

5.2. Establishment of a data quality initiative at the operational level

Another prospective strategy to consider is the establishment of a data quality initiative at the operational level, geared towards real-time data analysis and rectification. This approach reveals its significance in addressing data anomalies and discrepancies promptly, thereby maintaining the integrity of the information ecosystem.

This operational-level data quality initiative would involve deploying advanced algorithms and automated processes that continuously monitor incoming data streams. By leveraging real-time analytics, this system can instantaneously identify deviations from predefined data quality benchmarks. In the event of discrepancies, automated corrective measures can be applied, ranging from data enrichment through external sources to flagging erroneous entries for manual review.

A critical aspect of this initiative would be its proactive nature. Instead of relying solely on retrospective audits, it would function in an anticipatory mode, precluding the propagation of erroneous data into downstream processes. Timely alerts would be generated for immediate corrective actions, minimizing the risk of inaccurate insights, faulty decision-making, or downstream process disruptions.

Furthermore, such an operational-level data quality initiative would synergize with the existing data curation practices, forming a robust defense against data inconsistencies. This approach not only aligns with best practices in data governance but also positions the Data Ocean architecture for greater reliability and value generation.

To execute this initiative effectively, collaboration across cross-functional teams, including data engineers, analysts, and domain experts, is crucial. Additionally, the establishment of clear workflows, data quality metrics, and continuous performance monitoring mechanisms will be pivotal to ensure its success. By integrating real-time data quality assurance into the Data Ocean, this initiative can significantly elevate the overall data ecosystem's reliability and usability.

Existing SAP Info Steward could be used.

5.3. Select and implement a Data Validation Tool

Conclude the thorough analysis of the identified tools and choose one for conducting a Proof of Concept (POC).

6. References

- Data Curation: https://en.wikipedia.org/wiki/Data_curation
- Data Validation: https://en.wikipedia.org/wiki/Data_validation
- Data Normalization: https://en.wikipedia.org/wiki/Data_normalization
- Data Quality: https://en.wikipedia.org/wiki/Data_quality
- "Great Expectations": https://greatexpectations.io/
- "Google Data Validation Tool (DVT)": https://cloud.google.com/blog/products/databases/automate-data-validation-with-dvt
- "Data Quality: The Accuracy Dimension" by Jack E. Olson, 2003.
- "Managing Data in Motion" by April Reeve, 2013.
- "Data Governance: How to Design, Deploy and Sustain an Effective Data Governance Program" by John Ladley, 2012.