DAU - Certified Data & Analytics Tester (CDAT)

Glossary

Version 1.0

Released 22-12-2021

Copyright Notice

Thank you to the main authors

- Rogier Ammerlaan, Jaap de Roos and Armando Dörsek

Thank you to the review committee

Alexis Jesús Herrera Colmenares, Ángel Rayo Acevedo, Arjan Brands, Aurelio Gandarillas Cordero, Beatriz López Botello, Benjamín Gillibrand H., Christine Green, Daniel Leo Lopez Romero, Egbert Bouman, Eibert Dijkgraaf, Emilie Potin-Suau, Fabiola Mero, Geoffrey Wemans, Héctor Ruvalcaba, Hermon Alfaro Fuentes, Huib Stoel, Jayapradeep Jiothis, Jean-Luc Cossi, Joan Gasull Jolis, José Díaz, Juan Pablo Rios Alvarez, Julie Gardiner, Kaan Sanli, Koos van Strien, Krishan Premcharan, Kyle Alexander Siemens, Márcia Araújo Coelho, Mario Alvarez Gómez, Mark Summers, Melissa Pontes, Miaomiao Tang, Miguel Angel de León Trejo, Nadia Soledad Cavalleri, Neriman Kocaman, Paula Castro, Petr Neugebauer, Dr. Ralph Elster, Ruth Margaret Florian Caipa, Santiago de Jesús Gonzalez Medellin, Samuel Ouko, Shreyansh Tewari, Silvie Schoenmaker, Dr. Srinivas Padmanabhuni, Thomas Cagley, Tim Koomen, Tobi Steenbakkers, Vanessa Islas Padilla, Vipul Kocher, Wim Decoutere

Revision History

Version	Date	Remarks
1.0	Dec 22nd, 2021	Initial release

Table of contents

Glossary

A

Aggregations
Source: PowerBI glossary
The reduction of rows in underlying data sources to fit in a model. The result is an aggregate.

Agile
Source: wikipedia
Is a set of practices intended to improve the effectiveness of software development professionals, teams, and organizations.

Agile testing
Source: Glossary ISTQB Foundation
Testing practice for a project using Agile software development methodologies, incorporating techniques and methods, such as extreme programming (XP), treating development as the customer of testing and emphasizing the test-first design paradigm.

Analytics
Comprises the combined corporate and external data sources for the creation of analytical models.

Anonymization
Data anonymization has been defined as a “process by which personal data is irreversibly altered in such a way that a data subject can no longer be identified directly or indirectly (for instance by using additional data sources)”.

Audit trail
Source: wikipedia
A security-relevant chronological record, set of records, and/or destination and source of records that provide documentary evidence of the sequence of activities that have affected at any time a specific operation, procedure, event, or device.

ASCII
ASCII is a 7-bit encoding technique which assigns a number to each of the 128 characters that are most frequently used in American English.

B

BI tester
See Reporting Tester

BI testing
Business Intelligence & Data Analytics Testing is the process of validating the data, format and performance of the ETL, analytics, reports, subject areas and security aspects of the BI & DA projects.

Big Data
Big data is a combination of structured, semi-structured and unstructured large amount of data collected by organizations that can be analyzed for predictive modeling and other advanced analytics applications.

Big Data environment
See data environment

Boundary Testing
Source: Glossary ISTQB Foundation
A black-box test technique in which test cases are designed based on boundary values.

Business Intelligence
Business Intelligence (BI) comprises the strategy, methods and technologies used to gather, store, report and analyze business data to help people analyze data to make business decisions.

Business rules
Source: wikipedia
Business rules are intended to assert business structure or to control or influence the behavior of the business. Business rules describe the operations, definitions and constraints that apply to an organization.

C

Causation
Causation explicitly applies to cases where action A causes outcome B.

Character set
See: Data Combination Test

CI/CD
Source: Wikipedia
Continuous Integration/ Continuous Delivery (or Deployment) bridges the gaps between development and operation activities and teams by enforcing automation in building, testing and deployment of applications.

Conversions
The process of changing data or causing data to change from one form to another.

Completeness Testing
Completeness testing is required to ensure that all relevant records are transferred from the source to the target and that the contents (format, precision) of each record (row) and value (column) is still correct.

Correlation
Correlation is simply a relationship between A and B. Action A relates to Action B — but one event doesn't necessarily cause the other event to happen.

Coverages
Source: ISTQB Foundation
The degree to which specified coverage items have been determined or have been exercised by a test suite expressed as a percentage.

D

DAMA
DAMA International is a not-for-profit, vendor-independent, global association of technical and business professionals dedicated to advancing the concepts and practices of information and data management. A working group of DAMA created a concise and useful set of data quality characteristics that can be well used in the BI & DA environment to establish the actual data quality.

Dashboard
A dashboard is a visual display of the most important information needed to achieve one or more objectives; consolidated and arranged on a single screen so the information can be monitored at a glance.

Data analytics
Data analytics, which mainly take place in the phase Reports, Dashboards, Analysis, comprises the combined corporate and external data sources for the creation of analytical models. There are three types of analytics: Descriptive Analytics, Predictive Analytics and Prescriptive Analytics.

Data Combination Test
Source: TMAP Body of Knowlede
The data combination test (DCoT) is a versatile technique for the testing of functionality both at detail level and at overall system level. In the embedded world, this technique is also known as the "Classification Tree Method".

Data environment
The complete business intelligence environment where data got a central place in the organization. The total of staging, data warehouses, transformations of data and reporting environment.

Data lake
A data lake is a large repository (Big Data) of internal and external data in its natural, unstructured form as a raw data reservoir.

Data lineage
Data lineage states where data is coming from, where it is going, and what transformations are applied to as it flows through multiple processes. It helps understand the data life cycle and it time aspects. It is one of the most critical pieces of information from a metadata management point of view.

Data Mart (DM)
A subset of the Enterprise Data Warehouse (EDW). The DM will hold a subset of the data in the EDW, focusing on a single subject area and it is oriented towards a specific task or department/user group, enabling them to perform queries fast and in an easy manner.

Data migration
Data migration is the process of moving data from one system to another.

Data migration tester
Like the ETL Tester, Data Migration Testers will need to show that data is moved to a different system in correct and complete fashion.

Data Mining
Data mining concerns the process of discovering new patterns from large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics and database systems.

Data Model
Data models define how the logical structure of a database is modeled. Data Models are fundamental entities to introduce abstraction in a database.

Data Profiling
Source: Johnson
Data Profiling is the process of examining the data available in an existing data source [...] and collecting statistics and information about that data

Data Quality
Data quality is generally considered high quality if it is fit for intended uses. Data Quality Characteristics are used to have quantitative measures for data quality.

Data Quality characteristics
Source: DAMA
Refers to both the characteristics associated with ... and to the processes used to measure or improve the quality of data

Data Quality tester
Testers that have extended knowledge in measuring the expected and perceived data quality in the system, at source systems (OLTP) or the Data & Analytics landscape.

Data Quality tools
Data quality tools (see chapter 5) often provide features that enable file/table comparisons. Of course, this is not the only function of tools, they will also inform their users on many aspects of the data quality like its frequencies and “odd” situations in the data sets.

Data set
A collection of related sets of information that is composed of separate elements but can be manipulated as a unit.

Datatype
A classification that specifies which type of value a variable has and what type of mathematical, relational or logical operations can be applied to it without causing an error.

Data Vault modelling
A database modeling method that is designed to provide long-term historical storage of data coming in from multiple operational systems.

Data visualization
Data visualization is the graphical representation of information and data.

Data Warehouse
A data warehouse, stores data from several sources, either direct or via a staging area (is an intermediate storage area used for data processing during the extract, transform and load process).

Decision Table Test
Source: ISTQB
A black-box test technique in which test cases are designed to exercise the combinations of conditions and the resulting actions shown in a decision table.

De-personification
see pseudonymization or anonymization

Descriptive Analytics
Reports, dashboards, alerts and scorecards are used to describe what has happened in the past. Descriptive Analytics may also be used to classify customers or other entities into groups that are similar on certain dimensions.

Dimensions
See also facts
Dimensions are companions to facts and describe the objects in a fact table.

Dimensional model
A data structure technique optimized for data storage in a Data warehouse.

Drill down
"Drilling down" into the data in an OLAP cube, means that the user is analyzing the data at a different level of summarization.

Drill through
When users "drill through" data, they are requesting the individual transactions that contributed to the OLAP cube's aggregated data at a lowest level of detail for a given measure value.

DTAP
Development, Testing, Acceptance and Production is a phased approach to software testing and deployment. Most cases it is connected to the test environments used in software testing.

E

E2E
Source: ISTQB
A type of testing in which business processes are tested from start to finish under production-like circumstances.

End-to-end testing
See E2E

Enterprise Data Warehouse (EDW)
See dataware house

Equivalence Partitioning
Source: ISTQB

A black-box test technique in which test cases are designed to exercise equivalence partitions by using one representative member of each partition.

ETL
ETL stands for: Extract, Transform and Load (data). Data is *extracted* from the source systems, *transformed* in a usable format and *loaded* into the (enterprise) data warehouse.

ETL Tester
In general, ETL Testers will have knowledge of logical database design, ETL-processing tools and their querying skills will enable them to create comparisons between source and target tables, incorporating aggregations and transformations according to (general) design rules and technical designs like Source-to-Target Mappings.

Extended ASCII (ISO/IEC 8859)
ISO/IEC 8859 is an 8-bit extension to ASCII developed by ISO (the international organization for standardization). Next to the 128 ASCII characters it covers characters like the Euro sign (#128), British Pound (#163) and the American Dollar Cent (#162) symbol. Several variations of the ISO/IEC 8859/Latin standard exist, to cover different language families.

F

Facts
A fact is a quantitative piece of information - such as a sale or a download. Facts are stored in fact tables and have a foreign key relationship with a number of dimension tables.

File comparison
File comparison tools will support testers when loading two files and reporting in which rows or columns differences occur.

G

GDPR
The General Data Protection Regulation is a regulation in EU law on data protection and privacy in the European Union (EU) and the European Economic Area (EEA.

Granularity
The most detailed unit of the data is a fact, a contract, invoice, spending, task, etc. Each fact might have a measure — _an attribute that can be measured, such as: price, amount, revenue, duration, tax, discount, etc. The “grain” (or granularity) of a fact table states the level of detail in the fact table.

H

Historical data
Source: wikipedia
Historical source (also known as historical material or historical data) is an original source that contains important historical information.

I

Index
A method to track the performance of some group of assets in a standardized way.

Intersect
A query that can facilitate a much easier comparison of source vs target tables.

ISTQB
International Software Testing Qualification Board, Internationally most common and known. A test approach on how to setup a structured test process in different software development life cycles (agile, scrum, waterfall or V-model approaches).

L

Logical Data Model (LDM)
A logical data model (LDM) provides an overview of the entire set of data created and maintained by an organization and a diagrammatic presentation of the organization's data.

M

Measures
Measures are the numeric values that users want to slice, dice, aggregate, and analyze; they are one of the fundamental reasons why you would want to build OLAP cubes using data warehousing infrastructure.

Meta data
Metadata is a description of a digital asset, 'data about data'. Terms associated with an object to describe it, for instance the date a record was created or added, the source where the information comes from, the type or category of the data.

Minus
A querie that can facilitate a much easier comparison of source vs target tables.

Multidimensional Expressions (MDX)
A query language for online analytical processing (OLAP) using a database management system. Much like SQL, it is a query language for OLAP cubes. It is also a calculation language, with syntax similar to spreadsheet formulas.

N

NoSQL
Non-relational database systems or so called NoSQL ("Not only SQL") databases.

O

OLTP
Operational systems often also called On Line Transaction Processing (OLTP), is used for support of primary business activities like processing customer requests, invoice handling, processing lab results, planning courses and filling or filing tax forms (mainly 'writing' and 'updating' data).

OLAP
On Line Analytical Processing (OLAP) is used to create value from the existing business data, by judging measures like company results over several dimensions like Time, Pricing Categories, Regions and Customer Groups (mainly 'reading' data).

OLAP Cube
An OLAP cube, also known as multidimensional cube or hypercube, is a data structure which offers ways to display and sum large amounts of data and provide searchable access to any data points.

Outlier detection
Outliers are extreme values that deviate from other observations on data, they may indicate a variability in a measurement.

P

Partition
A section of a storage device, such as a hard disk drive or solid state drive.

Performance Testing
Performance testing may be introduced to validate that the cubes can be loaded within the time window that is made available.

Predictive Analytics
With Predictive Analytics, using data from the past to create models the future is predicted.

Prescriptive Analytics
In prescriptive analytics data can have a correlation with each other, for instance where two or more things (or data elements) have a mutual relationship or connection, or causation, that is the influence of causes on effects in a later state of the data.

Pseudonymization
The process of obscuring data with the ability to re-identify it later is also called pseudonymization.

Q

Quality
Source: ISTQB
The degree to which a component or system satisfies the stated and implied needs of its various stakeholders.

Quality Characteristics
Source: ISTQB
A category of quality attributes that bears on work product quality.

Queries
A request for data or information from a database table or combination of tables.

R

Raw Data
Primary data directory from the source without any transformations done before use. Think of data from a source, but also social media feeds in its rough form.

Report
Source: Madan
A document that presents information in an organized format for a specific audience and purpose.

Relational database
A relational database is a digital database based on the relational model of data.

Reporting Tester
The Reporting Tester will focus mainly on the end user product, interacting with the developer of reports and dashboards and with the business.

Reports Tester
See Reporting Tester

Risk Based Testing
Source: ISTQB
Testing in which the management, selection, prioritization, and use of testing activities and resources are based on corresponding risk types and risk levels.

S

Scrum
Source: scrum.org
Scrum is a lightweight framework that helps people, teams and organizations generate value through adaptive solutions for complex problems.

Self-service BI
Tools that are used for OLAP and Reporting are grouped among the “Self Service BI (SSBI)” tools, tools that provide querying features for (business) end users.

Semantic test
Source: TMAP body of knowledge
The semantic test, together with the syntactic test, belongs among the validation tests, with which the validity of the data input is tested. In practice, the semantic test is often executed in combination with the syntactic test.

Slicing and Dicing
Slicing and Dicing refers to a way of segmenting, viewing and comprehending data in a database. Large blocks of data is cut into smaller segments and the process is repeated until the correct level of detail is achieved for proper analysis.

Slowly Changing Dimensions (SCD's)
Source: Kimball, Ralph; Ross, Margy. The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling
Slowly changing dimension is a dimension which contains relatively static data which can change slowly but unpredictably, rather than according to a regular schedule.

Source
Data that is used as operational data and forms the basis for the business intelligence process.

Snowflake
See Dimensional model

Staging
An intermediate storage area used for data processing during the extract, transform and load process.

Star scheme
See Dimensional model

STM
Source to Target Mapping, a mapping document that describes names of source and target tables, names (and data types) of columns in the source- and target tables, how to detect Inserts, Updates and Deletes and transformations of data.

SQL
Structure Query Language (SQL) is a language to operate databases; it includes database creation, deletion, fetching rows, modifying rows, etc. SQL is an ANSI (American National Standards Institute) standard language.

Syntactic test
Source: TMAP body of knowledge
The syntactic test, together with the semantic test, belongs to the validation tests, with which the validity of the input data is tested. This establishes the degree to which the system is proof against invalid, or 'nonsense' input that is offered to the system willfully or otherwise. This test is also used to test the validity of output data.

T

Transformations
The change that occurs on data from a source table to a target table. E.g. the concatenation of values, Source-columns that are combined in a single target column, aggregations or decoding conversions for integration of data or easier readability for end users.

Transformation Testing
Testing of the transformations that data transforms to from source to target situations.

Testing
Source: ISTQB Syllabus
The process consisting of all lifecycle activities, both static and dynamic, concerned with planning, preparation and evaluation of software products and related work products to determine that they satisfy specified requirements, to demonstrate that they are fit for purpose and to detect defects.

Test leader
Source: ISTQB
On large projects, the person who reports to the test manager and is responsible for project management of a particular test level or a particular set of testing activities.

Test process
Source: ISTQB
The set of interrelated activities comprising of test planning, test monitoring and control, test analysis, test design, test implementation, test execution, and test completion.

Test strategy
Source: ISTQB
Documentation aligned with the test policy that describes the generic requirements for testing and details how to perform testing within an organization.

Test techniques
Test Techniques are methods to derive test cases from source documentation in a structured way.

TMAP
A test management approach on how to setup a structured test process in different software development life cycles (agile, scrum, waterfall or V-model approaches).

U

Unicode
Unicode is an attempt by ISO and the Unicode Consortium to develop a coding system for electronic text that includes every written alphabet in existence.

V

Variability
Variability is often confused with variety, but where variety is concerned with the way data is stored and presented, variability is concerned with the actual meaning of the data.

Value
The value is in the analyses done on that data and how the data is turned into information and eventually turned into knowledge, which could be used for learning objectives or just to create more knowledge of a certain subject.

Variety
Variety refers to the many different formats of data. In the past, all data that was created was structured data: it fitted neatly in columns and rows and in text files.
Velocity
Velocity refers to the speed at which the data is created, stored, analyzed and visualized.

Veracity
Organizations need to ensure that the data is *correct* as well as the analysis performed on the data are correct.

Visualization
A graphical representation of data in such a way that the recipients understand the message of the story.

Volume
Volume refers to the amount of data is created, stored, analyzed and visualized.

V-model
The V-Model demonstrates the relationships between each phase of the development life cycle and its associated phase of testing.

W

Waterfall
The waterfall model is a breakdown of project activities into linear sequential phases, where each phase depends on the deliverables of the previous one and corresponds to a specialization of tasks.

Index

