Author: Orlando Lopez

Published on IVT Network (www.ivtnetwork.com)

GXP Volume 25, Issue 4 – July 2021

IVT NETWORK
INSTITUTE OF VALIDATION TECHNOLOGY
an informa business

PEER REVIEWED

E-RECORDS INTEGRITY – DATA WAREHOUSE AND DATA MART LAYER

INTRODUCTION

Once a vast quantity of data is being generated and stored, it is becoming important to preserve the integrity of the information that's collected. Understanding the basics of data integrity (DI) and how it works is the initial step in retaining the reliability of the data and keeping it safe. This article provides the DI issues in a Data Analytics environment.

HMA and EMA Regulatory Activity

The regulatory agencies are noticing the value of Big Data to medicine development. As an example, since 2017 the Heads of Medicines Agencies (HMA) and the European Medicines Agency (EMA) Task Force on Big Data are operating to find best practices and opportunities to increase the utility of big data in regulatory activities, from data reliability through study methods to assessment and decision-making (1). Data quality is one of the best practices under review in this task force.

Source: Oracle Corporation.

Big Data Analytics Configuration

In Figure 1 (2), Big Data Analytics are central repositories (databases) of data from one or more disparate sources. These sources are depicted in Figure 1 as Source Systems. The electronic records (e-records) stored in the source systems are manipulated, cleansed and loaded to the Big Data Analytics.

From the context of the Big Data Analytics environment, the raw data (3) is extracted from its Operational Application or source repository for e-records (4) locations. The raw data may be manipulated and/or cleansed by applying rules or executing programmatic functions. The raw data is converted to data (5) and then loaded into a final set of tables, called Data Marts, for the consumption of users.

The transportation of data from the source systems to a storage medium where it can be accessed, used, and analyzed by an organization is also known as data ingestion. The destination is typically a data warehouse, data mart, database, or document store.

In the source systems the correct data quality and, the precise manipulation and cleansing of the data are the two basic elements of these decision-making systems. DI is an element of the data quality (6). In a data warehouse and data mart environments, the e-records handling function ensures that data is stored, archived or disposed of safely and securely (7).

To set up the processes and infrastructure to handle the regulated e-records, a set of technical and procedural controls must be implemented to maintain the reliability of each e-record after the e-record is created.

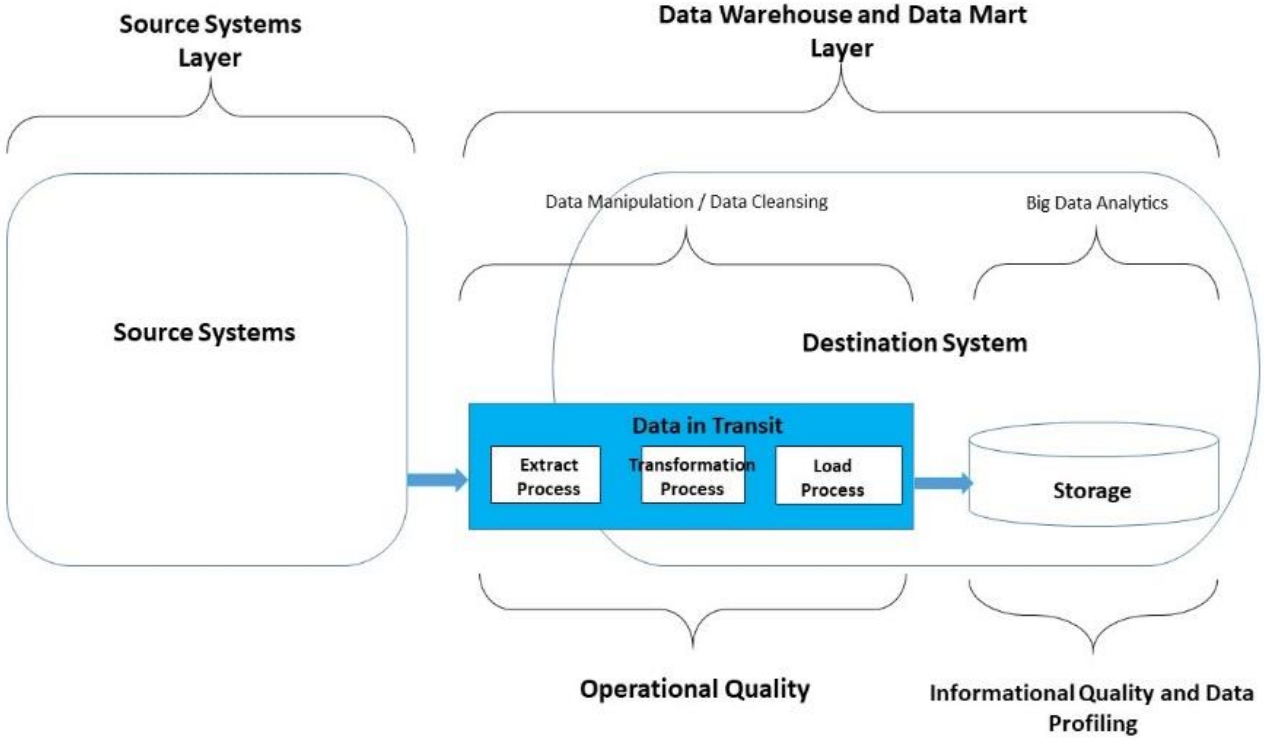


Figure 1 – Big Data Simplified Architecture

It is necessary to know where the e-record resides and understand the vulnerability (8) of each e-record and implement the correct e-records integrity-related controls.

Downstream reporting and analytics systems rely on reliable and accessible data. This paper describes the e-records integrity controls to assure the reliability of the e-records residing in the data marts.

It is supposed to on-premises infrastructures environment. In a cloud data warehouse environment, instead of an ETL is used a cloud-native ELT. As an example, a cloud-native ELT is built to leverage the best features of a cloud data warehouse such as parallel processing of many jobs at once.

The e-records integrity elements do not change if using an on-premise data warehouse or cloud data warehouse. The required e-records integrity controls in a data warehouse and data marts applications are safeguarded in four spaces: data creation, data in storage, during processing, data while in transit (9).

PROCESSING ENVIRONMENT

My first paper about Big Data e-records integrity covers data warehouses and business intelligence (BI) (10).

The scope of this paper is the e-records integrity in a Big Data-based processing environment. As depicted in Figure 1, this environment consists of data warehouses (11) and data marts (12) layers. These layers include data manipulation, data cleansing or data cleaning, and big data analytics. In the data manipulation and data cleansing levels, it includes Operational Quality. In the area of big data analytics, it includes Informational Quality and Data Profiling.

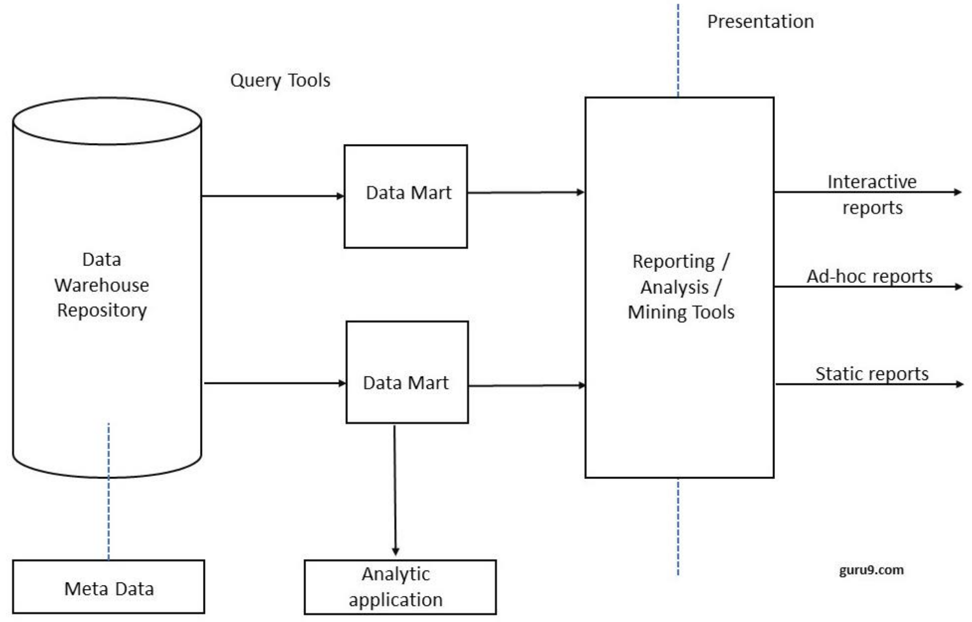


Figure 2 – Big Data Analytics Architecture (Courtesy of guru.com)

Figure 1 shows a single repository for e-records, but in effect, the warehouse and data mart are multiples repositories that are logically associated and represented as one big physical repository for e-records. Figure 2, provides a detailed architecture related to Big Data Analytics. Each section in this analysis includes the e-records vulnerabilities, a flowchart, and examples of the controls that can be implemented to mitigate each vulnerability.

E-RECORDS INTEGRITY

Consistent with any typical electronic systems (13), the required e-records integrity controls in a data warehouse and data marts applications are safeguarded in four spaces: data creation, data in storage, data during processing, and data while in transit (14). Refer to Figure 3.

As applicable, these controls must be integrated as part of the associated data warehouse and data marts application requirements document.
In addition to the typical technical and procedural DI controls delineated in the DI regulatory guidelines, the following databases DI controls (15) must be implemented in the big data environment databases.

If the integrity of data is maintained, it means that data values stored within the database are consistent concerning the data model and/or data type.

Entity Integrity - It depends on the making of primary keys or exclusive values that classify data items. The purpose is to make sure that data is not recorded multiple times, and the table has no null fields. Entity integrity is a critical feature of a relational database that stores data in a tabular format, which can be interconnected and used in a range of ways.

Referential Integrity - Referential integrity is a property of data stating that all its references are valid. In the context of relational databases, it requires that if a value of one attribute of a relation references a value of another attribute, then the referenced value must exist.

Domain Integrity - It's a collection of procedures that ensures the precision of every data item is maintained in a domain (16). Domain integrity encompasses rules and other processes that limit the format, type, and volume of data recorded in a database. It ensures that every column in a relational database is in a defined domain.

User-Defined Integrity - It comprises the rules defined by the operator to fulfil their specific requirements. At times entity, referential, and domain integrity are not enough to refine and secure data. Repeatedly, particular business rules must be considered and integrated into DI processes to meet enterprise standards.

SOURCE SYSTEMS (17)

In the context of Figure 1, the first set of e-record controls to discuss are related to source systems in big data architecture. It includes e-record controls between the Source Systems Layer interface and the Staging Area. The Source Systems Layer are heterogeneous sourced repositories. The majority of the source systems are repositories

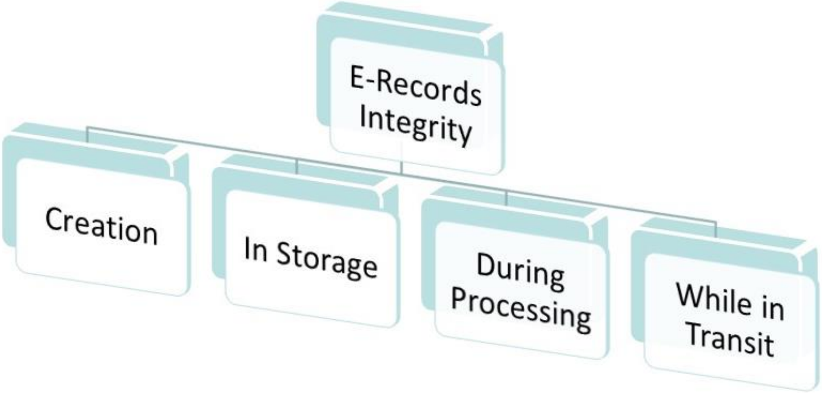


Figure 3 – E-records Integrity NIST

containing only e-records (18), including software as a service (SaaS). There are other source systems that, as part of the data in e-records, may contain other types of data (e.g., flat files, spreadsheets, or even information scraped from the internet).

On a typical electronic source system, e-records are stored and the data, as part of the e-records, can be processed to generate results. These are catalogued as the highest electronic systems (19). The vulnerabilities to these systems are higher than those of systems with lower categorization.

In the context of the Data Warehouse, there are two types of sources of raw data: data or e-records in the source systems and e-records that are created by the integration of e-records during the Data in Transit stage. Refer to Figure 1.

All source systems need to ensure the accurate (20) data at the time the information is ingested from the source system(s) and therefore available to the Transformation Process.

Source Systems Vulnerabilities

The main vulnerability of e-records in storage and the associated metadata is the accidentally or maliciously alteration, deletion, loss, re-creation or deliberate falsification of e-records, and the likelihood of detection of such actions.

Other vulnerabilities to source systems are:

- Infrastructure, not qualified.
- Media failure.
- Data management functions (21), not validated.
  - Write and/or retrieve incomplete records.
- General users having access to critical aspects of the software, e.g., system clocks, file deletion functions, and so on (22).
- Security levels scheme not ensuring segregation of duties.

- Not having access to the stored data in a readable format.
- Not having access to the audit trails and/or metadata.
- No audit trails functionality.
- No security levels functionality.
- No backup procedural control.
- Backup and restore functions not qualified.
- The system clock is not used to generate timestamps.
- Manipulation to the system clock.
- No reliable clock to the synchronization to the system clock.
- The memory containing data in transit allows edits to the data.

Source Systems E-Records Integrity Controls

The main control to e-records in storage is restricting the access to e-records and associated metadata to authorized personnel only. User access controls, both physical and logical, shall be configured and enforced to prohibit unauthorized access to, changes to and deletion of data. Individual Login IDs and passwords shall be set up and assigned for all staff needing to access and utilize the specific electronic system. Shared login credentials do not allow traceability to the individual who performed the activity; for this reason, shared passwords (even if justified for reasons of financial savings) must be prohibited (23).

If the access to the source system is not controlled, the e-records can be accidentally or maliciously modified, altered, or destroyed. Even people with access authorized to e-records should have access only to the e-records that are appropriate for their job role, and that actions are attributable to a specific individual.

As a function related to security, e-records integrity service maintains information exactly as it was inputted and is auditable to affirm its reliability. The e-records integrity service is implemented by a procedural control in which the scope is the data management functions such as creation, data review, data security, data reports, and data availability. The software infrastructure implementation of these functions must be qualified.

Any manual data entry to a source system should ensure that data accuracy (24). The interface between the manual data entry, data acquisition and recording systems shall be qualified to ensure the accuracy of data. During the system operational phase, an accuracy check (EU Annex 11 p6) to critical manual data entry may be executed by a second person or by a validated e-system.

Manual data entry shall be saved into volatile memory in a format that is not vulnerable to manipulation, loss or change (25). The time intervals before saving data should be minimized. Systems should be designed to require saving data to permanent memory (e.g., a repository for e-records) before prompting users to make changes.

The system shall rely upon appropriately controlled/synchronized clocks for recording timed events to ensure reconstruction and traceability, including information of the time zone where this data is used across multiple sites. The regulated user may store its e-records in the data as a service (DaaS) cloud and accesses those e-records through program interfaces. Data governance measures by a regulated user may be compromised by untrustworthy security service provided by the service provider. A service level agreement (SLA) shall be in place between the regulated user with the service provider capturing the responsibilities of the service provider, including the e-records integrity service and the associated security.

The initial and the periodic audits to the service provider and the DaaS environment must comprise DI risks and appropriate control measures.

Related to the interface between the Source Systems Layer and the Data Warehouse and Data Mart Layer, this shall be qualified to ensure the accuracy of data. After the deployment of the referenced layers, built-in checks (EU Annex 11 p5 and US FDA 21 CFR Part 211.68(b)) should be incorporated to ensure the accuracy of the exchanged data (I/Os) and completeness of data acquired, as any metadata associated with the data. The built-in checks maximize the mitigation associated with the inputs and outputs (I/Os) errors and recovery of lost data or corrupted data. Data while in transit shall be protected to avoid manipulation, loss or change. The data while in transit is stored as a temporary local file before transfer to a permanent storage location (e.g. server). During the period of 'temporary storage, there is often limited audit trail provision amending, deleting or recreating data. This is a DI risk. Reducing the period that data is stored in temporary memory reduces the risk of undetected data manipulation (26). A checksum is one of many implementations of a built-in check to detect errors that may have introduced during a file's transmission. Other built-in checks are encryption and secure transfer.

The impact on network-based technologies is that insufficient error checking at the point of transaction entry can result in incorrect transaction processing and DI risks. Integrity can be lost when data is processed incorrectly, or when transactions are incorrectly handled due to errors or delayed processing. As in any repository, the source systems shall rely upon appropriately controlled/synchronized clocks for recording timed events to ensure reconstruction and traceability, including information of the time zone where this data is used across multiple sites.

In addition to all of the above interface-related controls, computer infrastructure, network, and interfaces components must be verified periodically to ensure correct communication between components (27). These controls must be part of the management of physical media.

EXTRACT PROCESS

The intermediate storage location, also known as the data staging process, contains the Data Manipulation / Data Cleansing and has three major steps: extraction, transformation, and load (ETL) processes. Refer to Figure 4. The ETL provides a consolidated view of data, for easier analysis and reporting.

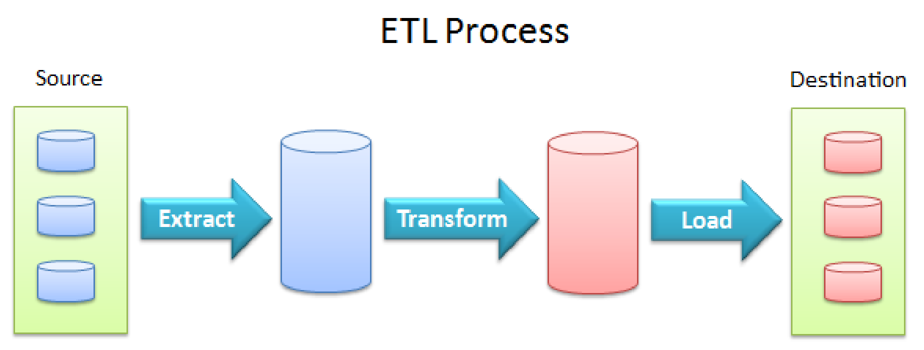


Figure 4 – Simplified ETL Process (Source: Applied Informatics, Inc.)

ETL is built for dealing with large volumes of data and this is ideal for moving data into a Data Warehouse or Data Mart but can certainly have other uses. The ETL is metadata-driven. The e-records created in the data staging come from the source systems. These e-records are accessed and transformed by the ETL and sent to the Data Warehouse.

This section covers the Extract Process. Data extraction is the process of obtaining data from an operational source system so that it can be replicated to a secondary data repository (28) designed to support online analytical processing. Refer to Figure 5. The copy of data stored in a secondary data repository is also called persisted data. The persisted data must contain appropriate metadata.

The persisted data must come from a source system that contains, at a minimum, a regulatory status equal to that of the status indicates for the secondary data repository. Persisted data should never be modified directly in a secondary data repository. If technical limitations require the data to be modified in the secondary data repository, the modification should have traceability to the same modification in the source system.

In ETLs some data do not require any transformation at all; such data are known as "direct move" or "pass-through" data. In the case a transformation is not performed, the key element to consider in ETLs is ensuring that each valid record from the source is loaded to the appropriate target table(s).

Extract Process Vulnerabilities

One of the vulnerabilities to the extract process to e-records is to have the incorrect representation of the creation and movement of data through the process including documentation of the systems used. The mapping of the Extract Process identifies data elements that are to be obtained from the associated source system.

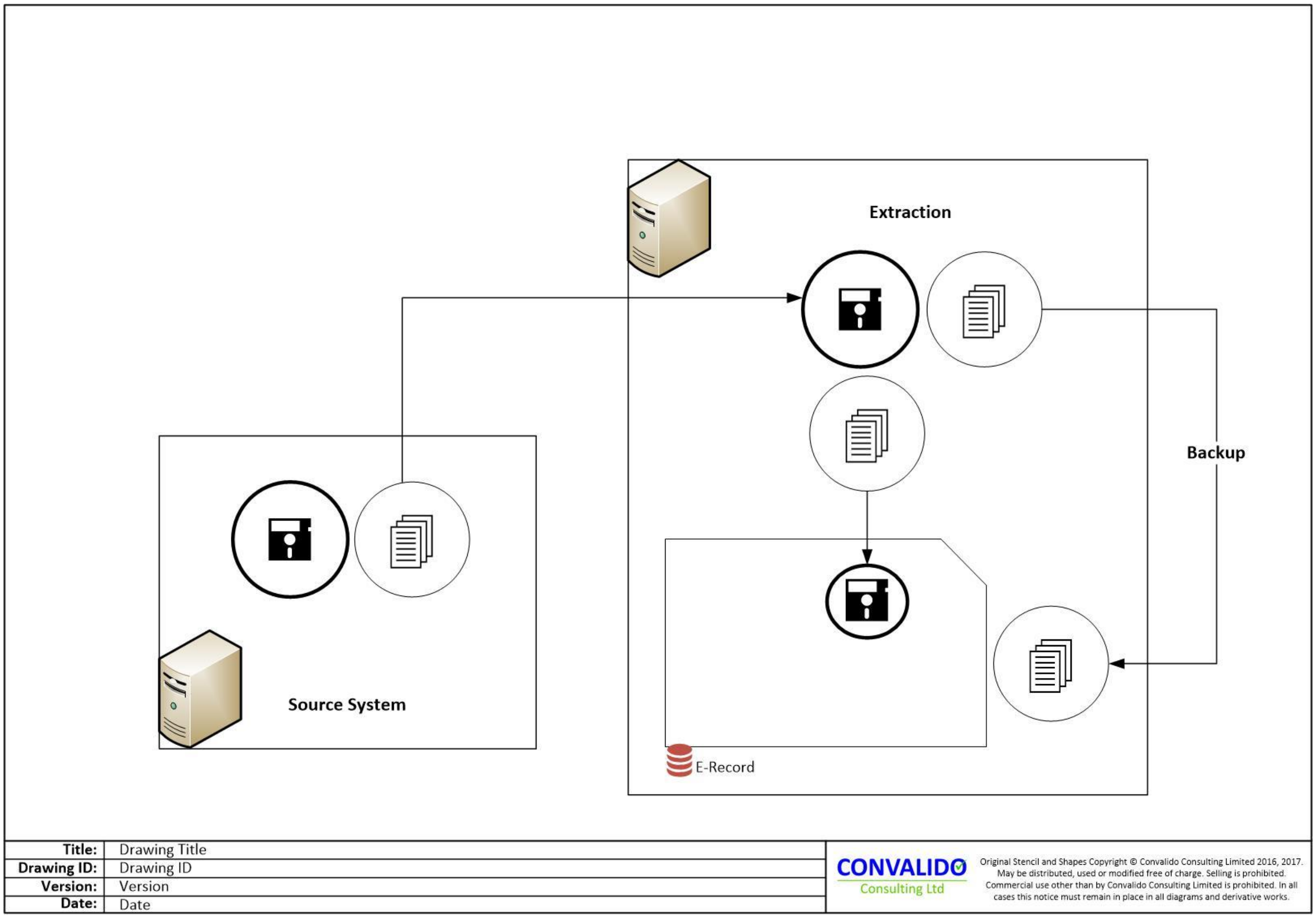

Extraction
Backup
Source System
E-Record
Title: Drawing Title
Drawing ID: Drawing ID
Version: Version
Date: Date
CONVALIDO
Consulting Ltd
Original Stencil and Shapes Copyright © Convalido Consulting Limited 2016, 2017. May be distributed, used or modified free of charge. Selling is prohibited. Commercial use other than by Convalido Consulting Limited is prohibited. In all cases this notice must remain in place in all diagrams and derivative works.

Figure 5 – Data Extraction

As the result of an inappropriate mapping of data processes, the extract process always exposes data quality issues that have been hidden within the operational source systems. Since data quality significantly impacts data warehouse credibility, it is needed to address these data staging related quality problems.

Data may be stored as a temporary local file before transfer to a permanent storage location (e.g. server). During the period of 'temporary storage, there is often limited audit trail provision amending, deleting or recreating data. This is a DI risk.

As depicted in Figure 5, Data Extraction, there are two additional key elements to consider. The first element is the DI controls associated with the data while in transit from the source system to the data warehouse. The second element to consider is the DI security controls of the data in storage.

Extract Process Integrity Controls

The data lineage is an analysis exercise that is verified during the design reviews. An automated tool is the best method to avoid manual errors. To correct any mapping error introduced during the Project Phase, the automated tool is executed to the deployed system. Any inconsistency between the mapping and the deployed system is evaluated and it is decided which element is correct.

As part of the data lineage, it is extremely relevant maintaining the correct mapping of data processes. The particular mapping must be part of the change control impact assessment. In addition to the data lineage, the interface between the two servers must be qualified. If technically feasible, a built-in check method must be implemented to secure the exchange of data between servers.

After the data is in the secondary data repository, the security-related procedures describe the physical and logical access controls to be implemented to the data. A system needs to be in place to control unauthorized access to systems.

To reduce the risk of losing the data in storage and guaranteeing data readiness for the users, periodic backups must be performed to guarantee that data is retrievable, reproducible and unaltered for the retention period of the record. Removing the use of temporary memory (or reducing the period that data is stored in temporary memory) reduces the risk of undetected data manipulation. (EMA Q&A GMP Data integrity, Aug 2016)

For extracted e-records, the system architecture needs to be accurately reviewed ensuring that the data is effectively timestamped, including the time zone as appropriate, and written to the corresponding repository for e-records (30). This DI control is in particular applicable to e-records to be written to the Source Systems, Staging Area, Warehouse, and Data Marts. Other DI controls associated with records in storage are Archiving and Migration. These are covered elsewhere (31).

TRANSFORMATION (32) PROCESS

After the extraction of the data from the source system(s), data may need to be verified or transformed to make it suitable for use in records, such as running analytics queries, running machine learning flows, or even just storing a subset of the data in a database. This may be involved activities to cleanse, scale, normalize, or lock data for its intended use (34). Transformation refers to the cleansing and aggregation that may need to happen to data to prepare it for analysis. Refer to Figure 6.

The typical architecture of the transformation process is the multistage data transformation. The extracted data is moved to a staging area where transformations occur before loading the data into the Data Warehouse. These records are transformed, and information may be inserted, updated, and/or deleted. Another example of transformation is data cleansing. During the data cleansing, it is detected and corrected corrupt or inaccurate e-records from an e-record set, table, or database. Incomplete, incorrect, inaccurate or irrelevant parts of the data are identified and replaced, modified, or deleted the dirty or coarse data. The source system(s) is(are) to be used to the e-records needing updates in the staging area(s).

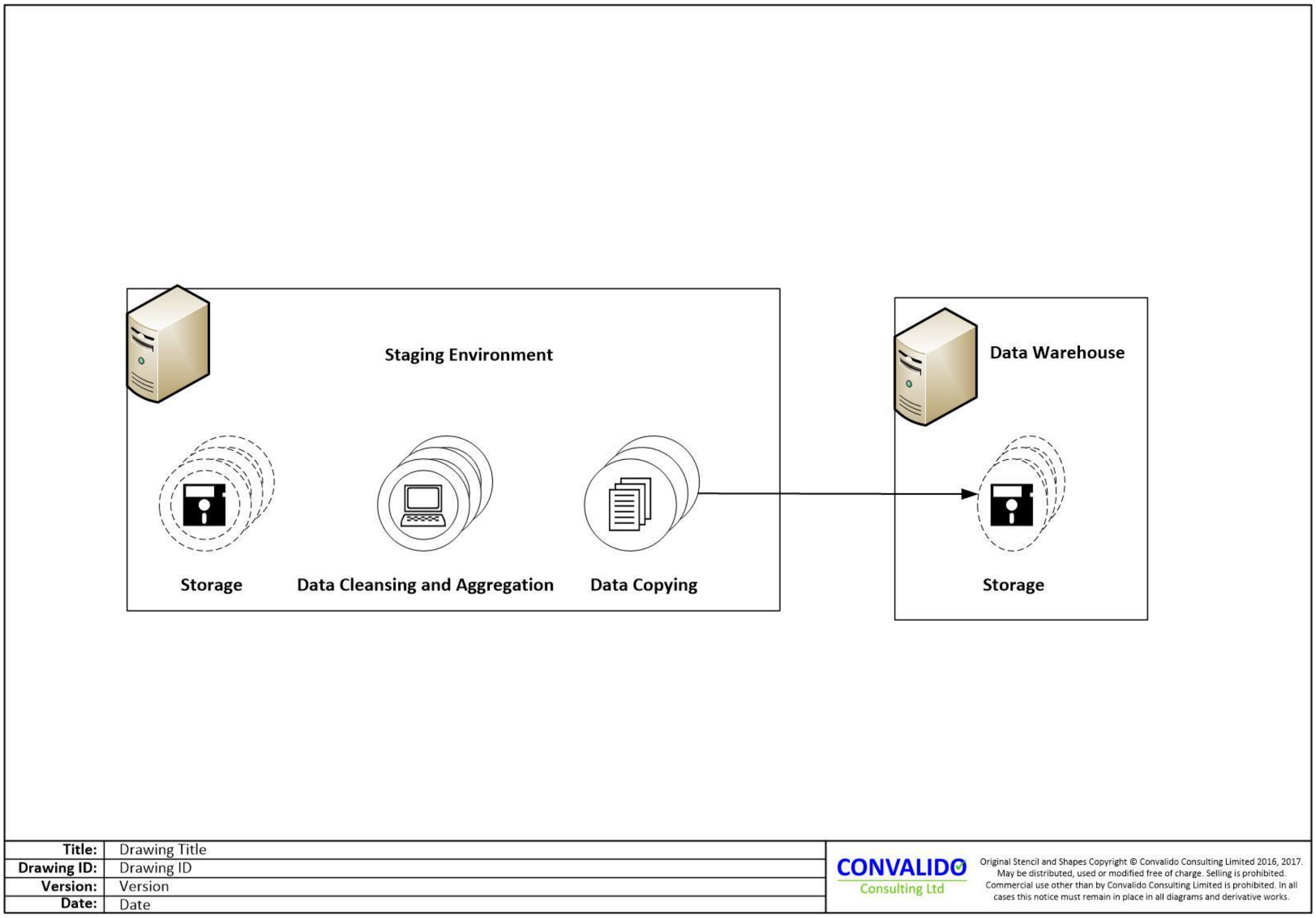

Staging Environment
Storage
Data Cleansing and Aggregation
Data Copying
Data Warehouse
Storage
Title: Drawing Title
Drawing ID: Drawing ID
Version: Version
Date: Date
CONVALIDO
Consulting Ltd
Original Stencil and Shapes Copyright © Convalido Consulting Limited 2016, 2017. May be distributed, used or modified free of charge. Selling is prohibited. Commercial use other than by Convalido Consulting Limited is prohibited. In all cases this notice must remain in place in all diagrams and derivative works.

Figure 6 – Data Transformation

This application interfaces directly with the source repository and, typically, interface as well with the secondary repositories called staging areas.

During the transformation stage, the DI integrity processing takes place as part of a series of rules or functions applied to the extracted data. These rules (e.g., operational checks) and functions are applied to prepare the data to be loaded into the end target (e.g., data mart).

- Basic transformations (33):
  - **Cleansing**: Mapping NULL to 0 or "Male" to "M" and "Female" to "F," date format consistency, and so on.

  - **Deduplication**: Identifying and removing duplicate records.
  - **Format revision**: Character set conversion, unit of measurement conversion, date/time conversion, and so on.
  - **Key restructuring**: Establishing key relationships across tables.
- Advanced transformations (33):
  - **Derivation**: Applying business rules to the data that derive new calculated values from existing data.
  - **Filtering**: Selecting only certain rows and/or columns.
  - **Joining**: Linking data from multiple sources.
  - **Splitting**: Splitting a single column into multiple columns.
  - **Data validation**: Simple or complex data validation – for example, if the first three columns in a row are empty then reject the row from processing.
  - **Summarization:** Values are summarized to obtain total figures which are calculated and stored at multiple levels as business metrics.
  - **Aggregation**: Data elements are aggregated from multiple data sources and databases
  - **Integration:** Give each unique data element one standard name with one standard definition. Data integration reconciles different data names and values for the same data element.

These activities should be suitably managed and documented if the transformed data is to be in records required for regulatory reasons (34).

Depicted in Figure 6, after the e-records reached the staging area, the ETL extracts selected e-records and, before loading to the Data Warehouse (Figure 2), these e-records are transformed and, the information of such e-records may be inserted, updated, and/or deleted. Some data do not require any transformation at all. This data is extracted from the source system and loaded to the Data Warehouse.

Transformation Process Vulnerabilities

During the transformations, data is processed by configuring applications. The main vulnerability to the Transformation Process is not executing these programs in the predefined order. The incorrect input to the transformation programs is another vulnerability in this process.

Transformation Process Integrity Controls

The implementation of the transformation processing must be performed via a system development life cycle, assessed via a validation process and associated configuration management after deployment. In the context of the transformation process, the effort is concentrated on the operational checks (21 CFR Part 11.10(f)).

The objective of operational system checks is to enforce the sequencing of steps and events as applicable to the transformation process. The application-dependent algorithms and sequencing of operations be followed within the e-system are encompassed in the computer program(s) which drive the electronic system. These applications dependent is defined in the requirements document, implemented as part of the Project Phase and, executed and maintained during the Operational Phase. The operational system checks would ensure that the proper sequence is followed. The above controls applicable to e-records processing are established, as appropriate, during the implementation of the electronic system and each control is re-evaluated during the periodic reviews.

LOAD PROCESS

When you're done moving your data into a “queryable” state, the newly transformed data is distributed into a new destination, which can be any data storage including a simple delimited flat file or a data warehouse. This can be done using tools, which requires some expertise and coding, or you can develop your orchestration tool by yourself. As depicted in Figure 7, the data is copied to the new storage location at the Data Warehouse.

Load Process Vulnerabilities

As the load phase interacts with a database, the constraints defined in the database schema apply, which also contribute to the overall data quality performance of the ETL process. Another vulnerability to the Load Process is the incorrect pathway between the staging area to the new destination. Program code to develop/compliment load-related application make it more complex the implementation project. This code must be going through the system development lifecycle (SDLC). Interfaces between Data Warehouse Repository and the Data Marts, present a risk whereby data may be inadvertently lost, amended or transcribed incorrectly during the transfer process.

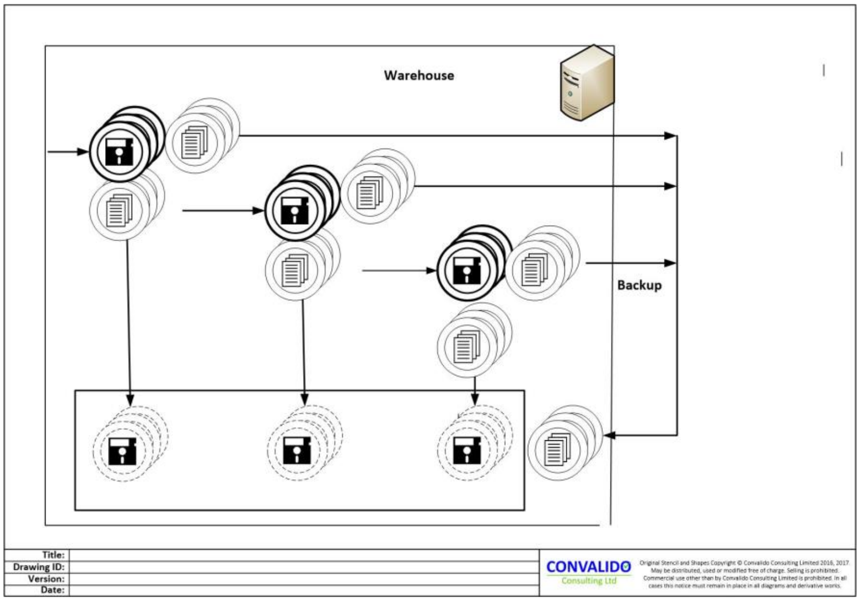

Warehouse
Backup
Title:
Drawing ID:
Version:
Date:
CONVALIDO
Consulting Ltd
Original Stencil and Shapes Copyright © Convalido Consulting Limited 2016, 2017. May be distributed, used or modified free of charge. Selling is prohibited. Commercial use other than by Convalido Consulting Limited is prohibited. In all cases this notice must remain in place in all diagrams and derivative works.

Figure 7 – Data Warehouse

Load Process Integrity Controls

The integrity controls consist of design correct data structures, physical and logical data models, and load mappings. The design and testing must be comprehensive to disallow, as much as possible, the errors introduced by the full development activity. During the Operational Phase, the correctness of data loaded to the data storage area periodically reconcile. The technique and tools to be used in the data reconciliation process, the frequency of data reconciliation, the rationale for the choice of subsets of data to reconcile and documentation of results of data reconciliation must be defined during the design and implemented according to the specifications. Instructions must be provided to the support team on ETL on data reconciliation requirements.

The interface between the originating system, data acquisition and recording systems shall be qualified to confirm the accuracy of the data. The users shall not be able to manipulate the data while in transit, from the staging area to the Data Warehouse.

DATA WAREHOUSE

Data warehouses are central repositories of integrated and/or transformed data from one or more disparate sources. It is used for reporting and data analysis and is considered a core component of BI.

The e-records are created from the staging area. The e-records are accessed or transformed by the ETL and sent to the Data Warehouse.

Most data warehouses have the following characteristics:

- Data access is typically read-only. The most common action is the selection of data for analysis. Data is rarely inserted, updated, or deleted. This is in contrast to most source systems which must be able to handle frequent updates as data is gathered. For more information on source systems, see Source systems for data collection.
- Data is aligned by subjects.
- Data formats are uniformly integrated using an ETL process.
- A data warehouse is populated with data from the existing operational systems using an ETL process, as explained in Extraction, Transformation, and Loading processes.

The way data is deleted from the warehouses is by the update of the persisted data in the staging area as a result of a change in the associated source system. By using the ETL loaded, the deleted data is deleted in the warehouse and, subsequently, the data associated mart(s). All data warehouses need to ensure accurate data quality after the information is loaded from the data staging.

As depicted in Figure 7, the Data Warehouse receive data from different staging areas. This data and associated metadata are saved in the respective e-records as mapped during the design. Backups are performed in case the backups are needed to recuperate accidentally or maliciously alteration, deletion, loss.

Data Warehouse Vulnerabilities

As in source systems, the main vulnerability of e-records in storage and the associated metadata is the accidentally or maliciously alteration, deletion, loss, re-creation or deliberate falsification of e-records, and the likelihood of detection of such actions. Another vulnerability in the Data Warehouse is the incorrect mapping between the staging area to the corresponding data warehouse. In addition, the data may be inadvertently lost, amended or transcribed incorrectly utilizing the interface during the loading process.

Data Warehouse Integrity Controls

As in any other databases in a Data Warehouse and Data Mart environments, the e-records storage areas must be qualified by following a process based on an approved validation/qualification and procedural controls. The validation/qualification should concentrate on the insertion, update, and deletion of e-records, and the sequencing of operations.

The data warehouse must be reasonably secure from intrusion and misuse and must be adhering to generally accepted security principles. The data warehouses are not accessed by users with a direct interest in the data. Only system administrators and developers should have access to the data warehouses. System Administrator rights, including permitting activities such as database amendment or system configuration changes.

All system administrators shall only have access to functionality within the system as required by their job role. User roles and responsibilities shall be pre-determining and documented in controlled documentation. The data warehouse applications will have open read-only access for database structures that do not contain data considered sensitive. Sensitive data will not have open access and will require additional approvals and restrictions for access according to the security plan.

To reduce the risk of losing the data in the storage and guarantee data readiness to the users, periodic back-ups must be performed. Integrity and accuracy of backup data and the ability to restore the data should be checked during the validation and monitored periodically. In addition, the capacity level of the storage must be monitored. Stored data should be checked periodically for accessibility, readability and accuracy. If changes are implemented to the computer infrastructure and/or the data warehouse application, then it is required to ensure and test the ability to retrieve data.

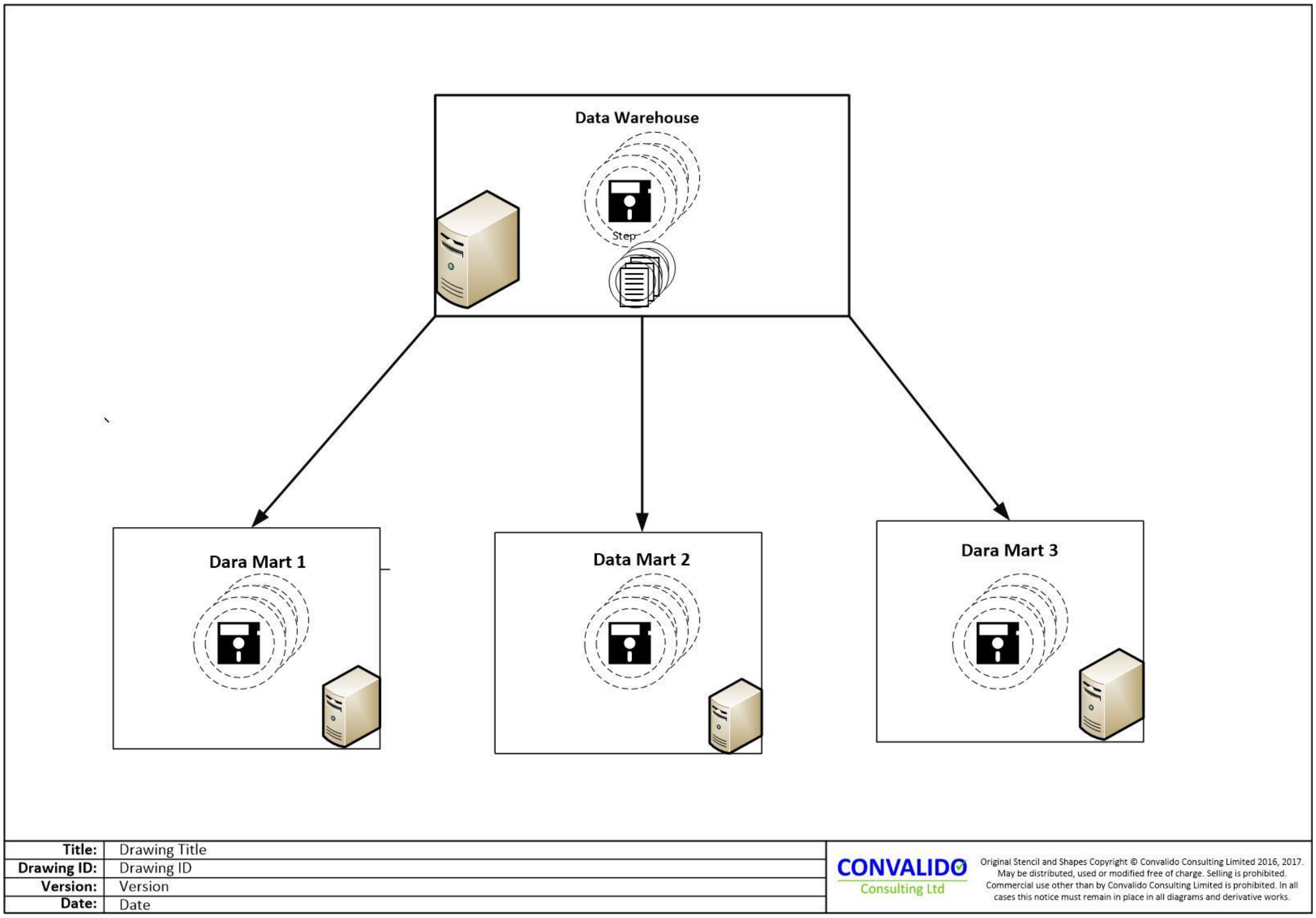
After completing the specified source system record retention requirements is reached, the respective e-records in the secondary repository must be physically deleted. The data warehouse shall rely upon appropriately controlled/synchronized clocks for recording timed events to ensure reconstruction and traceability, including information of the time zone where this data is used across multiple sites.

DATA MARTS

As depicted in Figure 2, a data mart is a structure/access pattern specific to data warehouse environments, used to retrieve client-facing data. A data mart is a subset of a data warehouse oriented to a specific business line. Data marts contain repositories of summarized data collected for analysis on a specific section or unit within an organization, for example, the Quality Control department.

The origins of the data contained in a Data Mart is from the Data Warehouse. The access to the data marts is read-only to the users. Physical and federated data marts will be under a separate schema from the data warehouse schemas. Data marts may be in the same database instance as the Data Warehouse if appropriate, but the default approach will be to keep data marts in an instance separate from the warehouse for upgrade independence.

The federated data mart is used to integrate key business measures and dimensions. The foundations of the federated data mart are the common business model and common staging area.



Title:	Drawing Title
Drawing ID:	Drawing ID
Version:	Version
Date:	Date

Figure 8 – Data Marts

Data Marts Vulnerabilities

All vulnerabilities associated with the Data Warehouse apply to the Data Mart.

Data Marts Integrity Controls

All DI controls associated with the data warehouse apply to the Data Mart.

AUDIT TRAILS

Persisted data is updated as a result of a change in the associated source system. Consequently, secondary data repositories are not required to have an audit trail. If technical limitations require the data to be modified in the secondary data repository, the modification must have traceability to the same change at the corresponding source system.

OPERATIONAL QUALITY

The data ingestion layer is the backbone of any analytics architecture. The data quality as an informational model has given way to operational data quality, being another element of assurance in the ingesting and ETL processes of customers. The data quality informational model allows analysis of a set of data quality requirements and their representation in terms of a conceptual schema, as well as accessing and querying data quality dimensions using a logical schema. It permits data lineage or tracking data from their source, through various manipulations that data can undergo, to their final usage.

As part of the operational quality, data cleansing must essentially start earlier the first step of building the ETL system. It is required to perform an all-inclusive data profiling analysis of the data sources during the up-front planning and design phase. The good data-profiling analysis takes the form of a specific metadata repository describing:

- Schema definitions
- Business objects
- Domains
- Data sources
- Table definitions
- Synonyms
- Data rules
- Value rules
- Issues that need to be addressed

This data profiling is a good quantitative assessment of the original data sources. Data provenance is the confidence of the data source systems. This confidence is established by instituting data quality rules at the source system (36). In case the data provenance is high, then there will be no data profiling activities at the staging area.

During the operations, the correctness of data loaded to the staging area must be periodically reconciled. The technique and tools to be used in the e-records reconciliation process, the frequency of e-records reconciliation, the rationale for the choice of subsets of e-records to reconcile and documentation of results of e-records reconciliation must be defined during the design and implemented according to the specifications. Instructions must be provided to the support team on ETL on e-records reconciliation requirements.

For additional Operational Quality controls, refer to Transformation Process Integrity Controls. The above-described controls improve downstream reporting and analytics systems rely on consistent and accessible data.

Informational Quality and Data Profiling.
"Informational quality" is a measure of the value which the information provided in Big Data Analytics (Figure 1) to the user of that information. "Quality" is often perceived as subjective and the quality of information can then vary among users and uses of the information.

Using the following standard core information quality dimensions, it is ensured the correctness and usefulness of data.

- accuracy - correct, unambiguous, consistent, and complete
- legibility – readable by users
- suitability – implementation of audit trails
- conformity – following standards
- completeness- the appropriate amount of information
- consistency – adherence to a given set of rules
- provenance – assurance of the data quality of the source system,
- timeliness - available when needed
- attributability – traceable to a source
- security - access limitations in place and information integrity maintained
- validity – syntax correctness

After establishing high data quality, data profiling is the process of examining the data quality available from an existing information source (e.g., a database or a file) and collecting statistics or informative summaries about that data. The purpose of these statistics may be to find out whether existing data can be easily used for other purposes.

Data profiling is a detailed analysis of source data. It tries to understand the structure, quality, and content of data and its relationships with other data.

SUMMARY

Figure 9 depicts a summary of the logical organization of the Data Warehouse and Data Mart Layers. In the context of DI, the required e-records integrity controls in a data warehouse and data marts applications are safeguarded in four spaces: data creation, data in storage, during processing, data while in transit.

In the context of Big Data Analytics, the data is created in the source systems. The DI controls associated with this activity include:

- Data must be saved in a durable media at the time of performance to create a record in compliance with CGMP requirements, including, original records and contemporaneous.
- The infrastructure is qualified to ensure the accuracy of created data; and,
- Built-in checks (EU Annex 11 p5 and 211.68(b)) to ensure the completeness of the data created.
- The e-records in Source Systems, Staging Areas, Data Warehouses, and the Data Marts storages:
- Must be periodically replicated in case any of these need to be restored.
- Must implement and maintain security-related controls to all e-records storages as a mean of ensuring e-records protections.
- Except for the e-records in source systems, there is no changes and associated change control applicable to e-records in the staging area, data warehouse, and the data marts.
- E-records migration and archiving processes are established to maintain the reliability of such records during the transfer.
- E-records must be periodically verified for accessibility, readability, and integrity.

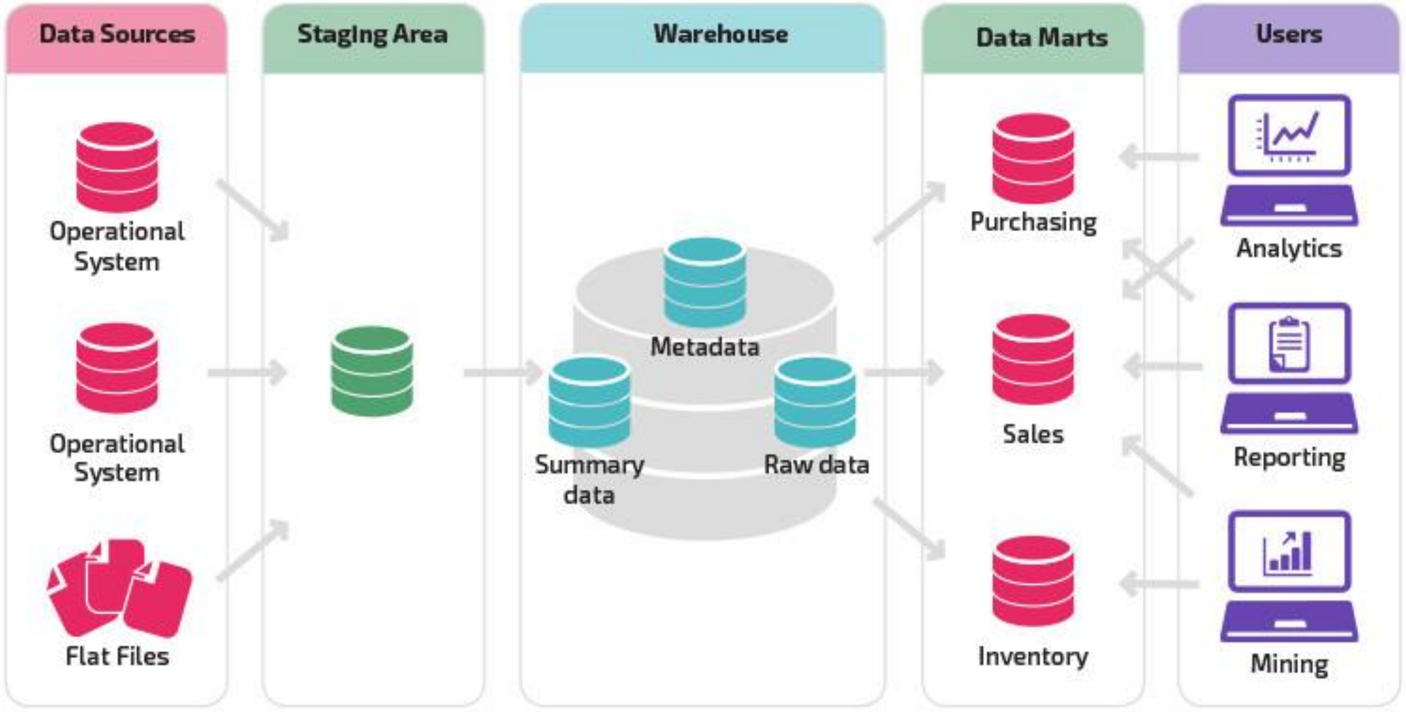


Figure 9 – Data Warehouse and Data Mart Layers (Source: Panoply Ltd,)

The processing in the Big Data Analytics is performed at the Staging Areas and the Data Warehouses. The main DI control during processing is the operational checks. These are used to enforce allowed sequencing. The purpose of performing operational checks is to ensure that operations are not executed outside of the predefined order established by the operating organization. Should it be necessary to create, delete, or modify e-records in a particular sequence, operational system checks would ensure that the proper sequence and processing is followed to minimize DI risks.

Records are in transit on the multiple interfaces in the Big Data Analytics: from data sources to Staging Areas, to the Data Warehouse, and the Data Marts. In addition, there are data movements within the Staging Area. The controls associated with moving e-records must verify that e-records have remained unaltered while in transit. This principle is applicable as well to e-records from creation to reception. E-records integrity controls while in transit consist of:

- record mappings - records mapping defines how to integrate the formats of the source application's data to fit the schema of the destination application. The data mapping provides the link and transfers with each field value in the source data to its corresponding field in the destination application record schema.
- qualification of the infrastructure - appropriate checks should be incorporated into qualification and validation work to ensure the integrity of all data obtained (37).
- built-in checks - computer systems exchanging data electronically with other systems should include, if technically feasible, appropriate built-in checks for the correct computer inputs and outputs (I/Os).

REFERENCES

1. https://www.ema.europa.eu/en/about-us/how-we-work/big-data#hma/ema-big-data-steering-group-section
2. Figure courtesy of Aqtiva.
3. Raw Data - Original records and documentation, retained in the format in which they were originally generated (i.e., paper or electronic), or as a 'true copy'. (MHRA)
4. A repository for e-records is a direct access device on which the e-records and metadata are stored.
5. Data - A basic unit of information that has a unique meaning and can be transmitted.6. López, O., "*Introduction to Data Quality*," in Ensuring the Integrity of Electronic Health Records, López, O., Eds (Routledge, Boca Ratón, FL, 1st ed., 2021), pp. 2019-230.
7. https://ori.hhs.gov/education/products/n_illinois_u/datamanagement/dhtopic.html
8. A vulnerability is a condition or weakness in (or absence of) security procedures, technical controls, physical controls, or other controls that could be exploited by a threat. (NIST, An Introduction to Computer Security: The NIST Handbook, SP 800-12)
9. López, O., "*A Computer Data Integrity Compliance Mode*l", Pharmaceutical Engineering, Volume 35 Number 2, March/April 2015.
10. López, O., "*Electronic Records Integrity in Data Warehouse and Business Intelligence*," in Best Practices Guide to Electronic Records Compliance, O. Lopez, Eds. (CRC Press, Boca Raton, FL, 1st ed., 2017), pp. 341-351.
11. A data warehouse is a system used for reporting and data analysis. These are central repositories of integrated data from one or more disparate sources.

12. A data mart is the access layer of the data warehouse environment that is used to get data out to the users. The data mart is a subset of the data warehouse that is usually oriented to a specific business line or team. Data marts are small slices of the data warehouse.
13. Electronic systems mean systems, including hardware and software, that produce e-records.
14. López, O., “*A Computer Data Integrity Compliance Mode*l”, Pharmaceutical Engineering, Volume 35 Number 2, March/April 2015.
15. Naeem, T., What is Data Integrity in a Database. Why Do You Need It? [Internet]. Astera; 17/05/2021. Available from https://www.astera.com/type/blog/data-integrity-in-a-database/
16. A domain is defined as a set of suitable values that a column is permitted to enclose.
17. Source systems refer to any system or file that captures or holds data of interest. (https://www2.microstrategy.com/producthelp/10.4/ProjectDesignGuide/WebHelp/Lang_1033/Content/ProjectDesign/Source_systems_for_data_collection.htm)
18. A repository to e-records is a direct access device on which the e-records and metadata are stored.
19. CEFIC, “*Practical risk-based guide for managing data integrity*,” March 2019 (Version 1).
20. Accurate Data - Data that is correct, unambiguous, consistent, and complete.
21. The process of ensuring that data is stored, archived or disposed of safely and securely during and after the decommissioning of the computer system. (https://ori.hhs.gov/education/products/n_illinois_u/datamanagement/dmtopics.html)
22. Russian Federal State Institute of Drugs and Good Practices, “Data Integrity & Computer System Validation,” (Draft) August 2018.
23. Russian Federal State Institute of Drugs and Good Practices, “Data Integrity & Computer System Validation,” (Draft) August 2018.
24. Data accuracy means that the data is correct, unambiguous, consistent, and complete.
25. Russian Federal State Institute of Drugs and Good Practices, “Data Integrity & Computer System Validation,” (Draft) August 2018.
26. EMA Q&A GMP Data integrity, Aug 2016.
27. US FDA CPG Sec. 425.400 Computerized Drug Processing; Input/Output Checking.
28. A Secondary Data Repository is any data repository that is used for storing a copy of data obtained from a source record repository. Data Warehouse and Data Marts are examples of s secondary data repository.
29. Data lineage includes the data origin, what happens to it and where it moves over time. Data lineage gives visibility while greatly simplifying the ability to trace errors back to the root cause in data analytics.

30 https://www.gmp-compliance.org/gmp-news/alcoa-what-does-it-mean?utm_source=Newsletter&utm_medium=email&utm_campaign=ECA+GMP+Newsletter+-+2021+-+KW20+-+CEU

31. López, O., “*Electronic Records Controls: Records Retained by Computer Storage*,” in Ensuring the Integrity of Electronic Health Records, López, O., Ed. (Routledge, Boca Ratón, FL, 1st ed., 2021), pp. 169-177.

32. Transformation is an intentional alteration of the appearance or meaning of the data. Transformational includes aggregation, filtering, reformatting and so on.

33. https://www.stitchdata.com/etldatabase/etl-transform/

34. ISPE/PDA, "*Good Practice and Compliance for Electronic Records and Signatures. Part 1 Good Electronic Records Management (GERM)*". July 2002.

35. https://www.gqjournal.com/2010/10/greg-margolin-every-day-we-as-a-society-depend-more-and-more-on-information-technology-to-perform-our-regular-activit.html

36. López, O., "*Data Quality*," in Ensuring the Integrity of Electronic Health Records, López, O., Ed. (Routledge, Boca Ratón, FL, 1st ed., 2021), pp. 219-228.

37. EudraLex, The Rules Governing Medicinal Products in the European Union Volume 4, EU Guidelines for Good Manufacturing Practice for Medicinal Products for Human and Veterinary Use, Annex 15: Qualification and Validation, March 2015.