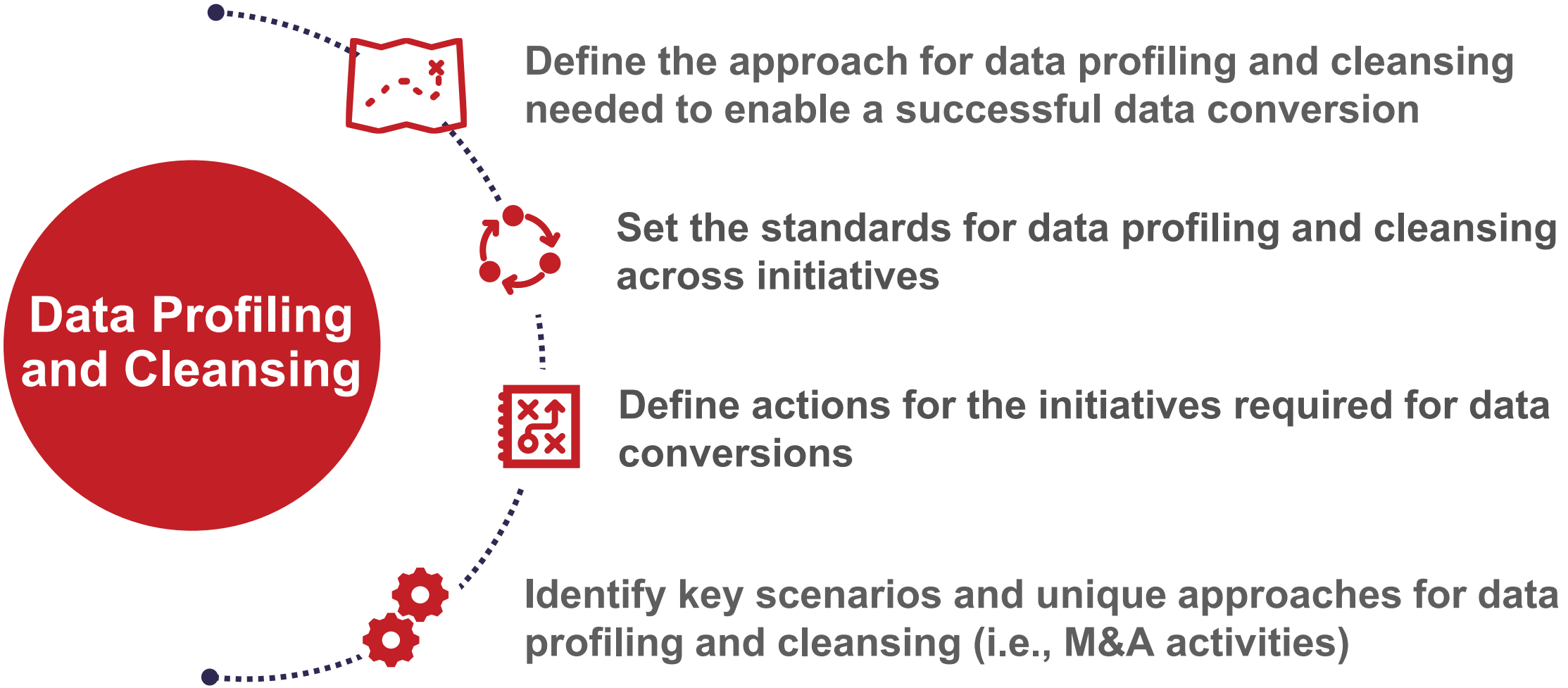
Data Profiling and Cleansing Conversion COE

Contents

Data Profiling/Cleansing - Objectives

The objectives of the profiling and cleansing approach document are outlined below:

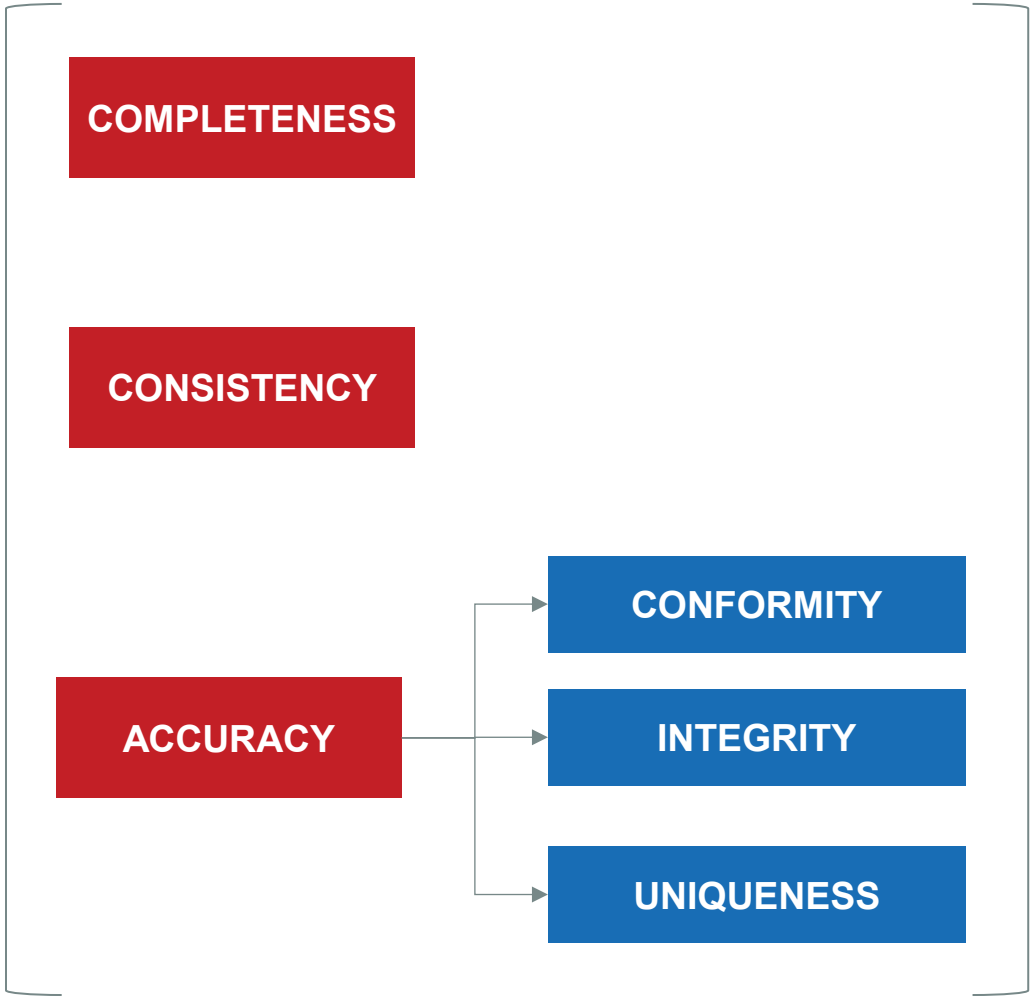


Contents

Data Profiling and Cleansing – Purpose

Data profiling and cleansing aim to improve the overall data quality and underpins the success of a migration/conversion initiative based on several quality dimensions

- **Completeness** – Degree to which the complete data set is available, or the fields are populated
- **Consistency** – Degree to which data is the same in its definition, business rules, format and value across systems
- **Accuracy** – Degree to which the information contains errors and meets business rules



- **Conformity** – Degree to which data is stored in defined format
- **Integrity** – Degree to which data relationship linkages are present
- **Uniqueness** – Degree to which data records are not duplicated / redundant

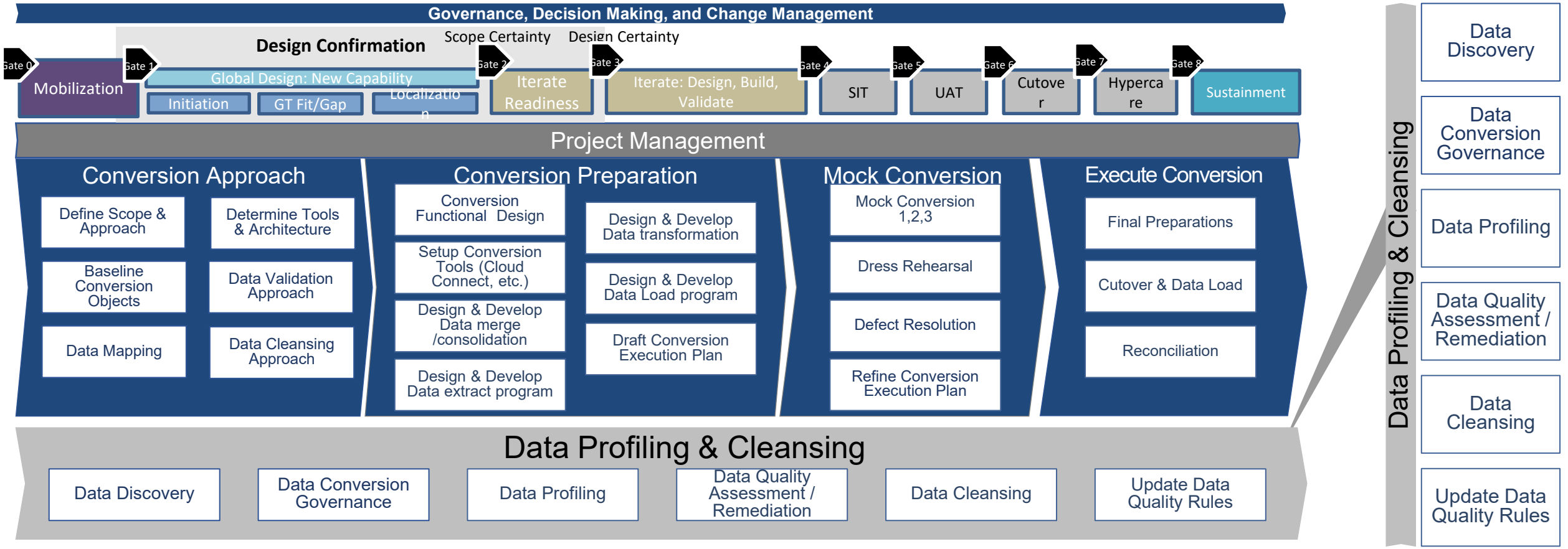
Data Profiling and Cleansing – Quality Dimensions

Data profiling and cleansing aims to improve the overall data quality and underpins the success of a migration/conversion initiative based on several quality dimensions. These dimensions will provide examples into the data quality rules:

Dimension	Description	Questions	Metrics Used
Uniqueness	Data is unique and not duplicated in the same database	What data records or attributes are duplicated?	Percent of records or data elements that are duplicated or non-unique
Completeness	Fields that are mandatory to have a value	What data is missing or unusable?	Percent of data having values entered into them
Accuracy	Default value established or data is valid compared to an external database, i.e. USPS, Royal Mail, HMRC	What data is incorrect or out of date?	Percent of values that are correct
Integrity	Data attributes are maintained correctly to ensure that all data can be traced and connected	What data values or relationships are missing or not referenced?	Percent of data values or relationships missing or not referenced.
Consistency	Agreement or logical coherence among data that frees it from variation or contradiction. Consistency of data across data sets.	What data values give conflicting information? What data values don't match across data sets?	Percent of matching value conditions or derived conditions satisfied
Conformity	Data items that are in accordance to standard formats, i.e. length, data type, format	What data is stored in a non-standard format?	Percent of data values violating standard formats or structure

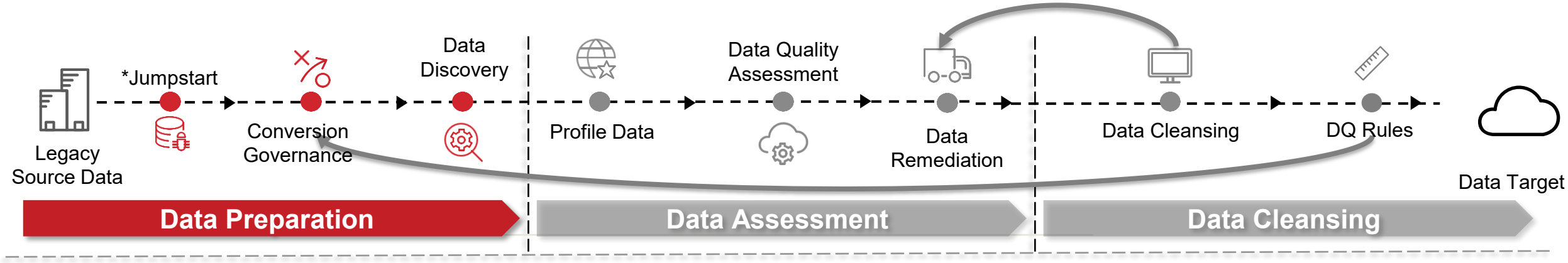
Data Profiling/Cleansing – Methodology

Highlighted below is the overall data conversion methodology with a specific call out to the components of data profiling and cleansing.



Data Profiling/Cleansing – Data Preparation

The primary purpose of Data Assessment is to identify the team members (conversion governance) and data needs (data discovery) for a successful conversion



Inputs

Key stakeholders

Pre-discovery data collection
- Data Conversion designs (if complete)
- Data security requirements

Current state of data quality

Current data quality KPIs & metrics

Previous data quality rules and profiling results

Team Involvement

Data Steward(s) / Data Owner - Provides recommendations on availability and access to business data across the enterprise. Identifies, defines, and analyzes how information assets drive business outcomes

Legacy System SME(s) / IT - Extracts data and prepares for data discovery and profiling

Data Owner - Provides input in preparation for discovery and profiling. Defines data quality rules across the business

Data Analyst(s) - Provides feedback for the data profiling and cleansing; involved throughout data profiling and cleansing

Outputs

Prioritization of data objects and focus areas

Named resources to work on data profiling and cleansing

Identified tools for profiling

Legacy data sources for profiling

Data Profiling/Cleansing – Methodology

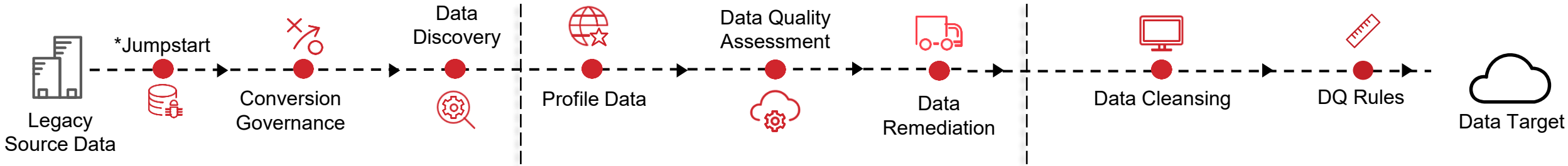
Highlighted below are the data profiling and cleansing tasks as part of the overall data conversion methodology. Discovery, governance, and profiling can be completed prior to mapping completion.

Data Profiling & Cleansing

Data Discovery	Collect and analyze legacy data for usage and data quality issues. Gather input based on previous data challenges and existing data conversion mapping. This will be refined.
Data Conversion Governance	Identify the key resources required to support the data conversion in defining business rules, profiling data and remediating issues.
Data Profiling	Based on the data discovery input start to profile the data, analyzing specific challenges and specific data quality issues.
Data Quality Assessment / Remediation	Assess the data quality issues and define the action plan to clean up the data.
Data Cleansing	Cleanse data as outlined in the remediation plan. Track progress against the targets.
Update Data Quality Rules	Evaluate and (re)define data quality rules and thresholds with our business partners and across our systems/tools.

Data Profiling/Cleansing – Example

Highlighted below are key tasks of profiling and cleansing. This is meant to be iterative, with additional details through each cycle.



Data Preparation

Conversion Governance

- Identify key resources to support the data conversion – specific roles include business owner, data steward, legacy system SME and data engineer

Data Discovery

- Document Known Pain Points & Data Issues
- Define Data Requirements (Data Quality)
- Define Business Rules (Data Relevancy Rules)
- Compare with target state Data Conversion Mapping

**Setup team to jumpstart data gathering prior to kickoff*

Data Assessment

Profile Data

- Identify and Prioritize Critical Data Elements (CDE) for Profiling
- Analyze Relevancy Rules
- Define CDEs Monitoring Frequency & Metrics
- Execute Profiling
- Compile/Provide Profiling Results for Review

Data Quality Assessment

- Review Profiling Results
- Review/Prioritize "Flagged" Data to be Cleansed
- Create DQ Scorecard/Dashboard

Data Remediation Plan

- Define Approach to Standardize Data
- Review/Approve Approach with Business

Data Cleansing

Data Cleansing

- Identify CDE for Cleansing
- Define Cleansing Reason/Requirements
- Conduct Cleansing Impact Analysis
- Execute Cleansing

Update Data Quality Rules

- Identify Need to Improve Data Quality Rules
- Identify Conflict and Complete Root Cause Analysis

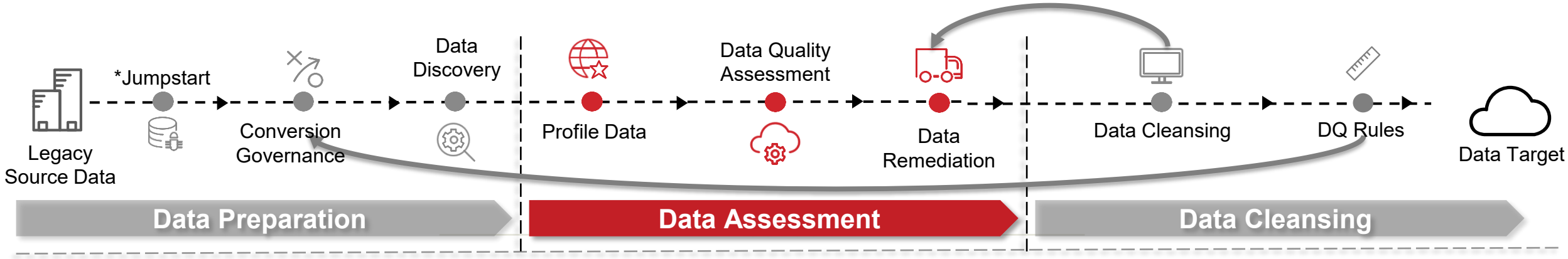
Data Profiling/Cleansing – Profiling Results

Data profiling provides insight to usage of the data in the legacy system application and profiles specific pain points or challenges based on the input from the affiliate team and IT teams.

Purchase Orders	
Total created in last 12 months	50,000
Total closed in 12 months	40,000
Total currently open	10,000
Open not updated in last 2 years	1,000
Incomplete (based on challenges)	500
Standard POs in last 12 months	20,000
Blanket POs in last 12 months	5,000

Data Profiling/Cleansing – Data Assessment

The primary purpose of data evaluation is to “flag” data values that may differ from expected or allowed values (data profiling) to determine where data needs to be updated or enriched (data quality assessment), and the approach to update or enrich the “flagged” data (data remediation)



Inputs

Data conversion policies, standards, and business rules

Data conversion metrics

Data relevancy rules

Legacy data source

Team Involvement

Legacy System SME(s) / IT - Extracts data from source system(s)

Profiling / Conversion Team - Executes profiling, defines remediation approach for flagged data to be review with business

Data Analyst(s) - Applies relevancy rules to the profiled data, executes data quality assessment, creates assessment results (data quality scorecard/dashboard) for business review

Data Owner / Data SME(s) - Reviews data quality assessment, reviews/approves remediation approach for flagged data. Identifies resources to support remediation

Outputs

Data profiling results

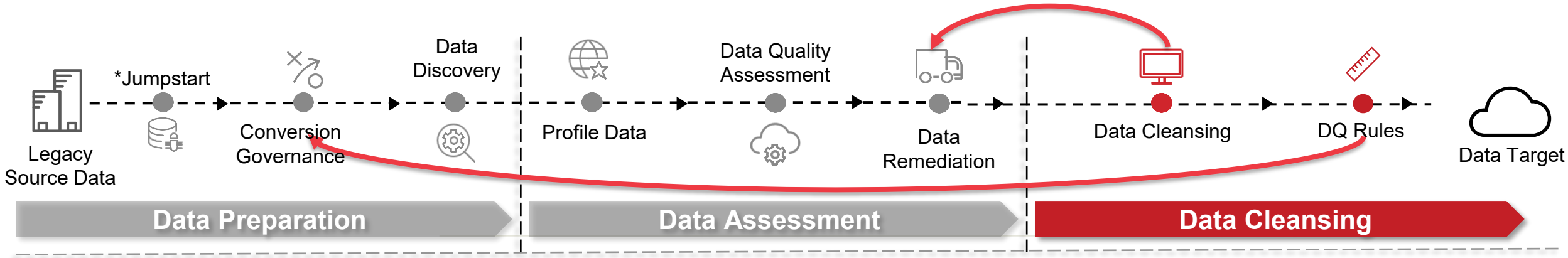
Data quality assessment (data quality scorecard and dashboard)

Data remediation approach

Data values for cleansing

Data Profiling/Cleansing – Data Cleansing

The primary purpose of data correction is updating data values (data cleansing) to ensure they align with our expected data value thresholds (DQ Rules)



Inputs

Data profiling results

Data quality assessment (data quality scorecard and dashboard)

Data remediation approach

Data values for cleansing

Team Involvement

Data Steward(s) / Data SME(s) - provides input on how best to make cleansing updates; potential to make the updates

Data Analyst(s) – supports any data updates and transformations required

Data Owner / Data SME(s) - tracks the progress of required data quality updates

Outputs

Updated data in source systems

Identification of data quality rules

Identification of data quality rules that can be used in data conversion

Data Profiling/Cleansing – Data Quality Assessment

Outlined below is an example of the insight available following our data quality assessment. Our data profiling results can be represented through a data quality scorecard/dashboard.

DATA QUALITY SCORECARD

Scorecards provide a mechanism to track progress towards defined data quality targets and help create DQ remediation plans

Data Object	Condition	Total Source Count	Identified Error Count	% of Records in Error	Total Loaded Count	% of Records Loaded	Score	Remediation
Item Master	Duplicate Manufacturer Name	51,165	5,231	10.2%	45,934	89.7%	🟡	Business to correct & provide revised extract file
Supplier Master	Incomplete Supplier Addresses	5,372	999	18.6%	4,373	81.4%	🔴	Business to correct & provide revised extract file

EXAMPLE

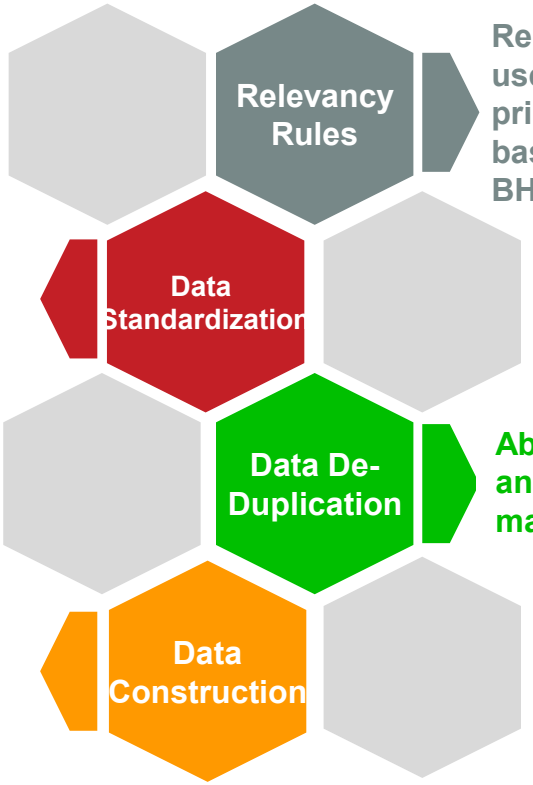
🟢 >= 95%
🟡 85% - 95%
🔴 <= 85%

Data Profiling/Cleansing – Data Cleansing

Data Cleansing is a pragmatic approach that aims to achieve high quality data while minimizing the data cleansing efforts

Addresses, postal codes, Country codes standardization is key for successful data integration

Standardization with support of third party data



Relevancy rules are used to filter data and prioritize master data based on impact to BHE

Check data inconsistencies across the whole enterprise

Ability to identify and analyze duplicate master data records

Data de-duplication tool allows the business user to easily manage duplicate records and identify the "survivor"

A data construction tool and/or approach is available for an easy and simple way to handle new data

Ability to build the data that does not exist in the legacy system

Data Profiling/Cleansing – Data Cleansing

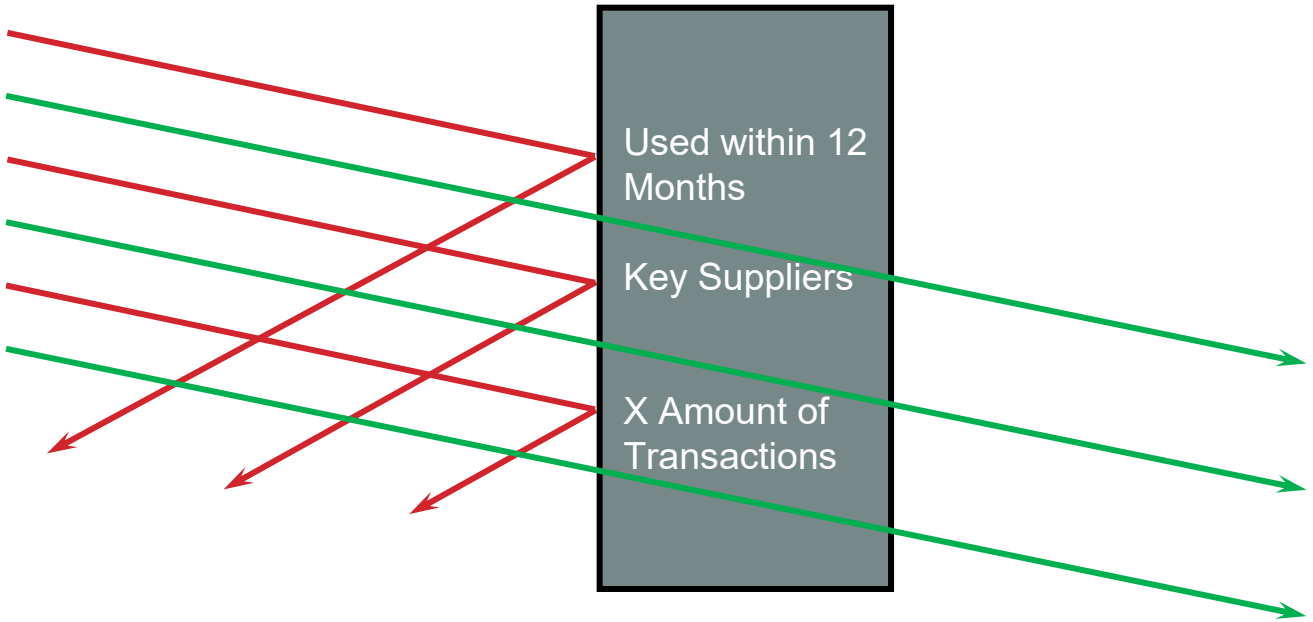
There are several approaches that can be used to cleanse data across multiple iterations to achieve the expected level of data quality

	Relevancy Rules	Standardize Data	Data De-duplication	Data Enrichment	Data Construction	Data Reporting
Definition	Relevancy rules are used to filter and prioritize data based on its usage and importance to the business	Standardization is performed by checking addresses, postal codes, country codes against third party data (e.g. United States Postal Services)	Ability to identify and analyze duplicate master data records. Matching criteria and fuzzy logic are applied to enable the identification of duplicate records.	Value mapping (parameters) table can be used to map legacy values into target values avoiding the need to hard code conversion rules.	Ability to fix existing data or build missing information before the data is loaded into the target system.	Before loading the data into the target system validate the data and number of records in a spreadsheet.
How it works	Relevancy rules are the initial filter to pre-select the data that should be loaded in the target system Example: active vendors in the last 2 years	The relevant data goes through a standardization process to assist in the next step of the data cleansing process	Similar records can be identified and grouped together to help the user select the “survivor” record and tag the rest as duplicates.	For example, plant ABCD in the legacy system can be mapped to 1234 in the target system.	A data construction tool and/or approach is available for an easy and simple way to build or fix data in Excel.	Business users can easily pre-validate the full set of data to be loaded in the target system either in a tool or export it to an Excel sheet.

Data Profiling/Cleansing – Data Quality Rules

By cross-checking master with transactional data it is possible to eliminate unnecessary data and make sure all the active master records and exceptions are considered in the process. The example below outlines that if data is not used within the last 12 months it is excluded from the conversion process.

Data that doesn't pass relevancy rules does not get loaded



Only relevant data passes through and gets loaded

Data Profiling/Cleansing – Data Quality Rules

The following is an example of the process from initial identification of resources to updating the data quality rules

#	Action	Who	Output	Example
1	Identify named resources to support the data profiling	Affiliate Lead	Conversion governance team	• Create conversion governance roster
2	Workshop to identify data challenges	Data Owner Data Steward	List of known data challenges	• Addresses are a known issue • Some duplicates associated w. suppliers Tax ID
3	Define the approach for profiling the data - options including extracts and database	Legacy System SME(s) / IT and Data Analyst(s)	Approach for data profiling	
4	Profile data and create DQ assessment results	Legacy System SME(s) / IT and Data Analyst(s)	Initial overview of data updates Analysis of specific challenges	• Supplier data: 10,000 creations in last 12 months, 5000 inactive; 783 addresses are in complete
5	Review data quality assessment to identify records for remediation	Data Owner / Data Steward	Summary of data quality challenges to be updated	• Update addresses for 665/783 suppliers with bad addresses
6	Determine how we're going to fix invalid data records	Data Owner / Data Steward	Data remediation approach	• Manually update invalid addresses in source • Create a conversion rule to populate supplier type with a default value
7	Update data identified in remediation plan	Data Analyst	Data cleansing results	• Update supplier addresses
8	Update data quality business rules as needed based on remediation approach	Data Steward	DQ business rules updated	• Addresses must not contain special characters • Inactive suppliers will be excluded from conversion

Contents

Data Profiling/ Cleansing – M&A Differences

As part of any future acquisitions, data profiling and cleansing will need to be updated due to likely differences in the data access and overall conversion process.

Area	Difference
System access	Project team will not have access to the existing legacy systems. Extract files will be provided from the company.
Data profiling	Data profiling will need to be executed from the files provided. BHE will need the ability to land, profile and report on the data.
Data quality	Quality of the data may not be influenced by the project team. Depending on the agreement, BHE will likely need to complement data quality and cleansing activities with their owns tools and capability. There is the potential to do an initial data quality assessment and verify completeness.
Data remediation	Data remediation may or may not be allowed in the system. If it is allowed, then the project team will need to coordinate with the selling company to address issues in the system. If this is not allowed, then the project team will have to remediate data from the data files provided from the selling company.
Extract timing	Timing of the extracted data will be dependent on the agreement between the selling company and BHE. Recommendation is to agree on the history of data that is needed by BHE and on specific timing to send the data for migration.

Data Profiling/Cleansing – Evolution Across Affiliates

Data cleansing and profiling will have differences from one affiliate to another, however the data profiling, data quality tools and solutions can be updated and used as affiliates continue to join the program. Below are steps that projects can take to account for changes:

#	Scenario	Re-use
1	New affiliate with source system (i.e., EBS) that has already been converted; significant reuse across all of the work done from extracts to profiling and business rules. Minor updates will be needed to reflect affiliate specific data and usage.	Significant across all components of data profiling and cleansing Expected re-use is 50+%
2	New affiliate with a new source system (i.e., Peoplesoft) that has not been converted. Moderate reuse of the existing data profiling and data cleansing work.	Moderate reuse – significant updates required for extracts and profiling Expected re-use is 30+%

Data Profiling/Cleansing – Complex Buying

There are some business events that may result in infrequent but important usage of master data. For example, in the event of buying a turbine. There are some suppliers that are engaged in an infrequent manner and as a result may fall outside of the data. Below are steps that can be taken to identify these type of suppliers:

#	Scenario
1	As part of the data profiling, business users should identify unique requirements for master and transaction data
2	Identify any unique attributes about the supplier(s) – for example, type, name, ID, PO type or dollar amount – that could help identify complexity and anomalies that needed to be accounted for in the data conversion
3	Update the relevancy rules to keep those suppliers that should be added to the conversion data
4	Validate the suppliers are part of the data load thru the data validation process

Contents

Data Profiling/Cleansing – Data Quality Rules

The data conversion document provides an additional input to the data quality rules associated with required fields, expected formats and size. Outlined below is an example of a data conversion document:

AutoSave Off DMDD_PAC_Asset.xlsx - Read-Only

SOURCE				TARGET					
Common People Table				MAXIMO Data Definition					
Message / Parent Component Name	Component Name	Component Element	Profiling Transformation	Component Element	Data Type	Length	Key?	Domain	Req?
				Asset Main Info					
			(autogenerate via new sequence)	ASSETNUM	UPPER	25	ASSETNUM		1
Category			Only convert E & M						
Super ord Equip			Needed only for Meter,Relay,LTCs assets	PARENT	UPPER	25			0
Manf. Serial No		MANUSERIALNO		SERIALNUM	UPPER	64			0
				ASSETTAG	ALN	64			0
Functional Location		floc(in spreadsheet extract)	it is substr of(1,8) of functionallocation/SUPFLOC(from functional_location extract)-To get the substation name	LOCATION	UPPER	12			0
Description	EQKT	EQKTX (Description in spreadsheet extract)	if description contains '(PS)'/'RETIRED' then do not convert	DESCRIPTION	ALN	100			0
				FAILURECODE	UPPER	8			0
				ISRUNNING	YORN	1			1
				ITEMNUM	UPPER	30			0
MAINT PLANT		MAINTPLANT	inherit from location PAC_SITEID(Mapping) 2005 - 2460 = RMP, 2620 - 2875 = PP	SITEID	UPPER	8			1
			default = PAC	ORGID	UPPER	8			1
				PRIORITY	INTEGER	12			0
objecttype / Technical Ident No.	EQUI	EQTYP	ASSET.TYPE (ASSETTYPE)	ASSETTYPE	UPPER	15		ASSETTYPE	0
System Status			If system status like INAC or DLFL or AVLB or RETR or SOLD then do not convert. If INST or ASEQ and userstatus does not contain NIS, DESP or SPAR then map to In-Service, IF INST or ASEQ and userstatus contains ONO then Not Ready IF INST or ASEQ and userstatus contains NIS or SPAR or DESP then DE-ENERGIZED	STATUS	ALN	20		LOCASSETSTATUS	0
User Status	JEST	STAT	refer BHE_USAGE tab user status like DUP or RETR then do not convert	BHE_USAGE	ALN	20		BHE_USAGE_D	0
objecttype	KLAH	CLINT	ASSET.TYPE (ASSETTYPE)	CLASSSTRUCTUREID	UPPER	20			0
				BINNUM	ALN	8			0

Data Profiling/Cleansing – Key Activities

The initiative can support the data cleansing activities by automating some of the data cleansing activities. Outlined below are examples of the data cleansing activities:

- Data cleansing that can't apply a rule to update the data is generally the affiliate responsibility. Examples include:
  - Updating a customer address that is incomplete
  - Updating names or descriptions of a key master data object (i.e., updating a supplier name)
  - Updating data that does not have a standard rule (i.e., project name and number)
- Data cleansing that can be automated as part of the conversion build can be enabled by the initiative team – examples include:
  - Populating a default value for a field (i.e., blank in legacy and default is required in Oracle)
  - Creating a rule to map value "A" in legacy to value "123" in Oracle
  - Creating a rule to close all POs that have been open since 2019 and earlier

Data Profiling/Cleansing –Example

Object	Business Rule Name	Rule Type	Description
Supplier+E	Tax ID	Quantitative	Requiring Tax ID/FEIN for all supplier - Need this as part of business requirements - Canada, format is either XX-XXXXXXX or for individuals that are suppliers it can be formated as XXX-XX-XXXX - UK unique combiniation of letters and/or numbers. for most individuals this will be their National Insurance Number,
Supplier	Phone Number	Quantitative	Business requires a phone number to contact supplier - Canada will have dialing code of +1 plus 10 digits - UK will have dialing code of +44 plus 10 digits
Customer	Mailing Address	Quantitative	Field is required by the Business - to actively communicate with customer
Customer	Zip Code	Quantitative	Field is required by the Business to ship - UK format is 7 alphanumeric charcters - Canada format is 6 alphanumeric characters
Supplier	Bank Account Info	Quantitative	Bank Information is required by the business for payments - Canada will have a bank account number - UK will use IBAN number format
Supplier	Part Number Details	Quantitative	Part Number needs to be populated for Business to identify the parts
Product	Currency	Quantitative	Requires that the Product is prices, use the 3 character ISO code - Canada ISO code is CAD - UK ISO code is GBP
Product	Dimensions	Qantitative	Weight is a business required field - UK use metric measurements - Canada use metric measurements
Product	Duplicate Check	Qualitative	Removes all duplicate data within a data set