**Anglicare-SA**

****

**Data Warehouse**

**Data quality requirements**

**Phil Crick**

Document Control

|  |  |  |  |
| --- | --- | --- | --- |
| **Date** | **Nature of Change** | **Author** | **Doc version** |
| 11 Apr 2017 | First draft for review and feedback | Phil Crick | V1 |
|  |  |  |  |

Contents

[Introduction 3](#_Toc479079220)

[Purpose and Overview 3](#_Toc479079221)

[Scope 3](#_Toc479079222)

[Assumptions 3](#_Toc479079223)

[Constraints 3](#_Toc479079224)

[Dependencies 3](#_Toc479079225)

[Functional Requirements – High Level **Error! Bookmark not defined.**](#_Toc479079226)

[Functional Requirements – Detailed **Error! Bookmark not defined.**](#_Toc479079227)

[Non-Functional Requirements **Error! Bookmark not defined.**](#_Toc479079228)

[Appendix A – Example statement **Error! Bookmark not defined.**](#_Toc479079229)

# Introduction

## Purpose and Overview

The purpose of this document is to provide a complete up-to-date list of all the data quality checks that are built into the data warehouse. This list provides clarity to both business users and developers so that data quality checks can be implemented quickly and the outcome of the checks understood.

Data quality checks are necessary as data from source systems are never perfect. Errors in data may not be realised by business end users, so this system can add a valuable insight into the quality of the data in source systems and allow the data to be addressed or cleaned.

## Scope

* Data included only in the data warehouse, the system will not check data outside of the warehouse
* The system will only fix/tag data in the warehouse, the business will need to address any data quality issues in the source systems

## Assumptions

* Business users will be able to clearly define data quality rules
* Business users are able to understand the data warehouse data models
* The business understand that the warehouse is only displaying the data it is given, even if it has data quality issues

## Constraints

* Quality checks are restricted to column, structure and rule types (see next section for explanation)

## Dependencies

# Overview of data quality checks

The data warehouse has a data quality control system; this system is responsible for checking data in three different ways according to a set of configurable rules.

## Column checks

These are the simplest checks; they check one column value against a rule. E.g. is the employee name blank? Other examples include numeric check, alphanumeric check, min/max lengths and format checks.

## Structure checks

Structure checks test the relationship across columns. Two or more attributes may be tested to verify they implement a hierarchy, such as a series of many-to-one relationships. Structure checks also test foreign/primary key relationships between two columns in two tables.

## Business rules

Business rules implement more complex tests that do not fit the simpler column or structure check categories. More complex multi-column checks are common (if this and this and this type rules) as are threshold type rules which check multiple rows for various values, e.g. are all our staff male?

When the data quality check system encounters errors, it has choices about what to do next:

1. Halt the process
2. Send offending records to a separate area and log them
3. Let all data pass through, but tag erroneous data

The third option is by far the best choice whenever possible. Halting the process is obviously a pain because it requires manual intervention to diagnose, restart the job or abort completely. Segregating erroneous data is often a poor solution because it is not clear when or if these records will be fixed and reintroduced. Until the records are restored to the data flow the overall integrity of the database is questionable because records are missing. The third option of tagging erroneous data often works well. Bad fact and dimension table data is tagged in the audit dimension so end users can easily see it in reports and the either report it or act on it.

# Data quality checks

Data quality checks for each data mart are held within separate spreadsheets.