**ECB-RESTRICTED**

Data Intelligence Service Centre

DISC Data Factory – Requirements for Data Load *<Data Source>*



**Table of contents**

0. Document Information 5

0.1 Versioning 5

0.2 Reference Documentation 5

1. Introduction 6

1.1 Data Description 6

1.2 Scope and Objectives 6

1.2.1 Scope 6

1.2.2 Business objectives 7

1.2.3 Technical objectives 7

1.3 Stakeholders 7

1.4 Data Flow & Architecture 8

1.5 Overview of Data Structures 9

2. Data ingestion 10

2.1 Load Data 10

2.2 Data Refresh/Load Cycle 12

3. Data integration 13

3.1 Logical Data Model 13

3.1.1 Landing Zone 14

3.1.2 Staging Store 15

3.1.3 Corporate Store 17

3.2 Target structure in HDFS 19

4. Automation and monitoring 20

5. Error Handling 21

6. Appendix 22

6.1 Staging Store Metadata Descriptions 22

6.2 Corporate Store Metadata Descriptions 22

**Table of Tables**

[Table 1: Version History 4](#_Toc533064683)

[Table 2: Services in Scope 5](#_Toc533064684)

[Table 3: Stakeholders 6](#_Toc533064685)

[Table 4: Data Structures 8](#_Toc533064686)

[Table 5: Data Loading 9](#_Toc533064687)

[Table 6: Data Refresh / Load Cycle 11](#_Toc533064688)

[Table 7: Data Processing in Corporate Store 16](#_Toc533064689)

[Table 8: Target Structure in HDFS 18](#_Toc533064690)

[Table 9: Automation and Monitoring 19](#_Toc533064691)

[Table 10: Error Handling 20](#_Toc533064692)

# Document Information

## Versioning

This document is in DARWIN: <LINK>

**Versioning:** v0.8 – ready for review; v0.9 – ready for sign-off; 1.0 sign-off

Table 1: Version History

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Version | Date | Who | Comment | Release |
|  |  |  |  |  |
|  |  |  |  |  |

## Reference Documentation

Throughout this document we shall make reference to a number of files that contain important details for the implementation of this dataset:

1. Field List – <LINK>
2. Test data – <LINK>
3. Naming conventions – <LINK>
4. <*Please add additional documentation if needed*>

# Introduction

## Data Description

Please describe the data and how you currently collect and process it. Also specify the main obstacles and pain points connected to this process.

## Scope and Objectives

### Scope

The services described in Table 2 are covered by this Functional Design:

Table 2: Services in Scope

|  |  |  |
| --- | --- | --- |
| Process | Explanation | Requested? |
| Data Load | Establishment of connection to the data source(s) and retrieve data sets | **YES** |
| Data storage in Landing Zone in their native format | **YES** |
| Propagation of data to the Staging Store and enrichment with metadata | **YES** |
| Implementation of de-normalisation as needed for performance on HDFS | **NO** |
| Offloading of vintage data from Oracle Exadata to Corporate Store Hadoop | **NO** |
| Scheduling of regular retrievals of data with batch/scheduling | **YES** |
| Corporate Store Creation | Implementation of data model in Corporate Store (Hadoop) | **YES** |
| Implementation of data model in Corporate Store (Oracle Exadata) | **NO** |
| Simple onboarding of data sets with minimal changes and additions to the original source structure | **YES** |
| Transformation and integration of data sets between DISC data stores based on complex data flows with various data sets as source and data products as target within data mart | **NO** |
| Scheduling of regular data flows with batch/scheduling | **YES** |
| Dissemination of single data sets or data products from DISC store | **NO** |
| Configuration of security settings as required | **YES** |
| Partitioning and performance tuning | **YES** |
| Reporting Services | Implementation of semantic layers for business objects and KPI definitions | **NO** |
| Creation of reports and dashboards for regular reporting and ad-hoc requests | **NO** |

### Business objectives

Describe the business case and objectives of the dataset to be managed. Add as many figures as required to facilitate the understanding of the business objectives.

### Technical objectives

Describe the technical objectives of the solution. Do you need parallel/distributed computing? Would a relational database be needed? Do you need to change data?

## Stakeholders

The following table provides a list of stakeholders with an interest in the dataset and relevant developments.

Table 3: Stakeholders

|  |  |
| --- | --- |
| Entity | Description |
| <Please Specify> | Owner of requirements and of data set in DISC corporate store. Prepare and execute user-acceptance test and sign-off for go-live. |
| <Please Specify> | Proclaimed interest in using the data set |
| DGIS/DAS | Coordination of development, test and release activities |

## Data Flow & Architecture

The following three stages describe the processes applied at DISC to integrate the data from source to target. This is a DISC general description – details for specific data sources may vary.

(1) **Source System to Landing Zone**

Data files are retrieved, unzipped and stored on the Landing Zone of the DISC platform by the bus system EXDI, Web Services or Direct Access. The landed data are a 1:1 copy of the original datasets without any structural changes.

(2) **Landing Zone** **to** **Staging Store**

At the Staging Store the data is supplemented with technical metadata and stored into one Hive managed table per data structure (format: Parquet). The data type for all functional attributes is string format. This data can be queried by authorised users only.

(3) **Staging Store to Corporate Store**

The Corporate Store is presenting the access layer for end users. Data will be loaded 1:1 from the Staging Store transforming the variables into the correct technical data types. In some cases, data quality checks are performed.

## Overview of Data Structures

Data structures are extracted from the data source and on-boarded onto the DISC platform. You can find the list of all data structures in Table 4. Details about the single structures can be found in the field list [1].

Table 4: Data Structures

|  |  |  |
| --- | --- | --- |
| Structure | Primary Key | Partition Key (if needed) |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

# Data ingestion

Data ingestion entails loading data from source into the landing zone. This process is shaped by various parameters related to the file and source characteristics, load method and access details.

## Load Data

Loading data entails establishment of connection to the data source(s) and retrieval of data sets. Data is stored in the landing zone.

Table : Data Loading

|  |  |  |  |
| --- | --- | --- | --- |
| AskMe | Topic | Req. ID | Specification |
| 1 | **Source data format** | Load-001 | File Format  txt  csv  xml  other:  Oracle Tables  … |
| 2 | **File compression (as received)** | Load-002 | Compression:  RAR  gzip  none  other: |
| 3 | **Multi-part containers** | Load-003 | Occurrence of multi-part containers:  yes  no |
| 4 | **Number of files/ structures** | Load-004 | <insert number> |
|  | **Data Volume** |  | Total Load Volume of   * uncompressed files: 12 * compressed files: |
|  | **Data Volume increase** |  | Constant load volumes  Annual volume increase (%):  Other: |
|  | **Pre-load historical data** |  | Yes, expected volume:  Not required  Data can be loaded from the data lab  Yes  No |
|  | **Interface** |  | EXDI  Direct Access  Web Service  other: |
|  | **Source System** |  | FTP Server  HTTPs  Oracle DB  Website:  other: |
|  | **Technical Account** |  | Technical User ID/Password  Certificate  Other: |

## Data Refresh/Load Cycle

Data refresh and load cycles are set according to the business needs.

Table 6: Data Refresh / Load Cycle

|  |  |  |
| --- | --- | --- |
| Topic | Req. ID | Specification |
| Frequency |  | Daily  Weekends  Holidays (if no, please specify which holidays to exclude)  Weekly  Monthly  Quarterly  Yearly  Other: |
| Scheduling |  | Event based  Time based  No scheduling  Details: <Please describe at what time the data should be loaded> |
| Data Refresh Method |  | Append  Truncate  Delayed Truncate  Snapshots  Daily Snapshots  Delta  Details: |

# Data integration

## Logical Data Model

The data model for the dataset is pre-scribed by the field list. Source data maps into a corresponding physical table in Hive/Impala in the Staging and Corporate Store. During the loading process, DISC specific technical metadata shall be added.

* The **field list** containing the target structures of the Landing Zone, Staging Store and Corporate Store is provided under [1]
* Names of the input structures will be retained adding a prefix in accordance to the DISC **naming conventions** [3]
* Example data to be used for testing purposes is available under [2]
* Data Type Conversions:
  + Staging Store: Variables shall be stored in **string** format
  + Corporate Store: Variables shall be transformed into the **correct technical data type** for Hive and Impala.

<Please describe any additional requirements>

### Landing Zone

The landing zone stores a 1:1 copy of original data in the folder “/data/landing/<data source>”. It will not be accessible for business users.

|  |  |  |
| --- | --- | --- |
| **Topic** | **Req. ID** | **Specification** |
| **Transformation Rules** |  | Business Rules  1:1 Data Load  Data Type Conversions |
| **Transformation Rules** |  | Format  .csv  .txt  .xml  other: |
| **User Access** |  | Access  No Access  [If access required]  Business justification:  ARP Role: |
| **Retention Period** |  | none  6 month  12 months  until next data delivery |

### Staging Store

At the Staging Store, the data model from the source structures is supplemented with additional technical metadata. No further adjustments to the original data model are required at this step.

|  |  |  |
| --- | --- | --- |
| **Topic** | **Req. ID** | **Specification** |
| **Target Format and Compression** |  | Target Format  Parquet  Avro  other:    Compression  Compression  Snappy  Other:  No Compression |
| **Partition Keys** |  | No partition required  Specific partition key as specified in table 4 |
| **Technical Metadata**[[1]](#footnote-1) |  | At the end of each structure the following technical metadata will be added (Metadata / Value):   |  | | --- | | tec\_source\_system | | tec\_dataset | | tec\_ingestion\_date | | tec\_version\_id | | tec\_execution\_date | | tec\_run\_id | | tec\_current\_flag | | tec\_business\_date  tec\_crc | |
| **Filters** |  | No application of filters  Filter for delta load  Specification of filters: |
| **Transformation Rules** |  | Business Rules  1:1 Data Load  Details: please see chapter Data Model for additional information on data enrichment and modification |
| **Data Type** |  | String  Data Type Conversions as per field list [1] |
| **Data Quality Checks** |  | Required  Not required |
| **User Access** |  | Access  No Access  [If access required ]  Business justification:  ARP Role: |
| **Retention Period** |  | none  3 month  12 months  until next data delivery  other: |

### Corporate Store

In the Corporate Store, transformations are applied to data if needed. Data is stored using corresponding data types.

Table 7: Data Processing in Corporate Store

|  |  |  |
| --- | --- | --- |
| **Topic** | **Req. ID** | **Specification** |
| **Target Format and Compression** |  | Target Format  Parquet  Avro  other:    Compression  Compression  Snappy  Other:  No Compression |
| **Partition Keys** |  | No partition required  Specific partition key as specified in table 4 |
| **Historisation** |  | No bi-temporal historisation  Bi-temporal historisation |
| **Technical Metadata** |  | At the end of each structure the following technical Metadata will be added:   |  | | --- | | tec\_ingestion\_date | |
| **Filters** |  | No application of filters  Filter for delta load  Specification of filters: |
| **Transformation Rules** |  | Business Rules  1:1 Data Load |
| **Data Type** |  | String |
| **Data Quality Checks** |  | Required  Not required |
| **Standard Views** |  | no view  Current view  other:  Business Logic: |
| **User Access** |  | Confidentiality  Public  ECB Public  Non-Public  ARP Role (If one is already existing):  Read only access will be granted.  <Please specify all users to have initial access to data> |
| **Level of Access** |  | Data set level  Table level |
| **Retention Period** |  | 6 month  12 months  until next data delivery  No retention policies specified as of now – Data is not removed from Corporate Store. |

## Target structure in HDFS

Table 8 provides an overview of how the loaded data is stored within each stage.

Table 8: Target Structure in HDFS

|  |  |  |  |
| --- | --- | --- | --- |
|  | Landing Zone | Staging Store | Corporate Store |
| HDFS directory | data/landing/<data source> | data/staging/<data source> | data/corporate/<data source> |
| Hive Target DB | -- | stg\_<data source> | crp\_<data source> |

# Automation and monitoring

Table 9: Automation and Monitoring

|  |  |  |
| --- | --- | --- |
| **Topic** | **Req. ID** | **Specification** |
| Integration with DGIS workload automation tool | NF-001 | Not Required  Required  Type:  Data base  File based  Manual |
| DGIS Workflow monitoring | NF-002 | Not Required  Required  Success alerts  Warning alerts  Failure alerts  Notifications should be sent to: <please provide an email distribution list>  The flow will be scheduled at:   * <please insert scheduled flow days> * <please insert time when data becomes available> |
| Workflow automated re-start | NF-003 | Not Required  Required  Configurable wait time  Set wait time |
| Manual re-start option | NF-004 | Required  Not Required |

# Error Handling

It is defined that every record must be loaded into the respective target structure of each layer (Staging Store, Corporate Store). In case of errors, standardized processes are set in place for correct error handling and tracking.

Table 10: Error Handling

|  |  |  |
| --- | --- | --- |
| **Topic** | **Req. ID** | **Specification** |
| Invalid input files - Identification |  | Identify invalid input file in the Landing Zone  Required  Not Required  Handling:  Error Message and Ticket in ITSM manually  Other: |
| Invalid input files - Processing |  | Resiliency:  Valid files are processed even if there is an invalid file  Processing stops upon encountering an invalid file  Other: |
| Invalid input files - Definition |  | Invalid input files are defined as follows:  Empty file  Unexpected size (Please specify)  Wrong file format  File integrity  Duplicate files <please define>  Late files <please define>  Other: |
| Missing files |  | File is considered missing after: <please specify>  If a file is missing:  Ignore  Send a warning message and proceed  Send a warning message and abort  Other: |

# Appendix

## Staging Store Metadata Descriptions

|  |  |
| --- | --- |
| **tec\_source\_system** | This metadata stores the information from which source system the data has been retrieved. |
| **tec\_dataset** | This metadata stores from which data set the data has been retrieved. |
| **tec\_ingestion\_date** | The ingestion date is the date when data is ingested into the staging store. |
| **tec\_version\_id** | This metadata keeps track of multiple versions for one load cycle (Versioning), in case one data set is loaded more than once for one business date (e.g. re-load from source system). |
| **tec\_execution\_date** | The parameter for the execution date will be delivered by WLA. This is the date on when the workflow will be executed. |
| **tec\_run\_id** | This metadata is the unique ID per data load. This is from the run\_id generated by WLA. |
| **tec\_current\_flag** | The current flag is a binary status indicator (Y/N). All current data records have the value Y, whereas closed data records (business or technical closure) are marked by an N. |
| **tec\_business\_date** | The business\_date is the date when the data has been updated on the business side. |
| **tec\_crc** | The tec\_crc column includes the hash value of the concatenated content of each row. |

## Corporate Store Metadata Descriptions

|  |  |
| --- | --- |
| **tec\_ingestion\_date** | The ingestion date is the date when data is ingested into the corporate store. |

1. Definition of technical metadata in the appendix [↑](#footnote-ref-1)