**NBS IT / Business Analytics CoE**

**Doc. ID:** CIT.SAL.DS.001.DICE-TP.V1.0

**Title:**  **DICE-TP (Territory Potential)**

Application Clarity ID: 084677

**Design Specification**

|  |  |  |
| --- | --- | --- |
| **Author:**  **Arvind Chimta**  Big Data Engineer  EPAM | **Date:** | **Signature:** |
|  |  |  |
|  |  |  |
| **Aleksandr Fomin**  Solution Architect  EPAM |  |  |
| |  | | --- | |  | |  |  |
| **Approval:** | **Date:** | **Signature:** |
| |  | | --- | |  | | **PUNNANI, LAKSHMI KANTH**  Project Manager  Novartis | |  |  |

|  |  |  |
| --- | --- | --- |
| **Hitendra Patel**  Technical Design Expert  Novartis |  |  |
| **Ashish Sinha**  Solution Design Manager  Novartis |  |  |

**Document History**

|  |  |  |  |
| --- | --- | --- | --- |
| **Version** | **Date** | **Author** | **Comments** |
| 1.0 | 08-11-2018 | Elena Kostenko | Initial version |
| 1.1 | 14-11-2018 | Arvind Chimta | Final version |

**Table of Contents**

[1. Introduction 1](#_Toc536551672)

[1.1 Purpose 1](#_Toc536551673)

[1.2 Prerequisites 1](#_Toc536551674)

[1.3 Reference to Standards 1](#_Toc536551675)

[1.4 Definitions 1](#_Toc536551676)

[2. Overview 1](#_Toc536551677)

[2.1 Solution Architecture 1](#_Toc536551678)

[2.2 Ingestion Process Overview 1](#_Toc536551679)

[2.2.1.1 Quality checks 1](#_Toc536551680)

[2.2.1.1.1 Sales Records 1](#_Toc536551681)

[2.2.1.1.2 Currency Records 1](#_Toc536551682)

[2.2.1.2 Referenced data 1](#_Toc536551683)

[2.2.1.3 Error handling 1](#_Toc536551684)

[3. System Design & Customization 1](#_Toc536551685)

[3.1 General Specifications 1](#_Toc536551686)

[3.1.1.1 Landing area 1](#_Toc536551687)

[3.1.1.2 Input file format 1](#_Toc536551688)

[3.1.1.3 File structure 1](#_Toc536551689)

[3.1.1.4 File name convention 1](#_Toc536551690)

[3.1.1.5 File name and file structure validation 1](#_Toc536551691)

[3.1.1.6 Data Mart 1](#_Toc536551692)

[3.1.1.7 Data Model 1](#_Toc536551693)

[3.1.1.8 DDLs 1](#_Toc536551694)

[4. References, Attachments, Abbreviations/Acronyms 1](#_Toc536551695)

[4.1 References 1](#_Toc536551696)

[4.2 Abbreviations / Acronyms 1](#_Toc536551697)

[5. Publishing 1](#_Toc536551698)

# Introduction

## Purpose

This document is Design Specification (DS) for DICE TP. It contains details on customized items based on the user requirements specifications for the application as defined in the User requirements Specification(s)

* CIT.SAL.URS.001.DICETPv1.0

## Prerequisites

The document(s) listed below must be approved prior to the approval of this document

* User Requirements Specification(s) CIT.SAL.URS.001.DICETPv1.0 – Application

## Reference to Standards

For system design, the following standards are applicable:

* *IGM Standard “Security & Compliance by Design”*
* *IGM Security Baseline*

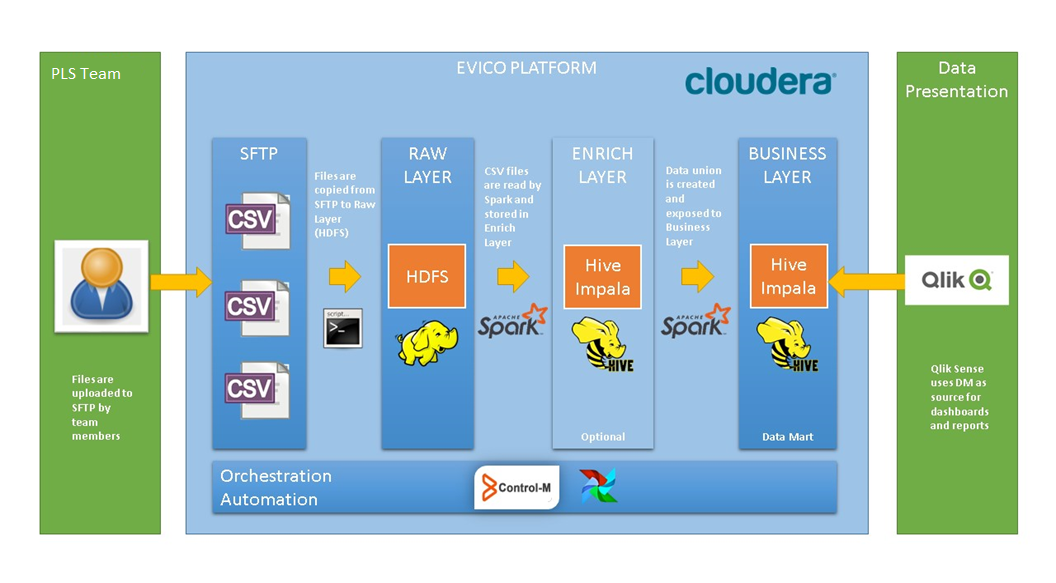
## Definitions

* Configuration: Setting of configuration parameters, no coding
* Customizing: Creation of custom code (new / changed functionalities)

# Overview

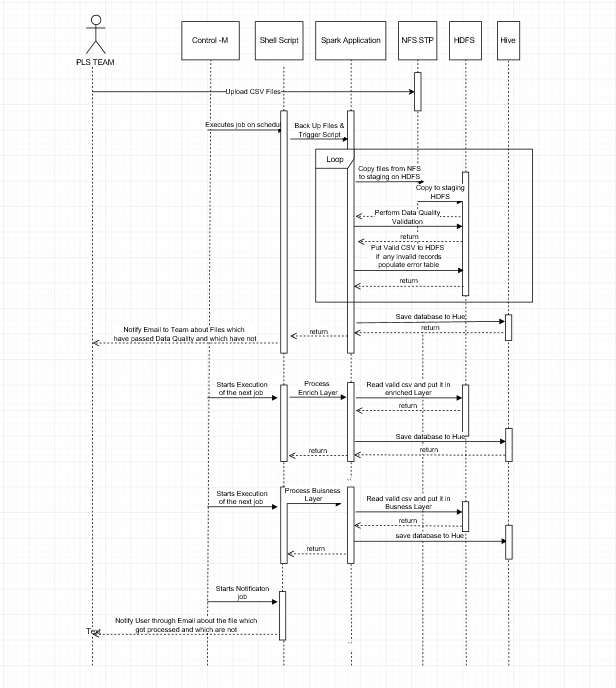
## Solution Architecture

At high-level, following project is to design and implement automated pipeline for ingesting DICE TP data and exposing it for Qlik Sense application for further consumption and visualization. To speed-up implementation process, existing frameworks and tools will be used and minimum coding required.



## Ingestion Process Overview

Please refer to below flow chart diagram for ingestion process steps



|  |  |  |
| --- | --- | --- |
| Object/Subject | Type | Description of usage |
| PLS Team | Team/Person | Team responsible for preparing input CSV files and uploading files to landing zone (FTP/NFS location) |
| Ops Team | Team/Person | Team responsible to start ingestion process on-demand using defined Control-M job (automated pipeline) |
| Control-M | Software/Tool | Orchestration tool. Defined job will be executed on-demand when new data is ready and available on FTP |
| Shell script | Software/Script | Bash script to process CSV files uploaded to FTP location. Bash script handles copying data from NFS to HDFS and running Spark application to process data |
| NFS/FTP | Software/Infrastructure | Data is uploaded to FTP by PMR Team members and is processed by automated pipeline. Please refer to it as **Landing Zone** |
| HDFS | Software/Infrastructure | Data is copied from NFS location to HDFS by Shell script/. Copying from NFS to HDFS is required by Spark processing engine. |
| Hive | Software/Infrastructure | Data is process by Spark engine and stored in Hive partitioned table. Data is accumulated and accessible for Qlik Sense application. |

#### Quality checks

Simple count checking will be implemented to ensure that number of records in input CSV file matches number of records stored in Enriched Layer and Business Layer.

Quality checking will be performed during runtime of bash script and Spark application. Quality checking will apply to columns in input CSV file for zeros (0) check according to MetricsToCheck.csv for sales records.

##### Sales Records

Data Quality Validations are applied per file.

If record fails, then file will not process the file to EDM. It stays in raw layer, error table will be populated with failed records.

Users will be notified through email regarding the files which have passed data validation check and files which have not.

Rest of the valid files continue to be processed to EDM and Business Layer.

##### Currency Records

If record fails, then records of that country and quarter will not process the file to EDM.

It stays in raw layer, error table will be populated with failed records.

Remaining valid records in the file are processed.

Users will be notified through email regarding the files which have passed data validation check and files which have not and the rest of the process continues without failing for EDM and Business layer.

For the rest of the column quality checking will apply to following fields/columns in input CSV file for NOT NULL check:

* Country
* Group Product Name
* Territory Code
* Territory Name
* Month Year

After the above process are completed in the Business layer, notify job sends out email to the user with files information which are processed and which are not.

Error Table contains the following information.

* Error In: Type of Files the error is in: Sales Data/ Currency Data / Metrics related data
* Error Row: Entire Row which is having the error
* Error Description: the column specific errors. For EG:- Value for Group Product Name does not exists
* Error Date: Date of error encountered

#### Referenced data

Master Metric mapping file

|  |  |  |
| --- | --- | --- |
| **Country** | **Group\_Product\_Name** | **Metric** |
| US | GILENYA | Market\_Share\_Units |
| US | COSENTYX | Market\_Share\_Units |
| US | ENTRESTO | Market\_Share\_Units |
| AUSTRALIA | COSENTYX | Market\_Share\_Value |
| AUSTRALIA | ENTRESTO | Sales\_Per\_Capita\_Units |
| AUSTRALIA | GILENYA | Market\_Share\_Value |
| AUSTRALIA | LUCENTIS | Market\_Share\_Value |
| AUSTRALIA | ULTIBRO | Market\_Share\_Value |
| ITALY | ENTRESTO | Sales\_Per\_Capita\_Units |
| ITALY | GALVUS | Market\_Share\_Value |
| ITALY | GILENYA | Market\_Share\_Value |
| ITALY | ULTIBRO | Market\_Share\_Value |
| ITALY | COSENTYX | Market\_Share\_Value |
| BRAZIL | ENTRESTO | Market\_Share\_Value |
| BRAZIL | ULTIBRO | Market\_Share\_Value |
| CHINA | ENTRESTO | Sales\_Per\_Capita\_Value |
| CHINA | EXFORGE | Sales\_Per\_Capita\_Value |
| CHINA | GALVUS | Sales\_Per\_Capita\_Value |
| CHINA | LUCENTIS | Sales\_Per\_Capita\_Value |
| CANADA | COSENTYX | Sales\_Per\_Capita\_Value |
| CANADA | ENTRESTO | Sales\_Per\_Capita\_Value |
| CANADA | LUCENTIS | Sales\_Per\_Capita\_Value |
| CANADA | ULTIBRO | Sales\_Per\_Capita\_Value |
| CANADA | XOLAIR | Sales\_Per\_Capita\_Value |
| UK | COSENTYX | Market\_Share\_Value |
| UK | LUCENTIS | Market\_Share\_Units |
| UK | ENTRESTO | Sales\_Per\_Capita\_Units |
| JAPAN | COSENTYX | Market\_Share\_Value |
| JAPAN | GALVUS | Market\_Share\_Value |
| JAPAN | EQUMET COMBO | Market\_Share\_Value |
| JAPAN | EXELON | Market\_Share\_Value |
| JAPAN | LUCENTIS | Market\_Share\_Value |
| JAPAN | ULTIBRO | Market\_Share\_Value |
| JAPAN | XOLAIR | Market\_Share\_Value |
| SPAIN | GILENYA | Market\_Share\_Value |
| SPAIN | ENTRESTO | Sales\_Per\_Capita\_Units |
| SPAIN | LUCENTIS | Sales\_Per\_Capita\_Units |
| SPAIN | XOLAIR | Sales\_Per\_Capita\_Value |
| SPAIN | COSENTYX | Sales\_Per\_Capita\_Value |
| SPAIN | GALVUS | Sales\_Per\_Capita\_Value |
| SPAIN | ULTIBRO | Market\_Share\_Value |
| GERMANY | COSENTYX | Market\_Share\_Value |
| GERMANY | ENTRESTO | Market\_Share\_Value |
| GERMANY | GILENYA | Market\_Share\_Value |
| GERMANY | LUCENTIS | Market\_Share\_Units |
| GERMANY | ULTIBRO | Market\_Share\_Value |
| FRANCE | GALVUS | Market\_Share\_Value |
| FRANCE | GILENYA | Market\_Share\_Value |
| FRANCE | LUCENTIS | Market\_Share\_Value |
| FRANCE | ULTIBRO | Market\_Share\_Value |
| FRANCE | ENTRESTO | Sales\_Per\_Capita\_Units |
| FRANCE | COSENTYX | Patient\_Share |

Metric calculation reference

|  |  |  |
| --- | --- | --- |
| **Metric** | **Numerator** | **Denominator** |
| Market\_Share\_Units | Product\_Sales\_Units | Market\_Sales\_Units |
| Market\_Share\_Values | Product\_Sales\_Value | Market\_Sales\_Value |
| Sales\_Per\_Capita\_Units | Product\_Sales\_Units | Market\_Sales\_Population |
| Sales\_Per\_Capita\_ Values | Product\_Sales\_Value | Market\_Sales\_Population |
| Patient\_Share | Product\_Sales\_Patients | Market\_Sales\_Population |

Currency conversion reference

|  |  |  |
| --- | --- | --- |
| Column Name | Description | Values |
| country | Name of the country | US, JAPAN, UK, AUSTRALIA, ITALY,BRAZIL, CANADA,CHINA, FRANCE, GERMANY, SPAIN |
| group\_product\_name | Global product group name | COSENTYX, GALVUS |
| quarter\_year | Quarter, year | 01\_2018 |
| USD\_conversion\_rate | Conversion rate from local currency to USD | 1.14 |

#### Error handling

Runtime errors will be handled by Control-M jobs based on exit codes returned by processes.

Control-M will update pipeline status accordingly and send notifications as configured.

Notify.sh job will save the error table into csv at /mnt/share/sdata/ph/com/r/gbl/dice\_tp/error\_csv location and this will be sent as a mail attachment to the PLS teamDL:

* **DICE\_TP\_File\_Processing\_Notifications\_PRJ\_GBL\_GBL@dl.mgd.novartis.com**

# System Design & Customization

Programming languages used

* Scala
* Bash

Framework(s):

* Apache Spark

Infrastructure needed:

* FTP server/client
* Apache Spark (cluster)
* Apache Hadoop (HDFS + YARN)
* Apache Hive
* Apache Impala

Code repositories:

https://gitlab.evico.novartis.net/project\_ph\_com\_dice\_tup/project\_ph\_com\_dicetp

## General Specifications

This section covers the design / customization made based on the requirements given in the URS CIT.SAL.URS.001.DICETP

#### Landing area

All files that are intended to be ingested, should be uploded to landing area.

Landing area will be implemented as a FTP location so users will be able to upload files using FTP client with secured connection.

FTP location: /mnt/share/sdata/ph/com/r/gbl/dice\_tp/pls\_input\_files

PLS team allowed members:

* JAINBH3
* MAHESRA1
* GURGNI1

#### Input file format

File format: Comma-separated values (CSV) file

File format description: CSV files are structured files containing fields (columns), separated by comma character. Excel spreadsheets can be directly converted to CSV files. CSV file consist of a header (first row) followed by data (starting from second row).

#### File structure

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Column Header Name** | **Data Type** | **Description** | **Values** | **Nullable** |
| country | String | Name of country in English | US, JAPAN, UK, AUSTRALIA, ITALY,BRAZIL, CANADA,CHINA, FRANCE, GERMANY, SPAIN | No |
| team\_name | String | Name of the team used to create summary | CV1 TEAM, SPECIALIST TEAM | Yes |
| primary\_team\_indicator | String | This will have value Y for team on which summary is created. For other team it will be N. If there is no team (example Italy), leave this column blank. | Y or N | Yes |
| product\_indication | String | Any product with indicator should be listed in this column | COSENTYX\_AS, COSENTYX\_PSA, COSENTYX\_PSO | Yes |
| group\_product\_name | String | Group name of product | COSENTYX, GALVUS | No |
| local\_product\_name | String | Local name of product | EQUA | Yes |
| territory\_code | String | Territory code | CALABRIA | No |
| territory\_name | String | Territory name | CALABRIA | No |
| month\_start\_date | Date | Start date of the month to which data belongs to. Format: MM-DD-YYYY | 1/1/2018 | No |
| product\_sales\_value | Decimal | Decimal (30,10): 10 decimal places | 1437.3345670 | Yes |
| product\_sales\_value\_curr | String | local currency | EUR | Yes |
| product\_sales\_units | Decimal | Decimal (30,10): 10 decimal places  Product sales in Units | 120.00 | Yes |
| product\_sales\_patients | IntegerType | Product sales: patient count | 200 | Yes |
| market\_sales\_value | Decimal | Decimal (30,10): 10 decimal places | 4300.5000000 | Yes |
| market\_sales\_value\_curr | String | local currency | EUR | Yes |
| market\_sales\_units | Decimal | Decimal (30,10): 10 decimal places Product sales in Units | 220.00 | Yes |
| market\_sales\_population | IntegerType | Total Patient count | 1200 | Yes |

Important remarks:

* If textual field contains comma in the value, whole value should be wrapped with quotes, example: "Some text, another some text, etc"
* Date values formatted as specified in Data Layout
* Each CSV file should have header with proper field/column ordering and naming
* All the values for Country, Team, Product Indication, Group Product Name, Local product name, Territory code, Territory Name and Currency should be in Upper case with underscore as applicable.

#### File name convention

|  |  |
| --- | --- |
| **File extension** | **Delimited text file (CSV)** |
| **File name pattern** | **<Country>\_<Quarter#>\_<Year>.csv** |
| **Remarks** | Ingestion year in YYYY format |
| **Examples** | ITALY\_Q1\_2018.csv, SPAIN\_Q2\_2018.csv |

#### File name and file structure validation

It’s assumed that file name and file structure is validated by responsible team/user before uploaded to FTP share location.

In case of wrong file structure:

Automated ingestion process will fail and/or yield improper data being ingested and exposed in Data Mart. Manual correction of data will be needed.

In case of wrong file name:

Automated ingestion process won’t fail if extension will be preserved but manual correction of data will be needed.

#### Data Mart

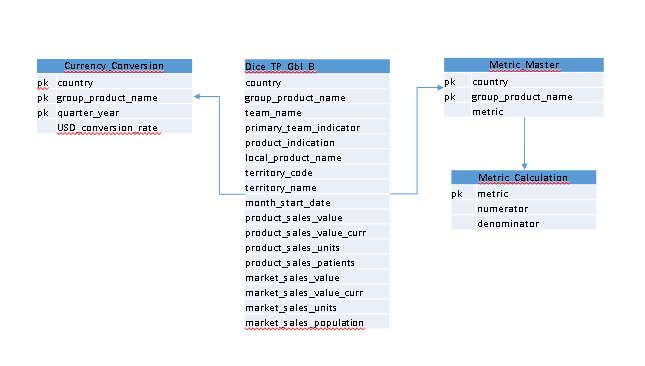
Ingested data will be stored and accumulated in single Apache Hive table and exposed for QlikSense for consumption.

Database name: **ph\_com\_b\_gbl\_dicetp**

Table name: **dice\_tp\_gbl**

#### Data Model

Data in table will be partitioned based on input file indentifier due to choosen reload startegy.



#### DDLs

|  |  |
| --- | --- |
| Raw | #-- Dropping Database if exists DROP DATABASE IF EXISTS ph\_com\_r\_gbl\_dice\_tp CASCADE;   #Creating hivedb database @ given location  CREATE DATABASE IF NOT EXISTS ph\_com\_r\_gbl\_dice\_tp comment 'ph\_com\_r\_gbl\_dice\_tp database ' location '/sdata/ph/com/r/ph\_com\_r\_gbl\_dice\_tp';   use ph\_com\_r\_gbl\_dice\_tp;  # Raw Layer # Create Table ph\_com\_r\_gbl\_dice\_tp.dice\_tp\_gbl( country string, team\_name string, primary\_team\_indicator string, product\_indication string, group\_product\_name string, local\_product\_name string, territory\_code string, territory\_name string, month\_start\_date string, product\_sales\_value Decimal(30,10), product\_sales\_value\_curr string, product\_sales\_units Decimal(30,10), product\_sales\_patients Decimal(30,10), market\_sales\_value Decimal(30,10), market\_sales\_value\_curr string, market\_sales\_units Decimal(30,10), market\_sales\_population Decimal(30,10), year\_month string ) PARTITIONED BY (icountry string, quarter string) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' stored AS PARQUET   #- Metrics Table --- create table ph\_com\_r\_gbl\_dice\_tp.metrics\_calculations( metric String, numerator String, denominator String) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' stored as PARQUET  #- Master\_Metric --- create table ph\_com\_r\_gbl\_dice\_tp.metrics\_master ( country string, group\_product\_name string, metric string ) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' stored as PARQUET  #- currency\_conversion --- create table ph\_com\_r\_gbl\_dice\_tp.currency\_conversion ( country string, group\_product\_name string, quarter\_year string, usd\_currency\_rate double) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' stored as PARQUET  # Error table Create Table error\_dice\_tp\_gbl( error\_in string, error\_row string, error\_reason string, error\_date string) stored AS PARQUET |
| Enriched | # Dropping Database if exists DROP DATABASE IF EXISTS ph\_com\_e\_gbl\_dice\_tp CASCADE;   #Creating hivedb database @ given location  CREATE DATABASE IF NOT EXISTS ph\_com\_e\_gbl\_dice\_tp comment 'ph\_com\_e\_gbl\_dice\_tp database ' location '/sdata/ph/com/e/ph\_com\_e\_gbl\_dice\_tp';   use ph\_com\_e\_gbl\_dice\_tp;   Create Table ph\_com\_e\_gbl\_dice\_tp.dice\_tp\_gbl( country string, team\_name string, primary\_team\_indicator string, product\_indication string, group\_product\_name string, local\_product\_name string, territory\_code string, territory\_name string, month\_start\_date string, product\_sales\_value Decimal(30,10), product\_sales\_value\_curr string, product\_sales\_units Decimal(30,10), product\_sales\_patients Decimal(30,10), market\_sales\_value Decimal(30,10), market\_sales\_value\_curr string, market\_sales\_units Decimal(30,10), market\_sales\_population Decimal(30,10), year\_month string ) PARTITIONED BY (icountry string, quarter string) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' stored AS PARQUET   #- Metrics Table --- create table ph\_com\_e\_gbl\_dice\_tp.metrics\_calculations( metric String, numerator String, denominator String) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' stored as PARQUET  create table ph\_com\_e\_gbl\_dice\_tp.metrics\_master ( country string, group\_product\_name string, metric string ) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' stored as PARQUET  create table ph\_com\_e\_gbl\_dice\_tp.currency\_conversion ( country string, group\_product\_name string, quarter\_year string, usd\_currency\_rate double) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' stored as PARQUET |
| DM | # Dropping Database if exists DROP DATABASE IF EXISTS ph\_com\_b\_gbl\_dice\_tp CASCADE;   #Creating hivedb database @ given location  CREATE DATABASE IF NOT EXISTS ph\_com\_b\_gbl\_dice\_tp comment 'ph\_com\_b\_gbl\_dice\_tp database ' location '/sdata/ph/com/b/ph\_com\_b\_gbl\_dice\_tp';   use ph\_com\_b\_gbl\_dice\_tp;  Create Table ph\_com\_b\_gbl\_dice\_tp.dice\_tp\_gbl( team\_name string, primary\_team\_indicator string, product\_indication string, group\_product\_name string, local\_product\_name string, territory\_code string, territory\_name string, month\_start\_date string, product\_sales\_value Decimal(30,10), product\_sales\_value\_curr string, product\_sales\_units Decimal(30,10), product\_sales\_patients Decimal(30,10), market\_sales\_value Decimal(30,10), market\_sales\_value\_curr string, market\_sales\_units Decimal(30,10), market\_sales\_population Decimal(30,10), year\_month string ) PARTITIONED BY (country string, quarter string) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' stored AS PARQUET    #- Metrics Table --- create table ph\_com\_b\_gbl\_dice\_tp.metrics\_calculations( metric String, numerator String, denominator String) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' stored as PARQUET    create table ph\_com\_b\_gbl\_dice\_tp.metrics\_master ( country string, group\_product\_name string, metric string ) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' stored as PARQUET    create table ph\_com\_b\_gbl\_dice\_tp.currency\_conversion ( country string, group\_product\_name string, quarter\_year string, usd\_currency\_rate double) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' stored as PARQUET |
|  |  |

# References, Attachments, Abbreviations/Acronyms

## References

| **#** | **ID** | **Title** | **Storage Location** |
| --- | --- | --- | --- |
|  | SOP-7037712 | GOP “Computerized System Validation” | ESOPS |
|  |  |  |  |

## Abbreviations / Acronyms

| **Abbreviation / Acronym** | **Description** |
| --- | --- |
| URS | User Requirement Specification |
| SIPOC | Suppliers/Inputs/Process/Outputs/Customers |
| SQL | Structured Query Language |
| TP | Territory potential |

# Publishing

This document is published in N/A.