**NBS/DCS/IT**

**Doc. ID:** CIT.SAL.DS.EV.001.DICE2AU.V1.0

**Title: Data Insights Customer Experience 2.0**

**DICE 2.0 AUSTRALIA**

**Application Clarity ID: 44730**

**Design Specification**

|  |  |  |
| --- | --- | --- |
| **Author:** | **Date:** | **Signature:** |
| Arvind Chimta  Software Engineer  EPAM Systems |  |  |
|  |  |  |
| **Reviewer:** | **Date:** | **Signature:** |
| Oleksandr Rudak  Lead Architect  EPAM Systems |  |  |
|  |  |  |
| **Approval:** | **Date:** | **Signature:** |
| Ng,Vivien  Integration PM  Novartis |  |  |

|  |  |  |
| --- | --- | --- |
| **Approval:** | **Date:** | **Signature:** |
| Omar KMA  IT System Owner  Novartis |  |  |

|  |  |  |
| --- | --- | --- |
| **Approval:** | **Date:** | **Signature:** |
| Prashant Mathpati  PQM  NBS IT |  |  |

**Document History**

|  |  |  |  |
| --- | --- | --- | --- |
| **Version** | **Date** | **Author** | **Comments** |
| 1.0 | 18-JUN-2019 | C. Arvind Kumar | Initial Version |

Table of Contents

[1. Introduction 4](#_Toc12967648)

[1.1 Purpose 4](#_Toc12967649)

[1.2 Prerequisites 4](#_Toc12967650)

[1.3 Reference to Standards 4](#_Toc12967651)

[1.4 Definitions 4](#_Toc12967652)

[2. Overview 4](#_Toc12967653)

[3. System Design & Customization 5](#_Toc12967654)

[3.1 General Specifications 6](#_Toc12967655)

[3.1.1 General Requirements 6](#_Toc12967656)

[3.1.2 Data Requirements 6](#_Toc12967657)

[3.1.2 Supported Spark Actions 7](#_Toc12967658)

[3.1.3 Notification service 8](#_Toc12967659)

[3.2 Data Flow & Data Model 8](#_Toc12967660)

[3.2.1 Data Flow Diagram 9](#_Toc12967661)

[3.2.2 Raw Layer Data Model 9](#_Toc12967662)

[3.2.3 Enhanced Layer Data Model 9](#_Toc12967663)

[3.2.4 Business Layer Data Model 11](#_Toc12967664)

[3.2.5 SI database change monitoring 12](#_Toc12967665)

[3.2.6 Solution folder structure 12](#_Toc12967666)

[3.3 Business Use Case / Module FS-GEN-04 Settings 12](#_Toc12967667)

[3.3.1 Feed to SI (PBS) 13](#_Toc12967668)

[3.4 Interfaces 14](#_Toc12967669)

[4. References, Attachments, Abbreviations/Acronyms 15](#_Toc12967670)

[4.1 References 15](#_Toc12967671)

[4.2 Attachments 16](#_Toc12967672)

[4.3 Abbreviations / Acronyms 16](#_Toc12967673)

[5. Publishing 16](#_Toc12967674)

# Introduction

## Purpose

This document is Design Specification (DS) for the DICE 2.0 – Australia project, which is being implemented on top of EVICO data platform.

## Prerequisites

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

* DICE 2.0 URS

## Reference to Standards

N/A

## Definitions

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

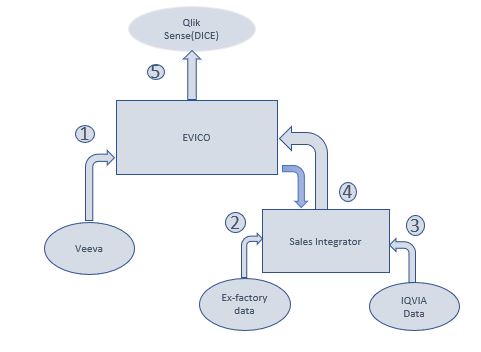
# Overview

The primary goal of the described solution is to introduce the Sales Integrator (SI), provide the one-way data channel between the SI and the EVICO platform, and to make the SI data marts available for the dashboard software QLIK Sense.

SI is a data warehouse, powered by Microsoft SQL Server (MSSQL), created and maintained by IQVIA company in partnership with the Novartis. Its primary goal is to integrate both Novartis’ own sales data and the third-party sales data from IQVIA sources; then to apply a particular set of business rules and calculate the required set of sales KPIs. The sales data, represented as a star-schema, and the pre-calculated KPIs represented shall be provided to EVICO for further consumption by either QLIK Sense and for ad hoc data analysis.

# System Design & Customization

Below is the data flow diagram that has been designed for DICE AU data processing.



1. The data on sales personnel and their activities, which has been introduced to EVICO in scope of DICE 1.0 project.
2. The sales data owned by Novartis. The process of ingestion of this data into SI is not covered here.
3. The sales data owned by IQVIA. The process of ingestion of this data into SI is not covered here.
4. The one-way data flow between EVICO and SI, that is the primary goal of the DICE 2.0 project. SI receives data from different sources and generates a star schema to be consumed by EVICO. This star schema includes required pre-calculated KPIs and are represented as data marts in EVICO. EVICO does automated monthly download of PBS data and supplies to SI.
5. QLIK Sense gets the KPI data, previously ingested to EVICO from the SI, represented as data marts.
6. System Design & Customization

Development is done in Agile methodology. Programming languages used

* Scala
* Bash

Frameworks

* Apache Spark
* Apache POI (to compose the Excel spreadsheet)

Infrastructure needed

* Apache Hadoop (HDFS + YARN)
* Apache Hive
* Apache Spark
* Sqoop

## General Specifications

This section covers the design / customization made based on the specifications given in the FS CIT.SAL.FS.EV.001.DICE2AU.V1.0

### General Requirements

| **DS-ID** | **Reference to FS** | **Function / Feature**  *(short title)* | **Detailed Specification** |
| --- | --- | --- | --- |
| DS-GEN-01 | FS-GEN-01 | Data Retention | EVICO to have Control-M task that executes Spark action From Si which is configured to ingest the data for the predefined period.  **Comments:**  For the details regarding configuration refer to Configuration File chapter. |
| DS-GEN-02 | FS-GEN-02 | Data Refresh | EVICO to have Control-M task that executes Spark actions ingest\_si\_to\_raw.sh ,raw\_to\_edm.sh and edm\_to\_dm.sh that is scheduled daily or can be triggered by ad-hoc request. |
| DS-GEN-03 | FS-GEN-05 | Download and copy PBS data into SI(sftp) | The Automation should be done to pull the data from the PBS website for the pharma and oncology item codes. |

### 3.1.2 Data Requirements

| **DS-ID** | **Reference to FS** | **Function / Feature**  *(short title)* | **Detailed Specification** |
| --- | --- | --- | --- |
| DS-DAT-01 | FS-DAT-01  FS-DAT-02  FS-DAT-03 | Build data marts for QS | EVICO to have DAGs which executes the following Spark actions:   * Ingest\_si\_to\_raw.sh * Register\_raw.sh * Raw\_to\_edm.sh * Register\_edm.sh * Raw\_to\_dm.sh * Register\_dm.sh   The tables are to be registered using Hive Table generation procedure.  **Comments:**  For data flows design approach and data transformations between the data layers refer to Data Flow & Data Model chapter.  For detailed data marts list and building logic for each refer to DICE Core chapter and the relevant Attachment. |
| DS-DAT-02 | FS-DAT-05 | Build feed for SI (PBS Data) | EVICO to have Control-M tasks that executes the following Spark actions:   * Run.sh   **Comments:**  For SI feed detailed design refer to Feed to SI chapter. |
| DS-DAT-04 | FS-DAT-04 | Notify data stewards | EVICO to trigger the Notification Service.  **Comments:**  Refer to Notification service chapter for the details. |
| DS-DAT-05 | FS-DAT-06 | Promote SI tables | EVICO to setup Control-M tasks that executes Spark action register\_dm.sh  **Comments:**  For SI tables promoting refer to Sales Integrator chapter. |

### Supported Spark Actions

#### Ingest\_si\_to\_raw.sh

We need to fetch data from SI, we do this through scoop job to fetch data from Sql server of SI Database.

#### Register\_raw.sh

We need to register the tables created from the above steps, this script registers the tables in impala in raw layer

#### Raw\_to\_edm.sh

We validate the data from raw layer by processing through the following steps

1. Data deduplication
2. Soft delete

After the above steps, we repartition the data according to the config parameters for repartition

And write it to the EDM Layer in parquet format.

#### Register\_edm.sh

We need to register the tables created from the above steps, this script registers the tables in impala in edm layer

#### Edm\_to\_dm.sh

We fetch the required data from raw area either transforming the data by joins or selecting the required columns and store in parquet format in datamarts.

#### Register\_dm.sh

We need to register the tables created from the above steps, this script registers the tables in impala in dm layer.

#### Memory Allocation

The memory for every Spark action is allocated via settings inside Shell script that run that action. The following settings are currently supported:

"spark.executor.memory": "10g",

"spark.driver.memory": "5g",

"spark.yarn.executor.memoryOverhead": "5g"

### Notification service

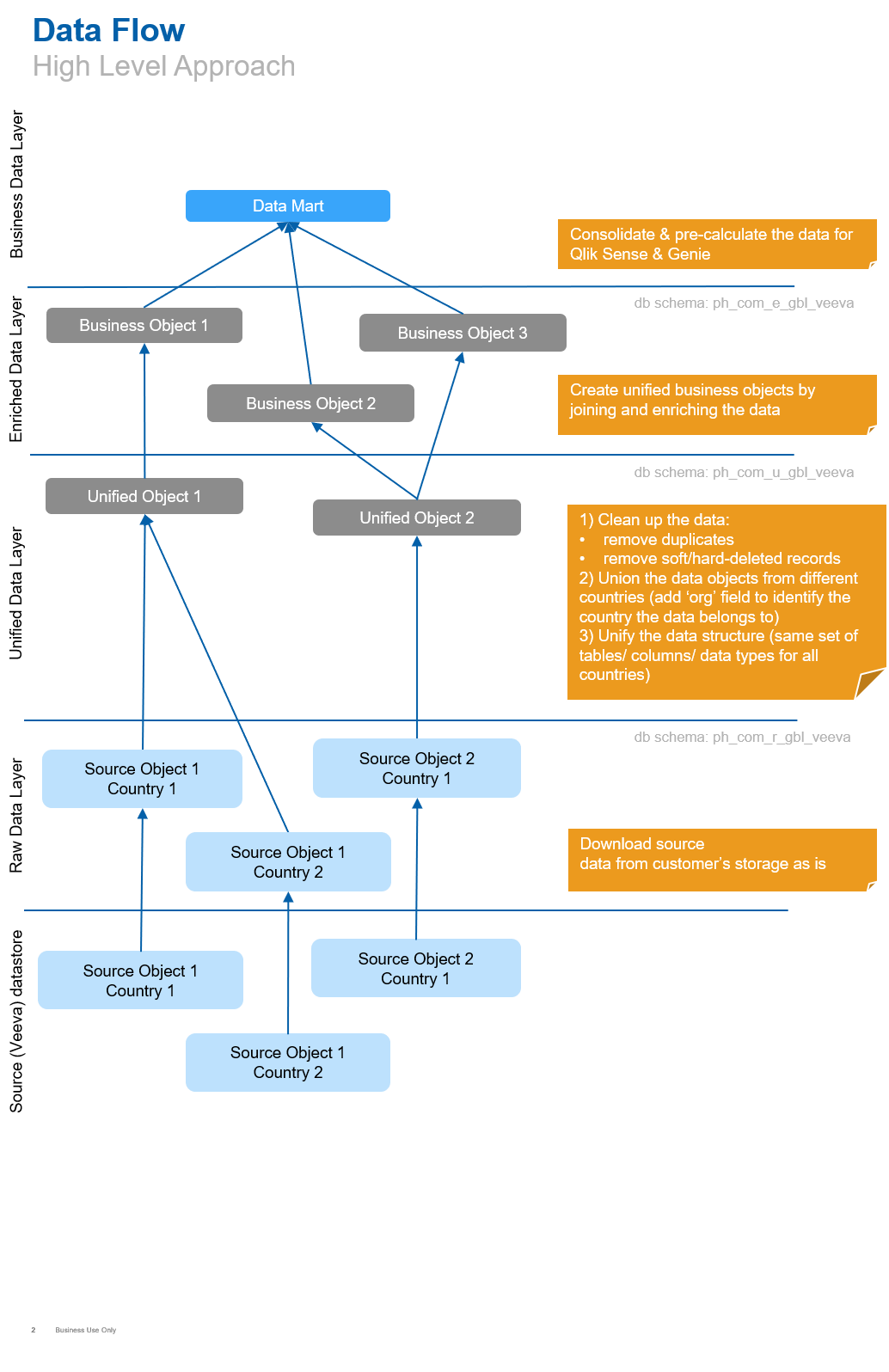
Control – M will notify the team of success or failure of the job with the logs attached to the required teams.

## Data Flow & Data Model

The system is designed to have three layers the data is transferred through. Each layer performs its own task clearing, modifying and structuring the data from source datastore (e.g. Veeva) up to business layer. This process makes the data possible to be used by the consumers (e.g. Qlik Sense) and further provided to the end users.

* Raw Data Layer – the data is copied as is from source SI datastore
* Enriched Data Layer - creates unified business objects by joining and enriching the data
* Business Data Layer - consolidate & pre-calculate the data for Qlik Sense & Genie

### Data Flow Diagram



### Raw Layer Data Model

The table below represents the new databases introduced in the EVICO Raw data layer.

| **Database** | **Description** |
| --- | --- |
| ph\_com\_r\_aus\_sales\_integrator | The raw data ingested from the Sales Integrator |

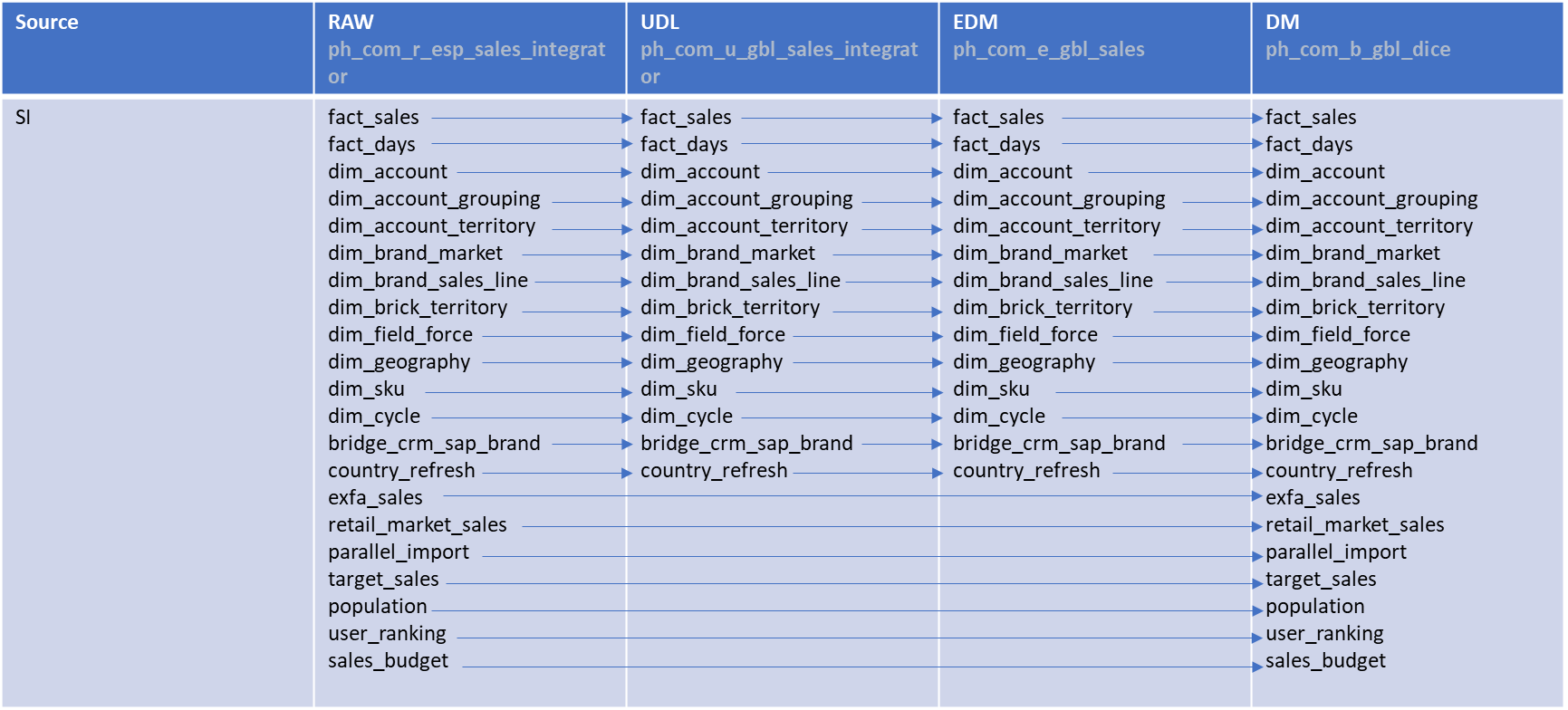
### Enhanced Layer Data Model

| **Database** | **Description** |
| --- | --- |
| ph\_com\_e\_gbl\_sales | * The tables with data derived from raw Sales Integrator data (initially, just the same tables copied to EDL without any transformation) * The data that needs to be pushed to SI: user-territory structure, product catalog, etc. * The cleansed Altas-Bajas data, and the results of the data cleansing / data quality checks: warnings, errors. |
| ph\_com\_b\_gbl\_dice | DICE core layer |

#### Sales Integrator

The table below represents the new databases introduced in the EVICO Enriched data layer.

| **Table Name** | **Description** |
| --- | --- |
| AU\_fact\_sales | Fact Sales - Sales data for all data sources |
| AU\_dim\_account | Dimension Account: account, address, banner group |
| AU\_Territory\_mapping | Territory Mapping details with territory\_code andsales\_line |
| AU\_dim\_product | Dimension Product: brand sku, brand, promoted\_flag, pack details and brand group |
| AU\_dim\_period | Dimension Period: cycle size, cycle start, cycle end, cycle name and reporting month |
| AU\_dim\_geography | Dimension Geography: brick, SRA, State and region |
| AU\_Market\_mapping | Dimension Market: market, business unit, product |
| AU\_crm\_product\_mapping | Mapping table to establish relation between crm and SI products |
| AU\_ims\_postcode\_to\_brick\_mapping | Mapping table for brick and related post code |
| AU\_ims\_territory\_structure\_onco\_pharma | Complete hierarchy at geography level |



#### DICE 2.0 Australia

Enhanced Data Model extends existing entities and creates new entities

| **Entity** | **Description** | **Design details** |
| --- | --- | --- |
| edm\_customer | Customer, new columns added | [R10] |
| edm\_product\_metrics | Product metrics, new entity | [ R10] |
| edm\_events | Events, new entity | [ R10] |
| edm\_hcp\_hco\_hierarchy | HCP and HCO hierarchy, new entity | [ R10] |
| edm\_tot | Time off territory, representatives added | [ R10] |
| edm\_survey | Survey, new entity | [R10] |
| edm\_survey\_response | Survey response, new entity | [ R10] |
| edm\_survey\_target | Survey target, new entity | [ R10] |
| edm\_survey\_question | Survey question, new entity | [ R10] |

### Business Layer Data Model

|  |  |
| --- | --- |
| **Database** | **Description** |
| ph\_com\_b\_dice | DICE business layer |

#### DICE Core

EVICO data marts created on the business layer

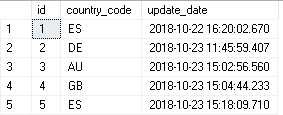
| **Data mart** | **Description** | **Design details** |
| --- | --- | --- |
| dm\_call | Historical calls based on account hierarchy and product | [R4] |
| dm\_call\_target | Historical call targets based on account hierarchy and product | [R9] |
| dm\_rep\_tot | Time off territory reported by representatives | [R7] |
| dm\_events | Medical events | [R3] |
| dm\_survey | Medical surveys | [R8] |
| dm\_flm\_tot | Time off territory reported by managers | [R7] |
| dm\_brand | Brand details | [R8] |
| dm\_brand\_metrics | Product details | [R8] |
| dm\_hcp\_hco\_hierarchy | Flattened representative, first line and second line managers hierarchy | [R5] |
| dm\_rep\_customer\_product\_call | Call information reference to rep,customer and calls | [R5] |
| dm\_rep\_hil | Representative Hil | N/A |
| dm\_flm\_hil | FLM | N/A |
| dm\_email | Details regarding Email | N/A |
| dm\_clm\_presentation | CLM Presentation Details | N/A |
| dm\_multichannel\_consent | Details regarding Multichannel consent | N/A |
| dm\_territory | Master Territory Details | N/A |
| dm\_customer | Customer Details | N/A |
| dm\_rep\_flm\_territory | Mapping Rep, FLM & Territory | N/A |
| dm\_customer\_portfolio\_tier | Details of Customer Portfolios | N/A |

### SI database change monitoring

SI country\_refresh is used to track data publishing on SI side. For the HyperCare period task to ingest SI data is scheduled via Control-M at agreed time.

#### Polling ingestion automation implementation

Control-M task start new Spark job, which polls table country\_refresh to look for today’s data publishing. Once data for the country is available, task starts data ingestion.



### Solution folder structure

|  |  |
| --- | --- |
| **Folder** | **Description** |
| Bin | The scripts, to launch every Spark job |
| Conf | The configuration files that have to be deployed along with the application |
| hive-scripts | Contains the scripts to create the tables in the EVICO Hive database.  Grouped by the data layers (raw, edl, etc). |
| Src | The source code for the Spark command-line application to launch the data pipelining and transformation jobs. |
| <root>/deploy.sh | The script to deploy the solution to the required folder |
| <root>/dist.xml | The descriptor file, required by the Maven plugin which creates the distribution archive. Contains the folders to take into the archive. |

## Business Use Case / Module FS-GEN-04 Settings

This requirement is described in the above section in details.

| **DS-ID** | **Reference to FS** | **Function / Feature**  *(short title)* | **Detailed Specification** |
| --- | --- | --- | --- |
| DS-GEN-01 | FS-GEN-04 | Ingest SI data model into EVICO | Data model build for Sales KPIs (in SI system) and for Activity KPIs (in EVICO system) should have data fields required to build QlikSense reports |

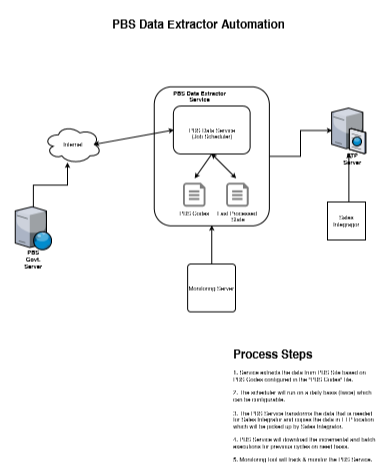
Source Code Repository : [R16]

| **DS-ID** | **Reference to FS** | **Function / Feature**  *(short title)* | **Detailed Specification** |
| --- | --- | --- | --- |
| DS-GEN-02 | FS-GEN-05 | Download and ftp PBS data into SI | The Automation should be done to pull the data from the PBS website for the pharma and oncology item codes. |

PBS Automation data flow diagram

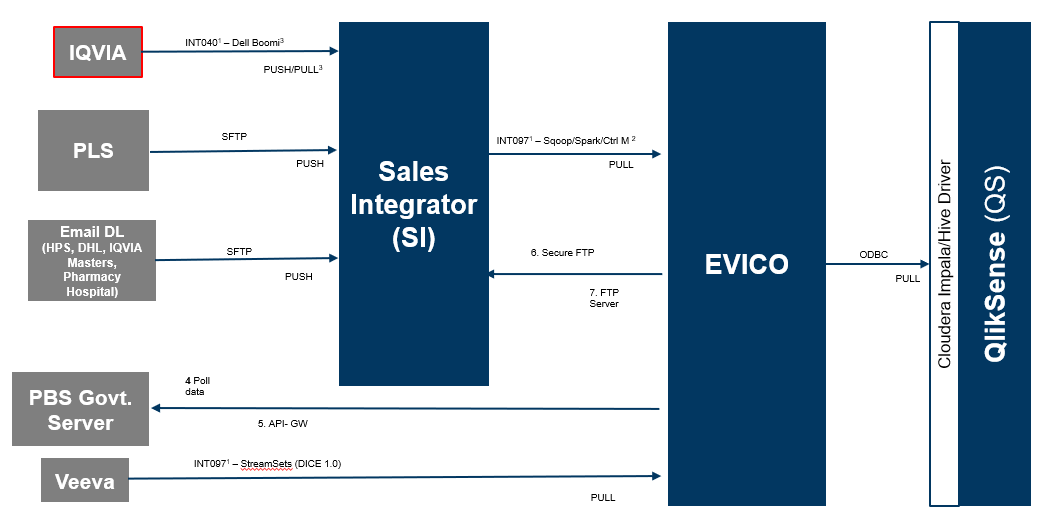
### Feed to SI (PBS)

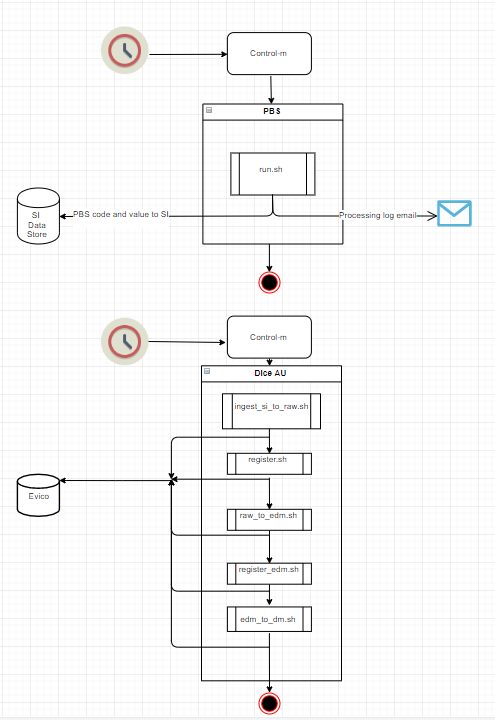
PBS application is a spring boot app designed to fetch the PBS codes and values from the website and feed it to the SI Datastore. Following are the actions performed

1. Read the PBS mapping file to fetch the required PBS codes to be downloaded.
2. Download PBS data for a batch of 20 PBS Codes.
3. Extract and make a report of the values downloaded
4. Write pbs sales data to ftp
5. Write the PBS processing log to the ftp.
6. Send the success or failure email with processing log attached to the mail to respective teams
7. The scheduling is done using control M jobs. 

## Interfaces

This section covers the interface design / customization made based on the specifications given in the FS CIT.SAL.FS.EV.001.DICE2AU.V1.0





| **DS-ID** | **Reference to FS** | **Function / Feature**  *(short title)* | **Detailed Specification** |
| --- | --- | --- | --- |
| DS-INT-01 | FS-INT-01 | Data flow: Veeva to EVICO | The data ingestion solution is covered by DICE 1.0 DS (CIT.ANA.DS.001.DICE.V1.0 chapter 3.2 Data Ingestion). |
| DS-INT-02 | FS-INT-02 | Data flow: SI to EVICO | EVICO to have Control-M task that executes Spark action ingest\_si\_raw. |

Source Code Repository: GIT\_REPO\_DICE\_AU [R16] & GIT\_REPO\_DICE\_PBS [R17]

# References, Attachments, Abbreviations/Acronyms

## References

| **#** | **ID** | **Title** | **Storage Location** |
| --- | --- | --- | --- |
|  | SOP-7037712 | GOP “Computerized System Validation” | ESOPS |
|  | CIT.SAL.URS.AU.001.DICE2.0 | User Requirements Specification | Project SharePoint |
|  | CIT.SAL.DS.EV.001.A001.DICE2ES.V2.0 | Events | Project SharePoint |
|  | CIT.SAL.DS.EV.001.A002.DICE2ES.V2.0 | Call | Project SharePoint |
|  | CIT.SAL.DS.EV.001.A003.DICE2ES.V2.0 | HCO & HCP Hierarchy | Project SharePoint |
|  | CIT.SAL.DS.EV.001.A004.DICE2ES.V2.0 | Brand & Brand Metrics | Project SharePoint |
|  | CIT.SAL.DS.EV.001.A005.DICE2ES.V2.0 | Time off Territory Representatives and Managers | Project SharePoint |
|  | CIT.SAL.DS.EV.001.A006.DICE2ES.V2.0 | Survey | Project SharePoint |
|  | CIT.SAL.DS.EV.001.A007.DICE2ES.V2.0 | Call Target | Project SharePoint |
|  | CIT.SAL.DS.EV.001.A005.DICE2ES.V2.0 | Time off Territory Representatives and Managers | Project SharePoint |
|  | CIT.SAL.DS.EV.001.A006.DICE2ES.V2.0 | Survey | Project SharePoint |
|  | CIT.SAL.DS.EV.001.A007.DICE2ES.V2.0 | Call Target | Project SharePoint |
|  | CIT.SAL.DS.EV.001.A008.DICE2ES.V2.0 | EDM Structure & mapping to UDL | Project SharePoint |
|  | CIT.SAL.DS.EV.001.A009.DICE2ES.V2.0 | SI Data Model | Project SharePoint |
|  | CIT.SAL.DS.EV.001.A010.DICE2ES.V2.0 | DICE 2.0 Data Flow Design | Project SharePoint |
|  | GIT\_REPO\_DICE\_AU | Git Repository | <https://gitlab.evico.novartis.net/project_ph_com_dice20/project_ph_com_dice20_pbs> |
|  | GIT\_REPO\_PBS | Git Repository | <https://gitlab.evico.novartis.net/project_ph_com_dice20/project_ph_com_dice20_pbs> |

## Attachments

The attachments are located in the same folder as the current document.

| **#** | **ID** | **Title** | **Storage Location** |
| --- | --- | --- | --- |
|  | CIT.SAL.FS.EV.001.DICE2AU.V1.0 | Functional Specification | [Attachment](https://my.novartis.net/:w:/r/personal/omarkm1_novartis_net/_layouts/15/Doc.aspx?sourcedoc=%7B0A9B0614-387B-4524-ACEA-2A6BB866BA04%7D&file=CIT.SAL.FS.EV.001.DICE2AU.V1.0.docx&action=default&mobileredirect=true&cid=c0506c12-254b-40b8-b431-2b6cf277d9fa) |

## Abbreviations / Acronyms

| **Abbreviation / Acronym** | **Description** |
| --- | --- |
| COTS | Commercial Off The Shelf |
| FS | Functional Specification |
| SIPOC | Suppliers/Inputs/Process/Outputs/Customers |
| SQL | Structured Query Language |
| URS | User Requirement Specification |

# Publishing

This document is published in [CIT PDR Documents :  Dashboard GCIT\_PH\_ES\_DICE 2](https://share.novartis.net/sites/CITPDR/CITPDRDocuments/Forms/CITPDRDocumentSet/docsethomepage.aspx?ID=1854&FolderCTID=0x0120D52000E76F6F1248E7B24997EBC1AE9398F50500839677B94FBF6940BE3C90EDADFA552F&List=d0504470-dcd6-411e-ad7c-d3e5df29d54a&RootFolder=%2Fsites%2FCITPDR%2FCITPDRDocuments%2FDashboard%20GCIT%5FPH%5FES%5FDICE%202&RecSrc=%2Fsites%2FCITPDR%2FCITPDRDocuments%2FDashboard%20GCIT%5FPH%5FES%5FDICE%202).