# Data quality: Step Description

This macro checks quality of data in DDS and creates a table with alerts about possible problems in the data.

## **Code Realization Requirements**

Use python

# 3. Input Data

## **3.1 IN\_PRODUCT\_ATTR**

|  |  |
| --- | --- |
| **DDS.IN\_PRODUCT\_ATTR** | |
| Column Name | Description |
| **PRODUCT\_ID** | Product ID (the lowest level of the product hierarchy) |
| **PRODUCT\_ATTR\_NAME** | Product attribute name |
| **PRODUCT\_ATTR\_VALUE** | Product attribute value |

## **3.2 IN\_LOCATION\_ATTR**

|  |  |
| --- | --- |
| **DDS.IN\_LOCATION\_ATTR** | |
| Column Name | Description |
| **LOCATION\_ID** | Location ID |
| **LOCATION\_ATTR\_NAME** | Location attribute name |
| **LOCATION\_ATTR\_VALUE** | Location attribute value |

## **3.3 IN\_CUSTOMER\_ATTR**

|  |  |
| --- | --- |
| **DDS.IN\_CUSTOMER\_ATTR** | |
| Column Name | Description |
| **CUSTOMER\_ID** | Customer ID |
| **CUSTOMER\_ATTR\_NAME** | Customer attribute name |
| **CUSTOMER\_ATTR\_VALUE** | Customer attribute value |

## **3.4 IN\_DISTR\_CHANNEL\_ATTR**

|  |  |
| --- | --- |
| **DDS.IN\_DISTR\_CHANNEL\_ATTR** | |
| Column Name | Description |
| **DISTR\_CHANNEL\_ID** | Distribution channel ID |
| **DISTR\_CHANNEL\_ATTR\_NAME** | Distribution channel attribute name |
| **DISTR\_CHANNEL\_ATTR\_VALUE** | Distribution channel attribute value |

## **3.5 PRODUCT LIFE-CYCLE**

|  |  |
| --- | --- |
| **PRODUCT LIFE-CYCLE** | |
| **Column Name** | **Description** |
| **PRODUCT\_ID** | Product ID |
| **LOCATION\_LVL\_ID** | Location hierarchy element ID |
| **CUSTOMER\_LVL\_ID** | Customer hierarchy element ID |
| **DISTR\_CHANNEL\_LVL\_ID** | Distribution CHANNEL hierarchy element ID |
| **PERIOD\_START\_DT** | Period start date |
| **PERIOD\_END\_DT** | Period end date |
| **PRODUCT\_SUCCESSOR\_ID** | Successor Product ID |
| **RELATION\_SHARE** | Number |
| **PERIOD\_TYPE** | Active', 'Blocked', 'End-of-life' |

## **3.6 LOCATION LIFE-CYCLE**

|  |  |
| --- | --- |
| **LOCATION LIFE-CYCLE** | |
| **Column Name** | **Description** |
| **LOCATION\_ID** | Location ID |
| **CUSTOMER\_LVL\_ID** | Customer hierarchy element ID |
| **PRODUCT\_LVL\_ID** | Product hierarchy element ID |
| **DISTR\_CHANNEL\_LVL\_ID** | Distribution CHANNEL hierarchy element ID |
| **PERIOD\_START\_DT** | Period start date |
| **PERIOD\_END\_DT** | Period end date |
| **LOCATION\_SUCCESSOR\_ID** | Successor Location ID |
| **RELATION\_SHARE** | Number |
| **PERIOD\_TYPE** | Closure reason: 'reconstruction', 're-branding' etc. |

## **3.7 CUSTOMER LIFE-CYCLE**

|  |  |
| --- | --- |
| **CUSTOMER LIFE-CYCLE** | |
| **Column Name** | **Description** |
| **CUSTOMER\_ID** | Customer ID |
| **LOCATION\_LVL\_ID** | Location hierarchy element ID |
| **PRODUCT\_LVL\_ID** | Product hierarchy element ID |
| **DISTR\_CHANNEL\_LVL\_ID** | Distribution CHANNEL hierarchy element ID |
| **PERIOD\_START\_DT** | Period start date |
| **PERIOD\_END\_DT** | Period end date |
| **CUSTOMER\_SUCCESSOR\_ID** | Successor Customer ID |
| **RELATION\_SHARE** | Number |
| **PERIOD\_TYPE** | active, blocked, end-of-life |

## **3.8 IN\_PRICE**

Information regarding planned price.

|  |  |
| --- | --- |
| **DDS.IN\_PRICE** | |
| **Column Name** | **Description** |
| **PRODUCT\_ID** | Product ID (the lowest level of the product hierarchy) |
| **LOCATION\_ID** | Location ID |
| **CUSTOMER\_ID** | Customer ID |
| **DISTR\_CHANNEL\_ID** | Distribution Channel ID |
| **PERIOD\_START\_DT** | Period start date |
| **PERIOD\_END\_DT** | Period end date |
| **PRICE** | Price |
| **PRICE\_TYPE** | Only one of the following price types: 1 - regular, 2 - promotional, 3 - markdown, 4 - other |

## **3.9 IN\_PROMO**

Promo history data containing the following fields is used as an input.

|  |  |
| --- | --- |
| **DDS.IN\_PROMO** | |
| **Column Name** | **Description** |
| **PROMO\_ID** | Promo ID |
| **PRODUCT\_ID** | Product ID (the lowest level of the product hierarchy) |
| **LOCATION\_ID** | Location ID |
| **CUSTOMER\_ID** | Customer ID |
| **DISTR\_CHANNEL\_ID** | Distribution Channel ID |
| **PERIOD\_START\_DT** | Period start date |
| **PERIOD\_END\_DT** | Period end date |
| **PROMO\_PRICE** | Promo price |
| **PROMO\_TYPE** | Promo type code |
| **PROMO\_MECHANICS** | Promo mechanics code (taken from the promo mechanics types table) |

## **3.10 SALES**

Sales information regarding the past till last known day of the history .

|  |  |
| --- | --- |
| **DDS.IN\_SALES** | |
| **Column Name** | **Description** |
| **PRODUCT\_ID** | Product ID (the lowest level of the product hierarchy) |
| **LOCATION\_ID** | Location ID |
| **PERIOD\_DT** | Date of sales (calendar day) |
| **SALES\_QTY** | Total sales in units per day (w/o returns) |

## **3.11 STOCK**

Inventory history data containing the following fields is used as an input:

|  |  |
| --- | --- |
| **DDS.IN\_STOCK** | |
| **Column Name** | **Description** |
| **PRODUCT\_ID** | Product ID (the lowest level of the product hierarchy) |
| **LOCATION\_ID** | Location ID |
| **PERIOD\_START\_DT** | Date of sales (calendar day) |
| **STOCK\_QTY** | Stock qty |

## **3.12 SELL\_IN**

Initial CPG sales history data

|  |  |
| --- | --- |
| **DDS.IN\_SELL\_IN** | |
| **Column Name** | **Description** |
| **PRODUCT\_ID** | Product ID (the lowest level of the product hierarchy) |
| **LOCATION\_ID** | Location from ID |
| **CUSTOMER\_ID** | Customer ID |
| **DISTR\_CHANNEL\_ID** | Distribution CHANNEL ID |
| **LOCATION\_TO\_ID** | Location to Id |
| **PERIOD\_DT** | Date of SELL-IN (calendar day) |
| **ORDER\_QTY** | Total ORDER in units per day (w/o returns) |
| **ORDER\_AMOUNT** | Total ORDER revenue (with VAT) per day (w/o returns) |
| **SHIPMENTS\_QTY** | Total SHIPMENT in units per day (w/o returns) |
| **SHIPMENTS\_AMOUNT** | Total SHIPMENT revenue (with VAT) per day (w/o returns) |
| **INVOICE\_QTY** | Total INVOICE in units per day (w/o returns) |
| **INVOICE\_AMOUNT** | Total INVOICE revenue (with VAT) per day (w/o returns) |
| **RETURNS\_QTY** | Total returns in units per day |
| **RETURNS\_AMOUNT** | Total returns amount (with VAT) per day |

|  |  |
| --- | --- |
| **DDS.IN\_SELL\_OUT** | |
| **Column Name** | **Description** |
| **PRODUCT\_ID** | Product ID (the lowest level of the product hierarchy) |
| **LOCATION\_ID** | Location from ID |
| **CUSTOMER\_ID** | Customer ID |
| **DISTR\_CHANNEL\_ID** | Distribution CHANNEL ID |
| **LOCATION\_TO\_ID** | Location to Id |
| **PERIOD\_DT** | Date of SELL-OUT (calendar day) |
| **ORDER\_QTY** | Total ORDER in units per day (w/o returns) |
| **ORDER\_AMOUNT** | Total ORDER revenue (with VAT) per day (w/o returns) |
| **SHIPMENTS\_QTY** | Total SHIPMENT in units per day (w/o returns) |
| **SHIPMENTS\_AMOUNT** | Total SHIPMENT revenue (with VAT) per day (w/o returns) |
| **INVOICE\_QTY** | Total INVOICE in units per day (w/o returns) |
| **INVOICE\_AMOUNT** | Total INVOICE revenue (with VAT) per day (w/o returns) |
| **RETURNS\_QTY** | Total returns in units per day |
| **RETURNS\_AMOUNT** | Total returns amount (with VAT) per day |
| **PROMO\_FLG** | Promo flag |
| **PROMO\_ID** | Promo ID |
| **COST** | Unit cost |

## **3.14 PRODUCT**

## **3.15 LOCATION**

|  |  |
| --- | --- |
| **DDS.IN\_PRODUCT** | |
| Column Name | Description |
| PRODUCT\_LVL\_ID1 | Hierarchy level ID 1 |
| PRODUCT\_LVL\_ID2 | Hierarchy level ID 2 |
| PRODUCT\_LVL\_ID3 | Hierarchy level ID 3 |
| PRODUCT\_LVL\_ID4 | Hierarchy level ID 4 |
| PRODUCT\_LVL\_ID5 | Hierarchy level ID 5 |
| PRODUCT\_LVL\_ID6 | Hierarchy level ID 6 |
| PRODUCT\_LVL\_ID7 | Hierarchy level ID 7 |
| PRODUCT\_ID | Product ID (the lowest level of the product hierarchy) |

## **3.16 CUSTOMER**

|  |  |
| --- | --- |
| **DDS.IN\_LOCATION** | |
| Column Name | Description |
| LOCATION\_LVL\_ID1 | Hierarchy level ID 1 |
| LOCATION\_LVL\_ID2 | Hierarchy level ID 2 |
| LOCATION\_LVL\_ID3 | Hierarchy level ID 3 |
| LOCATION\_LVL\_ID4 | Hierarchy level ID 4 |
| LOCATION\_LVL\_ID5 | Hierarchy level ID 5 |
| LOCATION\_ID | Location ID |

## **3.17 DISTR\_CHANNEL**

|  |  |
| --- | --- |
| **DDS.IN\_CUSTOMER** | |
| Column Name | Description |
| CUSTOMER\_LVL\_ID1 | Hierarchy level ID 1 |
| CUSTOMER\_LVL\_ID2 | Hierarchy level ID 2 |
| CUSTOMER\_LVL\_ID3 | Hierarchy level ID 3 |
| CUSTOMER\_LVL\_ID4 | Hierarchy level ID 4 |
| CUSTOMER\_LVL\_ID5 | Hierarchy level ID 5 |
| CUSTOMER\_ID | Customer ID |

|  |  |
| --- | --- |
| **DDS.IN\_DISTR\_CHANNEL** | |
| Column Name | Description |
| DISTR\_CHANNEL\_LVL\_ID1 | Hierarchy level ID 1 |
| DISTR\_CHANNEL\_ID | Distribution CHANNEL ID |

## **3.18 ASSORT\_MATRIX**

Assortment matrix Sales information regarding the past till last known day of the history.

|  |  |
| --- | --- |
| **DDS.IN\_ASSORT\_MATRIX** | |
| Column Name | Description |
| **LOCATION\_ID** | Location from ID |
| **PRODUCT\_ID** | Product ID (the lowest level of the product hierarchy) |
| **CUSTOMER\_ID** | Customer ID |
| **DISTR\_CHANNEL\_ID** | Distribution CHANNEL ID |
| **START\_DT** | Assort Start date (e.g. sales for Retailer/ shipment for CPG /shelf life for Fashion) |
| **END\_DT** | Assort End date (e.g. sales for Retailer/ shipment for CPG /shelf life for Fashion) |
| **STATUS** | Status (Active, No longer merchandised etc.) |

## **3.19 DQ\_PARAMETERS**

|  |  |
| --- | --- |
| **CNTL.DQ\_PARAMETERS** | |
| Column Name | Description |
| **check\_id** | id of check |
| **check\_name** | name of check |
| **client** | Name of client |
| **input\_table** | the table to which the check is applied |
| **var\_name** | Variable name for check |
| **var\_value** | Value of variable |

# 4. Algorithm Definition

**4.1 Check necessary tables in DDS**

Checking that DDS lib has all necessary tables. If the required table does not exist it is created.

List of tables for check:

* IN\_PRODUCT\_ATTR
* IN\_LOCATION\_ATTR
* IN\_CUSTOMER\_ATTR
* IN\_DISTR\_CHANNEL\_ATTR
* IN\_PROMO\_ATTR
* IN\_CUSTOMER\_LIFE
* IN\_LOCATION\_LIFE
* IN\_PRODUCT\_LIFE
* IN\_SALES
* IN\_STOCK
* IN\_SELL\_IN
* IN\_SELL\_OUT
* IN\_PRODUCT
* IN\_LOCATION
* IN\_CUSTOMER
* IN\_DISTR\_CHANNEL

**4.2 Check quality of data**

At this step the necessary data quality checks are applied. The output table contains information about possible problems with the data.

**4.2.1 Check values range**

Сhecks that the column values are not greater than the specified value (for example, that the prices are not negative)

4.2.1.1 IN\_PRICE

Checking for negative prices in IN\_PRICE table. Rows with negative values are written to the data\_quality\_output table.

4.2.1.2 IN\_PROMO

Checking for negative prices in IN\_PROMO table. Rows with negative values are written to the data\_quality\_output table.

4.2.1.3 IN\_SALES

Checking for negative sales in IN\_SALES table. Rows with negative values are written to the data\_quality\_output table.

4.2.1.4 IN\_SELLOUT

Checking for negative sales in IN\_SELLOUT table. Rows with negative values are written to the data\_quality\_output table.

4.2.1.5 IN\_SELL\_IN

Checking for negative sales in IN\_SELL\_IN table. Rows with negative values are written to the data\_quality\_output table.

4.2.1.6 IN\_STOCK

Checking for negative stocks IN\_STOCK table. Rows with negative values are written to the data\_quality\_output table.

**4.2.2 Check cross consistency**

Checks that there are no key fields in the first table that are missing in the second table.

Checking each pair of tables:

* IN\_SALES
* IN\_STOCK
* IN\_PRICE
* IN\_PRODUCT
* IN\_LOCATION
* IN\_CUSTOMER
* IN\_DISTR\_CHANNEL

4.2.2.1 Determine the intersection of the columns in two tables - name of key columns (product\_id, location\_id, customer\_id, distr\_channel\_id) which exists in both tables

4.2.2.2 Determine the set of column values that present in the first table and don’t exists in the second table

4.2.2.3 Adding rows from 4.2.2.2 to the data\_quality\_output table.

The original tables are not changed.

**4.2.3 Check time cross consistency**

Checks tables for time cross-consistency (i.e. finding prodict\_id - location\_id pairs that have been in the SALES table and not in the STOCK table for some period)

* Compare IN\_SALES and IN\_STOCK tables
* Compare IN\_STOCK and IN\_SALES tables
* Compare IN\_STOCK and IN\_ASSORT\_MATRIX tables
* Compare IN\_SALES and IN\_ASSORT\_MATRIX tables
* Compare IN\_STOCK and IN\_PRICE tables
* Compare IN\_SALES and IN\_PRICE tables

4.2.3.1. Determine the intersection of the columns in two tables (product\_id, location\_id or product\_id – location\_id) and time column (period\_dt or period\_start\_dt)

4.2.3.2 Determine the set of column values (product\_id – period\_dt, location\_id – period\_dt, period\_dt – location\_id – product\_id) which present in the first table and don’t present in the second table

4.2.3.4 Determine the line that are in the first table and are not in the second in more than threshold share of lines. Threshold read from parameters (var\_name for this check\_id)

4.2.3.4 Adding matching rows to the data\_qualuty\_output table

The original tables are not changed.

**4.3 Prepare output table**

Output table: data\_quality\_output

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |

|  |  |
| --- | --- |
| Column Name | Description |
| PRODUCT\_LVL\_ID | Element id of product hierarchy |
| LOCATION\_LVL\_ID | Element id of location hierarchy |
| CUSTOMER\_LVL\_ID | Element id of customer hierarchy |
| DISTR\_CHANNEL\_LVL\_ID | Element id of distr\_channel hierarchy |
| PRODUCT\_LVL | Level of product channel hierarchy |
| LOCATION\_LVL | Level of location channel hierarchy |
| CUSTOMER\_LVL | Level of customer channel hierarchy |
| DISTR\_CHANNEL\_LVL | Level of distribution channel hierarchy |
| PERIOD\_DT | Start period of date in al\_time\_lvl format |
| INPUT\_COLUMN | Column for check in the input table |
| INPUT\_TABLE | Input table vor check |
| INPUT\_VALUE | Value parametr |
| WARNING\_TYPE | Warning type |
| warning | Text of warning |