Yuvaraj Munirathinam

Yuvaraj.munirathinam@hcl.com]

LOW LEVEL DESIGN

NA TRANSFIX

Phase 3

Contents

[1. Overview 4](#_Toc22212959)

[2. Business process 4](#_Toc22212960)

[3. Scope 4](#_Toc22212961)

[3.1. In scope 4](#_Toc22212962)

[3.2. Out of scope 5](#_Toc22212963)

[4. Data Source 5](#_Toc22212964)

[4.1. File Extracts 5](#_Toc22212965)

[4.2. Internal BD Cluster 5](#_Toc22212966)

[5. Data model and data location 6](#_Toc22212967)

[5.1. Storage 6](#_Toc22212968)

[5.1.1. Permanent 6](#_Toc22212969)

[5.2. Tables 6](#_Toc22212970)

[5.2.1. Databases and sizing 6](#_Toc22212971)

[5.2.2. Staging objects 7](#_Toc22212972)

[5.2.3. Dictionary/static tables 7](#_Toc22212973)

[5.2.4. Target tables 7](#_Toc22212974)

[5.2.5. Views 8](#_Toc22212975)

[5.2.6. Logical/physical Data Model 9](#_Toc22212976)

[6. Data processing 10](#_Toc22212977)

[6.1. Processing triggering/scheduling 10](#_Toc22212978)

[6.1.1. Dependency 10](#_Toc22212979)

[6.2. Airflow process (DAG) 10](#_Toc22212980)

[6.2.1. PROD\_TransfixCalc 10](#_Toc22212981)

[6.2.2. PROD\_TransfixFap 12](#_Toc22212982)

[6.2.3. PROD\_TransfixRgvr 13](#_Toc22212983)

[6.2.4. PROD\_TransfixMd 13](#_Toc22212984)

[6.3. Airflow DAG variables 15](#_Toc22212985)

[6.3.1. Transfix variables 15](#_Toc22212986)

[6.3.2. Global variables 15](#_Toc22212987)

[6.3.3. Refresh tables on databricks user workspace variables 16](#_Toc22212988)

[6.4. Start Databricks job from airflow 16](#_Toc22212989)

[6.5. Running SQLs with hive params on spark (Databricks) 17](#_Toc22212990)

[6.6. Export data to excel 18](#_Toc22212991)

[6.7. Uploading excel file to SP 18](#_Toc22212992)

[6.8. Data archiving 19](#_Toc22212993)

[6.9. Email notification 19](#_Toc22212994)

[6.10. Passwords/tokens/secrets 19](#_Toc22212995)

[6.11. Refreshing tables on user databricks workspace 19](#_Toc22212996)

[7. Application interfaces 20](#_Toc22212997)

[8. Operational procedures 20](#_Toc22212998)

[4. Backup 21](#_Toc22212999)

[5. Contact list 22](#_Toc22213000)

[6. Hardware Platform 23](#_Toc22213001)

[7. Application user 24](#_Toc22213002)

[8. Service Level Agreement (SLA) Parameters 24](#_Toc22213003)

[9. Access and incident management 25](#_Toc22213004)

[10. Attachments 26](#_Toc22213005)

[10.1. Hive codes 26](#_Toc22213006)

[10.2. Python codes 26](#_Toc22213007)

Document revision history

|  |  |  |  |
| --- | --- | --- | --- |
| Date | Version | Author | Comment |
| 30-Mar-2018 | 1.0 | Robert Michalak | Initial version |
| 20-Jul-2018 | 1.2 | Robert Michalak | Added more details for processes |
| 05-Sep-2018 | 1.45 | Robert Michalak | Changed process diagram, updated monitoring chapter |
| 24-Oct-2018 | 1.50 | Robert Michalak | Removed parts from Transportation Visiblity project |
| 17-Jan-2019 | 1.60 | Robert Michalak | Removed parts from Transportation Visiblity project |
| 03-Jun-2019 | 1.70 | Robert Michalak | Merged documentation from phase1&2 and update Azure changes. |
| 22-Jul-2019 | 1.80 | Robert Michalak | Update document (Azure changes) |
| 12-Aug-2020 | 1.90 | Mark Besmonte | Update Document:  -Changed git export\_table\_2\_excel.py link to Azure DevOps Link  -Changed git upload\_file\_to\_sp.py link to Azure DevOps Link  -Changed git refresh\_tables\_metadata link to Azure DevOps  -Changed "What to do (easiest way)" value for Recreate Tables "Situation" in "Operational procedures" Gitlab to Azure DevOps  -Changed repository link in "Backup"  -Changed "Hardware Platform" Databricks Link  -Changed "Cluster configuration and component versions"  -Changed "Azure service information" Airflow Link  -Changed "Software GitLab link" |
| 17-10-2023 | 2.0 | Yuvaraj Munirathinam | Updated document ADF changes |

# Overview

This is extension of Transportation Visibility project, which adds needed report tables for current T-View reports in SAP BW. After Transfix, T-View reports on BW won’t be used anymore. Data gathered in BD will be used mostly in Tableau but in the future data can be accessed in other tools like Knime, Alteryx, Jethro for an further exploration.

Transfix delivers following data for reports:

* TFS (detailed and summary)
* TAC (detailed and summary)
* On Time Data
* LOT
* CSOT
* IOT
* Carrier Dashboard
* FAP
* VFR
* VFR US/CA
* RGVR
* Key Factors / Weekly Network Scorecard

Report usage and frequency: actual usage will vary by report – there are daily and weekly reports. Direct data usage in unknown at the time.

Criticality – all workflows should be completed every day because data is used by Tableau application.

# Business process

This application is using mainly Transportation Visibility tables to create reporting tables.

All tables are used by any stakeholder interested in transportation data. At this moment there is one business case transfix that reports data for NA Transportation Operation team.

Criticality of the data is low because it won’t stop the business work processes. However, any data outage should be resolved as quickly as possible, preferably within 24 hours (response within 8 hours). Source data should be up to date (max 24 hours data latency).

Data should be ready no later than 10 am EST (Cincinnati time) daily.

# Scope

This application is transforming internal BD tables to load target tables for business.

## In scope

Activities:

1. Data transformation to load target tables.

All activities would be maintained by P&G Big Data operation team.

## Out of scope

Any additional data transformations are out of scope.

# Data Source

## File Extracts

There are no source file extracts.

## Internal BD Cluster

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Database | Table/view name | ~Size [MB] | Type | Description |
| ap\_transfix\_tv\_na | operational\_tariff\_star | 1200 | Fact | Operation Tariffs |
| dp\_trans\_vsblt\_bw | shpmt\_cust\_898\_day\_fdim | 400000 | Dim | Customer 898 |
| dp\_trans\_vsblt\_bw | contract\_adherence\_na\_merged\_star | 1700 | Fact | CAD data |
| dp\_trans\_vsblt\_bw | freight\_stats\_na\_merged\_star | 5000 | Fact | TFS data |
| dp\_trans\_vsblt\_bw | on\_time\_arriv\_shpmt\_custshpmt\_na\_merged\_star | 2000 | Fact | OTD data |
| dp\_trans\_vsblt\_bw | otd\_vfr\_na\_star | 20000 | Fact | VFR data |
| dp\_trans\_vsblt\_bw | shipping\_location\_na\_dim | 13 | Dim | Shippings Locations |
| dp\_trans\_vsblt\_bw | tdcval\_na\_dim | 1 | Dim | TDC Val |
| dp\_trans\_vsblt\_bw | tender\_acceptance\_na\_merged\_star | 220 | Fact | TAC data |
| dp\_trans\_vsblt\_bw | transport\_type\_na\_dim | 1 | Dim | Transport type |
| dp\_trans\_vsblt\_bw | leo\_truck\_report\_lkp | 7 | Lkp | LEO TRUCK REPORT LKP table |
| dp\_trans\_vsblt\_bw | leo\_vehicle\_maintenance\_lkp | 1 | Lkp | LEO VEHICLE MAINTENANCE LKP table |
| dp\_trans\_vsblt\_bw | order\_shipment\_linkage\_zvdf\_lkp | 1 | Lkp | Linkage for SAP Order and Shipmen number |
| dp\_trans\_vsblt\_bw | tms\_unload\_method\_dest\_zone\_lkp | 1 | Lkp | TMS Method of Unload and Destination Zone |
| dp\_trans\_vsblt\_bw | sambc\_master\_lkp | 10 | Lkp | Service As Measured By Customer Indicators |
| dp\_trans\_vsblt\_bw | csot\_update\_reason\_lkp | 10 | Lkp | CSOT Updated Reasons |
| dp\_masterdata\_g11 | ausp\_characteristic\_values | 40 | Dim | Characteristic Values |
| dp\_masterdata\_g11 | cabn\_characteristic | 10 | Dim | Characteristic |
| dp\_masterdata\_g11 | klah\_class\_header\_data | 10 | Dim | Class Header Data |
| dp\_masterdata\_g11 | ksml\_characteristics\_of\_a\_class | 10 | Dim | Characteristics of a Class |
| dp\_masterdata\_g11 | kssk\_allocation\_table\_object\_to\_class | 10 | Dim | Allocation Table Object to Class |
| dp\_masterdata\_g11 | lfa1\_vendor\_master\_general\_section | 120 | Fact | Vendor Master (General Section) |
| dp\_masterdata\_g11 | makt\_material\_descriptions | 160 | Fact | Material Description |
| dp\_masterdata\_g11 | mara\_general\_material\_data | 80 | Fact | General Material Data |
| dp\_masterdata\_g11 | marm\_units\_of\_measure\_for\_material | 210 | Dim | Units of Measure for Material |
| dp\_masterdata\_g11 | ztxxpt0103\_allowed\_attribute\_values | 30 | Dim | Allowed Attribute Values |
| dp\_masterdata\_g11 | ztxxpt0104\_material\_attribute\_value\_assignment | 55 | Dim | Material Attribute Value Assignment |
| dp\_masterdata\_g11 | ztxxptatva\_material\_attribute\_values\_table | 20 | Dim | Material attribute values table |
| dp\_masterdata\_g11 | ztxxptcpik\_general\_pick\_list\_table | 1 | Dim | General Pick list Table |
| dp\_masterdata\_g11 | ztxxptmelf\_fpc\_external\_life\_cycle | 150 | Fact | FPC External Life Cycle |
| dp\_masterdata\_g11 | ztxxptnass\_product\_hierarchy\_nodes\_association | 460 | Fact | Product hierarchy |
| dp\_masterdata\_g11 | ztxx\_ptwfconfig\_product\_wf\_screen\_field\_configuration | 1 | Dim | Product Wf Screen Field Configuration |
| dp\_masterdata\_g11 | ztxx\_ptwfreq\_product\_workflow\_request\_table | 30 | Dim | Product Workflow Request Table |
| dp\_osi\_na\_ecc | adr6\_e\_mail\_addresses\_business\_address\_services | 40 | Dim | E-Mail Addresses (Business Address Services) |
| dp\_osi\_na\_ecc | adrc\_addresses\_business\_address\_services | 430 | Fact | Addresses (Business Address Services) |
| dp\_osi\_na\_ecc | kna1\_general\_data\_in\_customer\_master | 280 | Fact | General Data in Customer Master |
| dp\_osi\_na\_ecc | lfa1\_vendor\_master\_general\_section | 36 | Dim | Vendor Master (General Section) |
| dp\_osi\_na\_ecc | lfb1\_vendor\_master\_company\_code | 4 | Dim | Vendor Master (Company Code) |
| dp\_osi\_na\_ecc | t001k\_valuation\_area | 1 | Dim | Valuation Area |
| dp\_osi\_na\_ecc | t001w\_plants\_branches | 1 | Dim | Plants/Branches |
| dp\_osi\_na\_ecc | t001\_company\_codes | 1 | Dim | Company Codes |
| dp\_osi\_na\_ecc | t005\_countries | 1 | Dim | Countries |
| dp\_osi\_na\_ecc | tvko\_organizational\_unit\_sales\_organizations | 1 | Dim | Organizational Unit: Sales Organizations |
| dp\_osi\_na\_ecc | tvst\_organizational\_unit\_shipping\_points | 1 | Dim | Organizational Unit: Shipping Points |
| dp\_osi\_na\_ecc | twlad\_determination\_of\_address\_from\_plant\_and\_storage\_location | 1 | Dim | Determination of Address from Plant and Storage Location |
| dp\_osi\_na\_ecc | ztxxptatvs\_material\_attribute\_values\_table\_per\_sales\_area | 40 | Fact | Material attribute values table per Sales Area |
| dp\_osi\_na\_ecc | customer\_dim | 70 | Fact | Customer |
| RDS | cust\_hier\_dim | 10000 | Dim | Customer hierarchies |
| RDS | exchg\_rate\_fct | 17 | Fact | Exchange rates |
| RDS | cust\_dim | 430 | Dim | Customer Dimension |
| RDS | trade\_chanl\_hier\_dim | 1 | Dim | Trade Channel Hierarchy |

# Data model and data location

## Storage

### Permanent

|  |  |  |
| --- | --- | --- |
| ADLS Folder | Montly Delta | Current Size |
| abfss://etl@dppsdatahubx45bab07e.dfs.core.windows.net/transfix/ |  |  |
| abfss://lightrefined@dppsdatahubx45bab07e.dfs.core.windows.net/transfix/prod/final/ | 2 GB | 8 GB (for ph. II) |

## Tables

### Databases and sizing

|  |  |  |
| --- | --- | --- |
| Database Name | Description | ADLS Folder |
| ap\_transfix\_tv\_na | Stores application views and tables | abfss://lightrefined@dppsdatahubx45bab07e.dfs.core.windows.net/transfix/prod/final/ |

### Staging objects

None.

### Dictionary/static tables

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Database Name | Table Name | Function | Type | Description |
| ap\_transfix\_tv\_na | destination\_channel\_customer\_lkp | Lookup | Lkp | Destination Channel |
| ap\_transfix\_tv\_na | on\_time\_codes\_aot\_reason\_lkp | Lookup | Lkp | On Time Codes |
| ap\_transfix\_tv\_na | origin\_gbu\_lkp | Lookup | Lkp | Origin GBU code |
| ap\_transfix\_tv\_na | ship\_to\_pgp\_flag\_lkp | Lookup | Lkp | Ship To PGP flag |
|  |  |  |  |  |

### Target tables

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Database Name | Table Name | Function | Type | Description |
| ap\_transfix\_tv\_na | cust\_hierarchy656\_na\_lkp | Lookup | Lkp | Customer Hierarchy (656) |
| ap\_transfix\_tv\_na | cust\_na\_dim | Dimension | Dim | Customer Master Data |
| ap\_transfix\_tv\_na | plant\_na\_dim | Dimension | Dim | Plant Master Data |
| ap\_transfix\_tv\_na | prod1\_na\_dim | Dimension | Dim | Product1 Master Data |
| ap\_transfix\_tv\_na | sales\_org\_na\_dim | Dimension | Dim | Sales Organization Master Data |
| ap\_transfix\_tv\_na | shipping\_point\_na\_dim | Dimension | Dim | Shipping Point Master Data |
| ap\_transfix\_tv\_na | storage\_location\_na\_dim | Dimension | Dim | Storage Location data |
| ap\_transfix\_tv\_na | vendor\_na\_dim | Dimension | Dim | Vendor Master Data |
|  |  |  |  |  |
| ap\_transfix\_tv\_na | tac\_technical\_name\_star | Star / Fact | Star | Fact Table with TAC data and costs calculations |
| ap\_transfix\_tv\_na | tfs\_technical\_name\_star | Star / Fact | Star | Fact Table with TFS data and costs calculations |
| ap\_transfix\_tv\_na | tac\_tender\_star | Star / Fact | Star | Table calculated from KNIME script |
| ap\_transfix\_tv\_na | tac\_tender\_summary\_star | Star / Fact | Star | Table calculated from KNIME script |
| ap\_transfix\_tv\_na | tfs\_acsrl\_star | Star / Fact | Star | TFS Accesorial |
| ap\_transfix\_tv\_na | tfs\_subsector\_cost\_star | Star / Fact | Star | TFS Subsector Costs |
| ap\_transfix\_tv\_na | on\_time\_data\_hub\_star | Star / Fact | Star | On Time Data Hub |
| ap\_transfix\_tv\_na | vfr\_data\_hub\_star | Star / Fact | Star | VFR Data Hub (Vehicle Fill Rate) |
| ap\_transfix\_tv\_na | fap\_star | Star / Fact | Star | FAP Report (Freight Audit Payment) |
| ap\_transfix\_tv\_na | lot\_star | Star / Fact | Star | LOT Report (Loaded On Time) |
| ap\_transfix\_tv\_na | rgvr\_star | Star / Fact | Star | RGVR Report (Routing Guide Variance Reconciiliation) |
| ap\_transfix\_tv\_na | csot\_star | Star / Fact | Star | CSOT Report (Customer Specific On Time) |
| ap\_transfix\_tv\_na | iot\_star | Star / Fact | Star | IOT Report (Interplant On Time) |
| ap\_transfix\_tv\_na | tac\_shpmt\_detail\_star | Star / Fact | Star | Shipment Details Report from TAC data |
| ap\_transfix\_tv\_na | tac\_lane\_detail\_star | Star / Fact | Star | Lane Details Report from TAC data |
| ap\_transfix\_tv\_na | destination\_channel\_customer\_lkp | Lookup | Lkp | Destination - Channel – Customer |
| ap\_transfix\_tv\_na | on\_time\_codes\_aot\_reason\_lkp | Lookup | Lkp | On Time Codes - AOT reason |
| ap\_transfix\_tv\_na | origin\_gbu\_lkp | Lookup | Lkp | Origin – GBU |
| ap\_transfix\_tv\_na | ship\_to\_pgp\_flag\_lkp | Lookup | Lkp | Ship to - PGP Flag |
| ap\_transfix\_tv\_na | vfr\_load\_agg\_star | Star / Fact | Star | VFR Aggregated data |
| ap\_transfix\_tv\_na | weekly\_network\_sccrd\_star | Star/Fact | Star | Weekly Network Scorecard |
| ap\_transfix\_tv\_na | weekly\_network\_sccrd\_agg\_star | Star/Fact | Star | Weekly Network Scorecard Aggregated |

### Views

Views do not have any additional logic and are built to meet business expectations of columns naming convention.

|  |  |  |  |
| --- | --- | --- | --- |
| Database Name | View Name | Type | Description |
| ap\_transfix\_tv\_na | sales\_org\_md | View | Selects data from sales\_org\_md\_dim table |
| ap\_transfix\_tv\_na | prod1\_md | View | Selects data from product1\_md\_na\_dim table |
| ap\_transfix\_tv\_na | prod2\_md | View | Selects data from ztxxptmelf\_fpc\_external\_life\_cycle table |
| ap\_transfix\_tv\_na | country\_md | View | Selects data from t005\_countries table |
| ap\_transfix\_tv\_na | special\_attributes | View | Selects data from ztxxptatvs\_material\_attribute\_values\_table\_per\_sales\_area table |
| ap\_transfix\_tv\_na | shipping\_point\_md | View | Selects data from shipping\_point\_md\_na\_dim table |
| ap\_transfix\_tv\_na | storage\_location | View | Selects data from storage\_location\_na\_dim table |
| ap\_transfix\_tv\_na | cust\_md | View | Selects data from customer\_md\_na\_dim table |
| ap\_transfix\_tv\_na | plant\_md | View | Selects data from plant\_md\_na\_dim table |
| ap\_transfix\_tv\_na | vendor\_md | View | Selects data from vendor\_md\_na\_dim table |
| ap\_transfix\_tv\_na | tdcval\_md | View | Selects data from tdcval\_md\_na\_dim table |
| ap\_transfix\_tv\_na | transport\_type\_md | View | Selects data from transport\_type\_md\_na\_dim table |
| ap\_transfix\_tv\_na | shipping\_location | View | Selects data from shipping\_location\_na\_dim table |
| ap\_transfix\_tv\_na | cust\_hierarchy656\_md | View | Selects data from customer\_hierarchy656\_md\_na\_lkp table |
| ap\_transfix\_tv\_na | tac\_vw | View | Selects data from tac\_technical\_name\_star table |
| ap\_transfix\_tv\_na | accessorial | View | Selects data from tfs\_acsrl\_star |
| ap\_transfix\_tv\_na | subsectorcosts | View | Selects data from tfs\_subsector\_cost\_star |
| ap\_transfix\_tv\_na | tac\_tender\_pg | View | Selects data from tac\_tender\_star |
| ap\_transfix\_tv\_na | tac\_tender\_pg\_summary | View | Selects data from tac\_tender\_summary\_star |
| ap\_transfix\_tv\_na | otd\_data\_hub | View | Selects data from on\_time\_data\_hub\_star |
| ap\_transfix\_tv\_na | fap | View | Selects data from fap\_star |
| ap\_transfix\_tv\_na | lot | View | Selects data from lot\_star |
| ap\_transfix\_tv\_na | lot\_all\_shipments | View | Selects data from on\_time\_data\_hub\_star |
| ap\_transfix\_tv\_na | csot | View | Selects data from csot\_star |
| ap\_transfix\_tv\_na | csot\_vw | View | Selects data from csot\_star but without filtering |
| ap\_transfix\_tv\_na | Iot | View | Selects data from lot\_star |
| ap\_transfix\_tv\_na | iot\_vw | View | Selects data from iot\_star |
| ap\_transfix\_tv\_na | monster\_report\_vw | View | Monster Report |
| ap\_transfix\_tv\_na | rgvr | View | Selects data from rgvr\_star |
| ap\_transfix\_tv\_na | weekly\_network\_scorecard | View | Selects data from weekly\_network\_sccrd\_star |
| ap\_transfix\_tv\_na | weekly\_network\_scorecard\_agg | View | Selects data from weekly\_network\_sccrd\_agg\_star |
| ap\_transfix\_tv\_na | vfr | View | Selects data from vfr\_data\_hub\_star |
| ap\_transfix\_tv\_na | vfr\_load\_agg\_vw | View | Selects data from vfr\_load\_agg\_star |
| ap\_transfix\_tv\_na | vfr\_us\_ca | View | Selects data from vfr\_data\_hub\_star |
| ap\_transfix\_tv\_na | carrier\_dashboards\_tacshipmentdetailraw | View | Selects data from tac\_shpmt\_detail\_star |
| ap\_transfix\_tv\_na | carrier\_dashboards\_taclanedetailraw | View | Selects data from tac\_lane\_detail\_star |
| ap\_transfix\_tv\_na | carrier\_dashboards\_csotraw | View | Selects data from csot\_star |
| ap\_transfix\_tv\_na | carrier\_dashboards\_iotraw | View | Selects data from iot\_star |
| ap\_transfix\_tv\_na | carrier\_dashboards\_lotraw | View | Selects data from csot\_star |

### Logical/physical Data Model



Please see the azure devops repository for latest version.

# Data processing

## Processing triggering/scheduling

Transfix processing runs every day. There are only 1 data sets that run once a month:

* RGVR – second day of month

### Dependency

The availability of data pipe source tables are checked in the *load\_process\_control.load\_process\_plc* table. The table contains dates when each data pipe table has been loaded.

There is also dependency between Transfix DAGs:

|  |  |  |
| --- | --- | --- |
| **Parent DAG/task** | **Child DAG** | **Description** |
| PROD\_TransfixCalc  transfixOtdDataHubLoadTask | PROD\_TransfixFap | Child DAG will run if task transfixOtdDataHubLoadTask from PROD\_TransfixCalc DAG will finish successfully for the same day. |
| PROD\_TransfixCalc.  transfixOtdDataHubLoadTask | PROD\_TransfixRgvr | Child DAG will run if task transfixOtdDataHubLoadTask from PROD\_TransfixCalc DAG will finish successfully for the same day. |
| PROD\_TransfixMD  transfixMdCustomerLoadTask  transfixMdVendorLoadTask  transfixMdCustomerHier656LoadTask | PROD\_TransfixCalc | Child DAG will run if tasks in PROD\_TransfixMD DAG will finish successfully for the same day. |

## ADF pipeline

There are Transfix 3 processes:

|  |  |  |
| --- | --- | --- |
| **DAG name** | **Description** | **Schedule** |
| Transfix\_calc\_prod | Main Transfix processing | 0 4 \* \* \* |
| TransfixRgvr\_prod | RGVR processing | 0 4 2 \* \* |
| Transfix\_MD\_prod | Master Data processing | 0 4 \* \* \* |

### Transfix\_calc\_prod

Below you can find detailed description of the pipeline:

|  |  |  |
| --- | --- | --- |
| **Type** | **Activity name** | **Description** |
| Start | N/A | Start action |
| sensorTariff | Sensor activity | Waits until ap\_transfix\_tv\_na.operational\_tariff\_star is loaded for a particular day. The task uses timeouts and intervals from global airflow variables. |
| sensorCfrCust898 | Sensor activity | Waits until dp\_direct\_shpmt\_cfr.shpmt\_cust\_898\_day\_fdim is loaded for a particular day. The task uses timeouts and intervals from global airflow variables. |
| sensorEccVendor | Sensor activity | Waits until dp\_osi\_na\_ecc.Vendor\_dim is loaded for a particular day. The task uses timeouts and intervals from global airflow variables. |
| sensorEccCustomer | Sensor activity | Waits until dp\_osi\_na\_ecc.Customer\_dim is loaded for a particular day. The task uses timeouts and intervals from global airflow variables. |
| sensorRdsExchgRate | Sensor activity | Waits until rds.exchg\_rate\_fct is loaded for a particular day. The task uses timeouts and intervals from global airflow variables. |
| sensorTvbCad | Sensor activity | Waits until dp\_trans\_vsblt\_bw.contract\_adherence\_na\_merged\_star is loaded for a particular day. The task uses timeouts and intervals from global airflow variables. |
| sensorTvbFreightStats | Sensor activity | Waits until dp\_trans\_vsblt\_bw.freight\_stats\_na\_merged\_star is loaded for a particular day. The task uses timeouts and intervals from global airflow variables. |
| sensorTvbOnTime | Sensor activity | Waits until dp\_trans\_vsblt\_bw.on\_time\_arriv\_shpmt\_custshpmt\_na\_merged\_star is loaded for a particular day. The task uses timeouts and intervals from global airflow variables. |
| sensorTvbTenderAcceptance | Sensor activity | Waits until dp\_trans\_vsblt\_bw.tender\_acceptance\_na\_merged\_star is loaded for a particular day. The task uses timeouts and intervals from global airflow variables. |
| sensorTvbOtdVfr | Sensor activity | Waits until dp\_trans\_vsblt\_bw.otd\_vfr\_na\_star is loaded for a particular day. The task uses timeouts and intervals from global airflow variables. |
| sensorTvbShippingLocation | Sensor activity | Waits until dp\_trans\_vsblt\_bw.shipping\_location\_na\_dim is loaded for a particular day. The task uses timeouts and intervals from global airflow variables. |
| transfixTfsLoadTask | Execute TFS  tfs\_processing.hql | Calculates TFS data. |
| sensorMDCustHier656 | Sensor MD | Waits until task transfixMdVendorLoadTask in dag PROD\_TransfixMD will finish for the same day. |
| sensorMDCust | Sensor MD | Waits until task transfixMdCustomerLoadTask in dag PROD\_TransfixMD will finish for the same day. |
| sensorMDVendor | Sensor MD | Waits until task transfixMdVendorLoadTask in dag PROD\_TransfixMD will finish for the same day. |
| transfixTacLoadTask | Execute TAC  tac\_processing.hql | Calculates TAC data. |
| transfixDsAccesorialLoadTask | Execute tfs acsrl  ds\_accessorial.hql | Calculates aggregated data form TFS data for accessorial. |
| transfixDsSubsectorLoadTask | Execute subsector  ds\_subsectorcosts.hql | Calculates aggregated data form TFS data for subsector costs. |
| transfixDsTacLoadTask | Execute Tax tender  ds\_tac\_tender\_pg.hql | Calculates aggregated tender data form TAC data. |
| transfix1stJoin | N/A | Join multiple processings. |
| sensorRdsCust | Sensor 2 | Waits until rds.cust\_dim is loaded for a particular day. The task uses timeouts and intervals from global airflow variables. |
| sensorRdsTradeChanlHier | Sensor 2 | Waits until rds.trade\_chanl\_hier\_dim is loaded for a particular day. The task uses timeouts and intervals from global airflow variables. |
| sensorTvbTmsUnloadMethodDestZone | Sensor 2 | Waits until dp\_trans\_vsblt\_bw.tms\_unload\_method\_dest\_zone\_lkp is loaded for a particular day. The task uses timeouts and intervals from global airflow variables. |
| sensorTvbOnTimeArrivShpmtCustshpmt | Sensor 2 | Waits until dp\_trans\_vsblt\_bw. on\_time\_arriv\_shpmt\_custshpmt\_na\_star is loaded for a particular day. The task uses timeouts and intervals from global airflow variables. |
| transfixOtdDataHubLoadTask | Execute OTD  main.py  load\_on\_time\_data\_hub\_star.py | Calculates On Time Data Hub. |
| transfixCsotLoadTask | Execute CSOT  main.py  load\_csot\_star.py | Calculates CSOT Mart. |
| transfixLotLoadTask | Execute LOT  main.py  load\_lot\_star.py | Calculates LOT Mart. |
| transfixVfrDataHubLoadTask | Execute VFR data  main.py  load\_vfr\_data\_hub\_star.py | Calculates VFR Data Hub. |
| transfixIotLoadTask | Execute IOT  main.py  load\_iot\_star.py | Calculates IOT Mart. |
| transfixVfrAggLoadTask | PROD\_transfix2.0\_jobs.json  main.py  load\_vfr\_agg\_star.py | Calculates VFR Aggregated table. |
| transfixCarrierLoadTask | Execute Tac carrier  main.py  load\_carrier.py | Calculates Carrier Dashboard data. |
| transfixWeeklyNetworkScorecardLoadTask | Execute weekly network  main.py  load\_network\_scorecard\_star.py | Calculates Weekly Network Scorecard Mart. |
| Execute TAC Detail | Execute TAC Detailmain.py  load\_network\_scorecard\_star.py | Calculates Tac detail |
| refreshTablesTask | refresh\_tables\_metadata.py | Refreshes all tables in Transfix Database |
| endTask | N/A | End action |

Detailed data processing diagrams:



### TransfixRgvr\_prod

Below you can find detailed description of the DAG:

|  |  |  |
| --- | --- | --- |
| **Task name** | **Script/JSON template** | **Description** |
| Start | N/A | Start action |
| sensorTfxOtd |  | Task will complete if task transfixOtdDataHubLoadTask from PROD\_TransfixCalc DAG will finish successfully for the same day. |
| transfixRgvrLoadTask | PROD\_transfix2.0\_jobs.json  main.py  load\_rgvr\_star.py | Calculates Rgvr Mart data. |
| refreshTableTask | refresh\_tables\_metadata.py | Refreshes Rgvr Star Table |
| End | N/A | End action |

Detailed data processing diagrams:



### Transfix\_MD\_prod

Below you can find detailed description of the DAG:

|  |  |  |
| --- | --- | --- |
| **Task name** | **Script/JSON template** | **Description** |
| Start | N/A | Start action |
| waitForClusterTask | main.py | Waits for the Transfix Cluster to Start |
| sensorG11Ausp | dep\_load\_process\_plc.sql | Waits until dd\_masterdata\_g11.Ausp is loaded for a particular day. The task uses timeouts and intervals from global airflow variables. |
| sensorG11Cabn | dep\_load\_process\_plc.sql | Waits until dd\_masterdata\_g11.Cabn is loaded for a particular day. The task uses timeouts and intervals from global airflow variables. |
| sensorG11Klah | dep\_load\_process\_plc.sql | Waits until dd\_masterdata\_g11.Klah is loaded for a particular day. The task uses timeouts and intervals from global airflow variables. |
| sensorG11Ksml | dep\_load\_process\_plc.sql | Waits until dd\_masterdata\_g11.Ksml is loaded for a particular day. The task uses timeouts and intervals from global airflow variables. |
| sensorG11Kssk | dep\_load\_process\_plc.sql | Waits until dd\_masterdata\_g11.Kssk is loaded for a particular day. The task uses timeouts and intervals from global airflow variables. |
| sensorG11Lfa1 | dep\_load\_process\_plc.sql | Waits until dd\_masterdata\_g11.Lfa1 is loaded for a particular day. The task uses timeouts and intervals from global airflow variables. |
| sensorG11Ztxx\_ptwfconfig | dep\_load\_process\_plc.sql | Waits until dd\_masterdata\_g11.Ztxx\_ptwfconfig is loaded for a particular day. The task uses timeouts and intervals from global airflow variables. |
| sensorG11Ztxx\_ptwfreq | dep\_load\_process\_plc.sql | Waits until dd\_masterdata\_g11.Ztxx\_ptwfreq is loaded for a particular day. The task uses timeouts and intervals from global airflow variables. |
| sensorG11Ztxxptatva | dep\_load\_process\_plc.sql | Waits until dd\_masterdata\_g11.Ztxxptatva is loaded for a particular day. The task uses timeouts and intervals from global airflow variables. |
| sensorG11Ztxxptcpik | dep\_load\_process\_plc.sql | Waits until dd\_masterdata\_g11.Ztxxptcpik is loaded for a particular day. The task uses timeouts and intervals from global airflow variables. |
| sensorG11Ztxxptnass | dep\_load\_process\_plc.sql | Waits until dd\_masterdata\_g11.Ztxxptnass is loaded for a particular day. The task uses timeouts and intervals from global airflow variables. |
| sensorG11ProdAttrValue | dep\_load\_process\_plc.sql | Waits until dp\_masterdata\_g11.prod\_attr\_value\_lkp is loaded for a particular day. The task uses timeouts and intervals from global airflow variables. |
| sensorG11ProdAttrValueAssign | dep\_load\_process\_plc.sql | Waits until dp\_masterdata\_g11.prod\_attr\_value\_assign\_dim is loaded for a particular day. The task uses timeouts and intervals from global airflow variables. |
| sensorG11FpcExternalLifeCycle | dep\_load\_process\_plc.sql | Waits until dp\_masterdata\_g11.Fpc\_external\_life\_cycle\_dim is loaded for a particular day. The task uses timeouts and intervals from global airflow variables. |
| sensorG11ProdDesc | dep\_load\_process\_plc.sql | Waits until dp\_masterdata\_g11.prod\_desc\_lkp is loaded for a particular day. The task uses timeouts and intervals from global airflow variables. |
| sensorG11Prod | dep\_load\_process\_plc.sql | Waits until dp\_masterdata\_g11.prod\_dim is loaded for a particular day. The task uses timeouts and intervals from global airflow variables. |
| sensorG11ProdUom | dep\_load\_process\_plc.sql | Waits until dp\_masterdata\_g11.prod\_uom\_lkp is loaded for a particular day. The task uses timeouts and intervals from global airflow variables. |
| transfixMdProduct1LoadTask | PROD\_transfix\_jobs.json  product1.hql | Calculates product 1 master data. |
| sensorEccPlant | dep\_load\_process\_plc.sql | Waits until dp\_osi\_na\_ecc.Plant\_dim is loaded for a particular day. The task uses timeouts and intervals from global airflow variables. |
| sensorEccVltnArea | dep\_load\_process\_plc.sql | Waits until dp\_osi\_na\_ecc.Vltn\_area\_dim is loaded for a particular day. The task uses timeouts and intervals from global airflow variables. |
| transfixMdPlantLoadTask | PROD\_transfix\_jobs.json  plant.hql | Calculates plant master data. |
| sensorEccShipPoint | dep\_load\_process\_plc.sql | Waits until dp\_osi\_na\_ecc.Ship\_point\_lkp is loaded for a particular day. The task uses timeouts and intervals from global airflow variables. |
| sensorEccAdrc | dep\_load\_process\_plc.sql | Waits until dd\_osi\_na\_ecc.Adrc is loaded for a particular day. The task uses timeouts and intervals from global airflow variables. |
| transfixMdShippingPointLoadTask | PROD\_transfix\_jobs.json  shipping\_point.hql | Calculates shipping point master data. |
| sensorEccTvko | dep\_load\_process\_plc.sql | Waits until dd\_osi\_na\_ecc.Tvko is loaded for a particular day. The task uses timeouts and intervals from global airflow variables. |
| sensorEccCompanyCode | dep\_load\_process\_plc.sql | Waits until dp\_osi\_na\_ecc.Company\_code\_dim is loaded for a particular day. The task uses timeouts and intervals from global airflow variables. |
| transfixMdSalesOrgLoadTask | PROD\_transfix\_jobs.json  sales\_org.hql | Calculates sales organization master data. |
| sensorEccVendor | dep\_load\_process\_plc.sql | Waits until dp\_osi\_na\_ecc.Vendor\_dim is loaded for a particular day. The task uses timeouts and intervals from global airflow variables. |
| sensorEccLfb1 | dep\_load\_process\_plc.sql | Waits until dd\_osi\_na\_ecc.Lfb1 is loaded for a particular day. The task uses timeouts and intervals from global airflow variables. |
| transfixMdVendorLoadTask | PROD\_transfix\_jobs.json  vendor.hql | Calculates vendor master data. |
| sensorEccTwlad | dep\_load\_process\_plc.sql | Waits until dd\_osi\_na\_ecc.Twlad is loaded for a particular day. The task uses timeouts and intervals from global airflow variables. |
| transfixMdStorageLocationLoadTask | PROD\_transfix\_jobs.json  storage\_location.hql | Calculates storage location master data. |
| sensorEccCustomer | dep\_load\_process\_plc.sql | Waits until dp\_osi\_na\_ecc.Customer\_dim is loaded for a particular day. The task uses timeouts and intervals from global airflow variables. |
| sensorEccAdr | dep\_load\_process\_plc.sql | Waits until dd\_osi\_na\_ecc.Adr6 is loaded for a particular day. The task uses timeouts and intervals from global airflow variables. |
| transfixMdCustomerLoadTask | PROD\_transfix\_jobs.json  customer.hql | Calculates customer master data. |
| sensorRdsCustHier | dep\_load\_process\_plc.sql | Waits until rds.cust\_hier\_dim is loaded for a particular day. The task uses timeouts and intervals from global airflow variables. |
| transfixMdCustomerHier656LoadTask | PROD\_transfix\_jobs.json  customer\_hier656.hql | Calculates customer hierarchy 656 master data. |
| refreshTableTask | refresh\_tables\_metadata.py | Refreshes Rgvr Star Table |
| End | N/A | End action |

### Transfix variables

Transfix DAG uses dedicated airflow variable named “transfix\_config”. The config contains JSON with the following parameters:

|  |  |  |
| --- | --- | --- |
| **Parameter Name** | **Value** | **Description** |
| cluster\_id | 0621-085219-grows735 | Databricks cluster ID dedicated for Transfix processing |
| db\_name | ap\_transfix\_tv\_na | Transfix database name |
| job\_config\_name | transfix\_prod\_config | Transfix python configuration file |
| job\_spark\_config\_name | transfix\_spark\_config | Spark configuration file |
| hive\_job\_config\_name | tranfix\_reload\_config | Transfix hive configuration file |
| hive\_md\_job\_config\_name | tranfix\_md\_reload\_config | Transfix hive configuration file |
| archive\_dir | dbfs:/mnt/dppsdatahubx45bab07e /etl/transfix/prod/archive | DBFS path to archive directory |
| dbfs\_code\_path | dbfs:/mnt/dppsdatahubx45bab07e/etl/transfix/prod/code | DBFS path to Transfix code |
| dbfs\_input\_dir\_path | dbfs:/mnt/dppsdatahubx45bab07e/lightrefined/transfix/prod/input | DBFS path to Transfix input/output directory (f.eg. for excels) |
| env\_name | PROD | Environment type |
| kv\_name | kv-psdatahub-p201 | Key vault name |
| kv\_secret\_app\_user\_pass | transfix-intranet-pass | Key vault secret name with password |
| kv\_secret\_databricks\_token | transfix-dbr-token | Key vault secret name with token |

Transfix Config uses also global variable named “global\_db\_config”. The config contains JSON with the following parameters

|  |  |  |
| --- | --- | --- |
| **Parameter Name** | **Value** | **Description** |
| rds | rds | RDS database name |
| direct\_shpmt\_cfr | dp\_direct\_shpmt\_cfr | Direct Shipments databricks database name |
| masterdata\_g11 | dp\_masterdata\_g11 | G11 masterdata databricks database name |
| Dp\_inventory | Dp\_inventory | Global BW databricks database name |
| osi\_na\_ecc | dp\_osi\_na\_ecc | ECC transactional SAP N6P databricks database name |
| osi\_la\_ecc | dp\_osi\_la\_ecc | ECC transactional SAP L6P databricks database name |
| osi\_eu\_ecc | dp\_osi\_eu\_ecc | ECC transactional SAP F6P databricks database name |
| osi\_ap\_ecc | dp\_osi\_ap\_ecc | ECC transactional SAP A6P databricks database name |
| osi\_ap\_anp\_ecc | dp\_osi\_ap\_anp\_ecc | ECC transactional SAP ANP databricks database name |
| osi\_ecc\_template | dp\_osi\_{}\_ecc | Template for ECC transactional SAP databricks database name |
| trans\_vsblt\_bw | dp\_trans\_vsblt\_bw | Transvisibility databricks database name |
| transfix\_tv\_na | ap\_transfix\_tv\_na | Transfix databricks database name |
| load\_process\_control | load\_process\_control | Metadata with loaded ECC transactional SAP tables databricks database name |

### Refresh tables on databricks user workspace variables

The DAGs use also global variable named “refresh\_tables\_on\_user\_dbr\_workspace\_config”. The config contains JSON with the following parameters:

|  |  |  |
| --- | --- | --- |
| **Parameter Name** | **Value** | **Description** |
| cluster\_name\_mask | PS\_TAC\_cluster | Databricks cluster names mask on which we would like to refresh tables  (e.g. ‘PS\_\w\*\_cluster’) |
| db\_name | ap\_transfix\_tv\_na | Databricks database name |
| kv\_name | ps-kv-prod-all | Azure Key Vault Name |
| kv\_secret\_dbr\_token | osi-dbr-user-token | Azure Key Vault Secret Name Databricks token to the user workspace |
| app\_user\_name | aptransfix.im | Application user name, owner of the token |
| dbr\_api\_host | eastus2.azuredatabricks.net | Databricks API host |

## Running SQLs with hive params on spark (Databricks)

First part of Transfix project was built to work on hive. For that reason, we built a python tool with a set of functions enabling replacing hive parameters, splitting and running the HQL/SQL statements on spark. The code of the tool you can find on GitLab:

<https://dev.azure.com/dh-platforms-devops/app-ps-datahub/_git/Tools-Azuredatabricks?path=%2FCode%2FDatabricks%2FPython%2Frun_spark_sql.py>

The functions are used by dedicated Transfix python wrapper:

* tranfix\_reload\_config.py – configuration file for main processing (TAC and TFS tables), current parameters are presented in the table below

|  |  |  |
| --- | --- | --- |
| **Parameter Name** | **Value** | **Description** |
| HIVE\_PARAMETERS | {  'database': 'ap\_transfix\_tv\_na',  'dbTransVsbltBw': 'dp\_trans\_vsblt\_bw',  'dbMasterDataG11': 'du\_masterdata\_g11',  'dbPrdMasterDataG11': 'dp\_masterdata\_g11',  'dbOsiNa': 'dp\_osi\_na\_ecc',  'dbOsiNaUat': 'du\_osi\_na\_ecc',  'dbRds': 'rds',  'dbDirectShpmtCfr': 'dp\_direct\_shpmt\_cfr'  } | List of hive parameters to replace in HQL scripts |
| HIVE\_SCRIPT\_DIR\_PATH | dbfs:/mnt/dppsdatahubx45bab07e /etl/transfix/code | DBFS path to hive scripts |
| ARCHIVE\_DIR\_PATH | dbfs:/mnt/dppsdatahubx45bab07e /etl/transfix/prod/archive/hive | DBFS path where the loaded excel are archived. If empty, then we do not archive them. |

* tranfix\_md\_reload\_config.py – configuration file for master data processing, current parameters are presented in the table below

|  |  |  |
| --- | --- | --- |
| **Parameter Name** | **Value** | **Description** |
| HIVE\_PARAMETERS | {  'database': 'ap\_transfix\_tv\_na',  'dbTransVsbltBw': 'dp\_trans\_vsblt\_bw',  'dbMasterDataG11': 'du\_masterdata\_g11',  'dbPrdMasterDataG11': 'dp\_masterdata\_g11',  'dbOsiNa': 'dp\_osi\_na\_ecc',  'dbOsiNaUat': 'du\_osi\_na\_ecc',  'dbRds': 'rds',  'dbDirectShpmtCfr': 'dp\_direct\_shpmt\_cfr'  } | List of hive parameters to replace in HQL scripts |
| HIVE\_SCRIPT\_DIR\_PATH | dbfs:/mnt/dppsdatahubx45bab07e /etl/transfix/code | DBFS path to hive scripts |
| ARCHIVE\_DIR\_PATH | "dbfs:/mnt/ dppsdatahubx45bab07e /lightrefined/transfix/prod/archive" | DBFS path where the loaded excel are archived. If empty, then we do not archive them. |

* run\_transfix\_sql.py – Transfix wrapper for running HQLs with parameters on spark. It requires the configuration file (one of above).

## Export data to excel

Excel is created with python pandas library. The code can be found on GitLab:

<https://dev.azure.com/dh-platforms-devops/app-ps-datahub/_git/Tools-Azuredatabricks?path=%2FCode%2FDatabricks%2FPython%2Fexport_table_2_excel.py>

Excels that are sent successfully are later archived into archive directory.

The tool input arguments:

|  |  |  |
| --- | --- | --- |
| **Argument** | **Required** | **Description** |
| -c | Y | Config module |
| -cd | Y | Config module FS dir path |
| -d | Y | Excel directory path |
| -w | Y | Excel sheet name |
| -rd |  | Report date (yyyymmdd) |
| -o |  | Output excel file name |
| -t | -s | Y | Source table for excel (db\_name.table\_name) | SQL to generate excel |

## Uploading excel file to SP

Transfix excel files (only FAP reports) are uploaded to SP location:

[https://pgone.sharepoint.com/sites/PSCNA/Lib3/Physical Distribution/Daily Operations/FAP Team/KPI's/3-RFT/RFT Raw Data/](https://pgone.sharepoint.com/sites/PSCNA/Lib3/Forms/AllItems.aspx?id=%2Fsites%2FPSCNA%2FLib3%2FPhysical%20Distribution%2FDaily%20Operations%2FFAP%20Team%2FKPI%27s%2F3-RFT%2FRFT%20Raw%20Data)

The files are created in Transfix input directory on ADLS:

dbfs:/mnt/dppsdatahubx45bab07e/etl/transfix/prod/input and sent to SP using python code:

<https://dev.azure.com/dh-platforms-devops/app-ps-datahub/_git/Tools-Azuredatabricks?path=%2FCode%2FDatabricks%2FPython%2Fupload_file_to_sp.py>

The tool input arguments:

|  |  |  |
| --- | --- | --- |
| **Argument** | **Required** | **Description** |
| -c | Y | Config module |
| -cd | Y | Config module FS dir path |
| -i | Y | Input file path |
| -spa | Y | Sharepoint server address (default=https://pgone.sharepoint.com/) |
| -sps |  | Sharepoint site path (default=sites/PSBigData) |
| -spd | Y | Sharepoint directory, eg. 'Hydra Uploads/Files ' |
| -spu | Y | Sharepoint site user, eg [user@pg.com](mailto:user@pg.com) |
| -dss | Y | Key Vault name |
| -dsk | Y | Key Vault secret name |
| -a |  | Archive directory |
| -d |  | Delete file after upload |

## Data archiving

Only excels are archived.

## Email notification

|  |  |  |
| --- | --- | --- |
| **Status/exception** | **Action** | **Target emails** |
| Process Success email (do not include transferring file from BigData to DOMO SFTP server) | Success, send an email | BD Support |
| Process Error email | Error, send an email and stop the process | BD Support |
| Status email | Beginning and end of workflow | BD Support |
| Status email | FAP report generated | BD Support |

# 

## Passwords/tokens/secrets

All of the needed passwords/tokens/secrets needed in the processing are kept as secrets in Azure KeyVault service named “ps-kv-prod-all” and in Airflow connections. The Azure KeyVault secrets are used also in Databricks (called Azure Key Vault-backed scopes) with the same name as KeyVault. Please remember that you have to manage Azure KeyVault and Databricks secrets ACL separately.

Azure KeyVault secrets have an expiration date set if they can expire. In Trans Visibility we use the following secrets:

* transfix-intranet-pass – application user intranet password (needed for SharePoint connection)

The following Airflow connections are used in Trans Visibility project:

* databricks\_transfix – Databricks connection, used to data processing on Databricks
* databricks\_transfix\_jdbc – Databricks JDBC connection, used for dependency checks in SQL sensors

## Refreshing tables on user databricks workspace

Final tables existing on user databricks workspace are refreshed through Databricks API & notebook. Full code can be found on Gitlab:

<https://dev.azure.com/dh-platforms-devops/app-ps-datahub/_git/Tools-Azuredatabricks?path=%2FCode%2FDatabricks%2FPython%2Frefresh_tables_metadata>

# Application interfaces

For now data from transfix application is used by only one application:

|  |  |  |
| --- | --- | --- |
| **Source application DB** | **Business application** | **Description** |
| ap\_transfix\_tv\_na | Tableau | BI application which connects to data in two modes:   * Live SQL (runs ad-hoc queries)   Extract (connects once and pulls data offline to later use) |

In the future there can be more applications like f.eg. KNIME.

# Operational procedures

The operational procedures are listed below in table:

|  |  |  |
| --- | --- | --- |
| **Situation** | **Action needed** | **What to do (easiest way)** |
| There was an error while DAG was running | Investigate logs and find reason. | 1. Open pipeline that failed 2. Check for the DBR link   https://eastus2.azuredatabricks.net/?o=4705345359376196#job/2860/run/1):     1. Copy and paste link in browser and login to Databricks. Check detailed logs:    * Spark UI    * Logs      1. Restart task if needed. |
| Resume DAG | Restart task | 1. Open pipeline that failed 2. Click rerun from failed activity: |
| No space on ADLS/BLOB | Cleanup space | Investigate if there are files in input directory that can be removed. If yes, remove files.  If needed, contact with administrators to increase space. |
| Bad source file | Contact with BW Team | Contact with BW Team to regenerate file. |
| Recreate tables | Run initial DDLs | All initial DDLs are on Azure DevOps. |
| JDBC error | Rerun task | To connect and execute query via JDBC cluster has to be running. If the cluster is not running try to start cluster and rerun failed task. |
| Token / password expired | Update value in key vault or connection | If token or password changed, update new value in:   * connection details (Airflow -> Admin -> Connections) * key vault secret |

# Backup

All files (workflow xml, hql, sh, py, etc.) are available on

Azure DevOps : <https://dev.azure.com/dh-platforms-devops/app-ps-datahub/_git/Application-Transfix-NA>

ADF : <https://dev.azure.com/dh-platforms-devops/da-ps-analytics/_git/Application-Transfix-NA_V2>

DBR : https://dev.azure.com/dh-platforms-devops/da-ps-analytics/\_git/Application-Transfix-NA-DBR-V2

Data in tables created in ap\_transfix\_tv\_na are fully refreshed every day so they don’t need additional backup.

# Contact list

|  |  |  |
| --- | --- | --- |
| **Role** | **Name** | **Contact email** |
| **Product Supply Analyst / Business Owner** | Scott Cassell | [cassell.se@pg.com](mailto:cassell.se@pg.com) |
| **Project Manager** | Matthew Minson | [minson.mw@pg.com](mailto:minson.mw@pg.com) |
| **IT Project Manager** | Dalila Vasquez | vasquez.d.6@pg.com |
| **Big Data Service Manager** | Matthew Minson  Belghazi, Hassan | [minson.mw@pg.com](mailto:minson.mw@pg.com)  belghazi.h@pg.com |
| **Developer** | Yuvaraj Munirathinam  HCL support | [Yuvaraj.munirathinam@hcl.com](mailto:Yuvaraj.munirathinam@hcl.com)  HCL\_PS\_BigData\_Supp@hcl.com |
| **Business users** | Matthew Minson  Ken Hamall  Brian Kosina  Aaron Meyerrenke [VFR]  Lauren Thomas [OTD, CSOT, IOT]  Tomas Navarro Fernandez [OTD]  Thomas Zobrosky [IOT]  Melissa Bowman [LOT]  Bill Bittner [RGVR] | [minson.mw@pg.com](mailto:minson.mw@pg.com)  [hamall.km@pg.com](mailto:hamall.km@pg.com)  [kosina.bm@pg.com](mailto:kosina.bm@pg.com)  [meyerrenke.a@pg.com](mailto:meyerrenke.a@pg.com)  [thomas.lc@pg.com](mailto:thomas.lc@pg.com)  [navarrofernandez.ta@pg.com](mailto:navarrofernandez.ta@pg.com)  [zobrosky.tj@pg.com](mailto:zobrosky.tj@pg.com)  [bowman.mc@pg.com](mailto:bowman.mc@pg.com)  [bittner.wb@pg.com](mailto:bittner.wb@pg.com) |
| **TFTS Setup Contact** | EAI TFTS Support Team | [eai-tfts@dxc.com](mailto:eai-tfts@dxc.com) |
| **JDA (Blue yonder) (source data)** | JDA Shipment Locations extract | HostingSC@blueyonder.com |
| **SAP BW**  **(source data)** | Jain Kshitiz (Architect)  BW Team:  BW support | [Kshitiz\_Jain01@infosys.com](mailto:Kshitiz_Jain01@infosys.com)  [RBW\_IMFR@infosys.com](mailto:RBW_IMFR@infosys.com)  [RBW\_Monitoring@infosys.com](mailto:RBW_Monitoring@infosys.com)  [RBW\_TV@infosys.com](mailto:RBW_TV@infosys.com) |
| **Tableau**  **(end user BI application)** | Balint Horvath (Tableau consultant)  TABLEAU Server support | Balint\_Horvath1@epam.com  [biselfserve.im@pg.com](mailto:biselfserve.im@pg.com) |

# Hardware Platform

Application is running on Databricks Cluster:

<https://eastus2.azuredatabricks.net/?o=570894300217#/setting/clusters/0621-085219-grows735/configuration>

Cluster configuration and component versions:

|  |  |
| --- | --- |
| **Property** | **Value** |
| Cluster name | PROD\_Transfix |
| URL | [Cluster Details - Databricks (azuredatabricks.net)](https://eastus2.azuredatabricks.net/?o=6168179505272179#setting/clusters/1216-052143-l9g4zl84/configuration) |

Azure subscription information:

|  |  |
| --- | --- |
| **Property** | **Value** |
| Subscription | PG-External-01 |
| Resource group | AZ-RG-ProductSupply-PROD-ETL-01 |
| Location | East US 2 |

Azure service information:

|  |  |  |  |
| --- | --- | --- | --- |
| **Service name** | **Comment** | **Link/IP/Name** | **Version** |
| DataBricks on Azure | Portal | https://eastus2.azuredatabricks.net/?o=6168179505272179# | n/a |
| Azure KeyVault | Safeguard cryptographic keys and other secrets | ps-kv-prod-all | n/a |
| Shared Scripts Azure Data Lake Storage (ADLS) | Storage  Mounted to workspace as:  /mnt/ dppsdatahubx45bab07e /etl/shared | ADLS: dppsdatahubx45bab07e  Container: shared | Gen2 |
| Application ETL Azure Data Lake Storage (ADLS) | Storage  Mounted to workspace as:  /mnt/ dppsdatahubx45bab07e /etl/transfix | ADLS: dppsdatahubx45bab07e  Container: transfix | Gen2 |
| Application Final Azure Data Lake Storage (ADLS) | Storage  Mounted to workspace as:  /mnt/ dppsdatahubx45bab07e /lightrefined/transfix | ADLS: dppsdatahubx45bab07e  Container: transfix | Gen2 |
| ADF | Orchestrator, scheduler | adf-productsupply-p203 | n/a |

Nonstandard DataBricks cluster python libraries:

|  |  |
| --- | --- |
| **Library name** | **Description** |
| xlrd | xlrd is a library for reading data and formatting information from Excel files, whether they are .xls or .xlsx files |
| databricks-api | This package provides a simplified interface for the Databricks REST API |
| mssql\_jdbc\_7\_0\_0\_jre8.jar | JDBC Driver for MSSQL |
| openpyxl | A library to read/write Excel 2010 xlsx/xlsm/xltx/xltm files |

Software GitLab link: <https://dev.azure.com/dh-platforms-devops/app-ps-datahub/_git/Application-Transfix-NA>

Azure python tools - python scripts developed during our projects can & used in many ours projects. They were moved to a separate GitLab repository: <https://dev.azure.com/dh-platforms-devops/app-ps-datahub/_git/Tools-Azuredatabricks>

# Application user

For now data from transfix application is used by only one application:

|  |  |
| --- | --- |
| **User name** | **Environment** |
| aptransfix.im | Production |

# Service Level Agreement (SLA) Parameters

# Access and incident management

Application related issues should be reported in SNOW. You should raise a [PS Datahub - Incident Escalation ticket](https://pgglobalenterprise.service-now.com/ITSM_Portal/catalog-item.do?sysparm_document_key=sc_cat_item_producer,375976c0db008450766be6650596196b).

Enter the following information:

* Are you requesting for someone else?
  + Yes / No
* What is the impact of this incident?
  + Choose one of the options
* Mandatory - must be populated before SubmitWhat is the urgency of this incident?
  + Choose one of the options
* When did issue first appear?
  + Choose the date when the issue occured
* Mandatory - must be populated before SubmitProblem Area
  + Application
* What kind of issue are you experiencing?
  + Choose one of the options
* Which Region is impacted?
  + Choose one of the options
* Provide short description
  + [TRANSFIX] Short description of the issue
* Provide detailed description
  + Detailed description of the issue

# Attachments

## Hive codes



## Python codes

