

PMO report - 3 - Technical - Data Pipeline

- [High-Level Design architecture \(HLD\)](#)
- [Low-Level Design architecture \(LLD\)](#)
- [Architecture Data Flow](#)
 - [DataPrep Flow](#)
- [Conceptual Data Model](#)
- [Steps descriptions](#)
 - [DataSource 1 - Accolade DB](#)
 - [Description](#)
 - [Tools](#)
 - [Access rights](#)
 - [Source](#)
 - [Location](#)
 - [Format](#)
 - [Destination](#)
 - [Location](#)
 - [Format](#)
 - [Sizing](#)
 - [Assessment](#)
 - [Scheduling](#)
 - [Timing](#)
 - [Criticality](#)
 - [Logging](#)
 - [Draft \(Notes\) for DataOps Understanding:](#)

High-Level Design architecture (HLD)

[Link.](#)

Low-Level Design architecture (LLD)

[Link.](#)

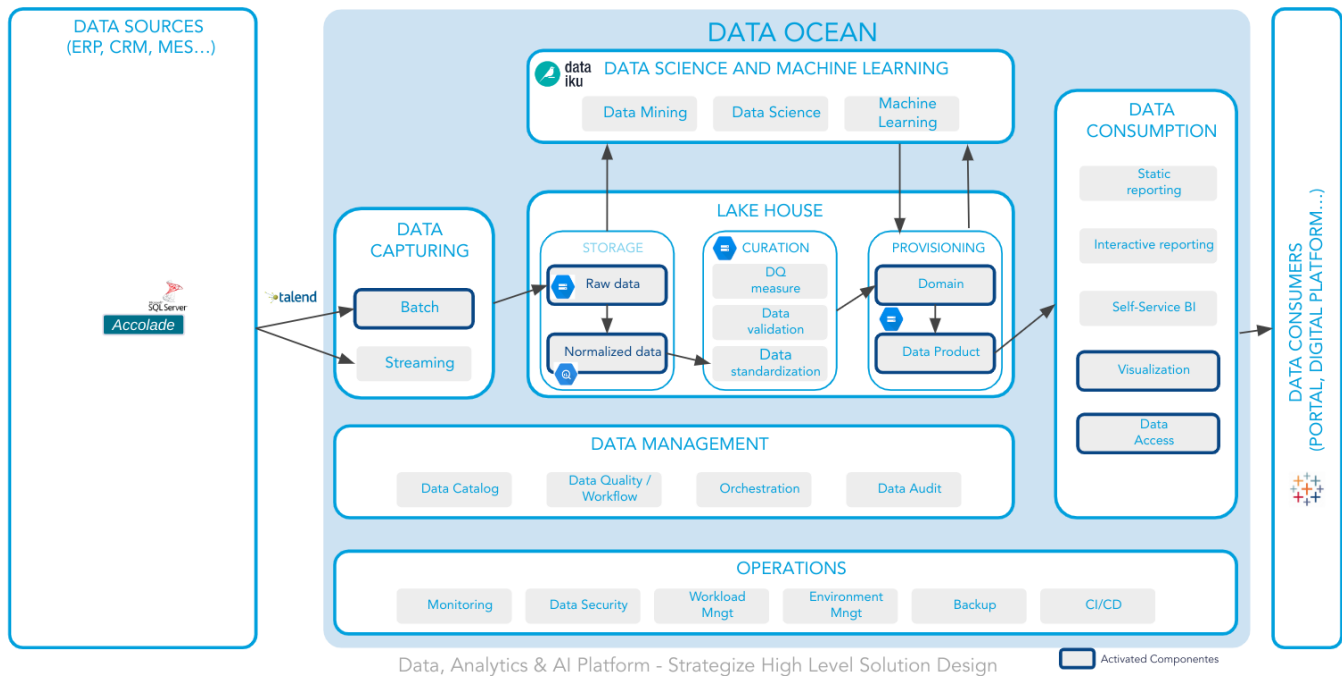
Architecture Data Flow

Here is a suggested template for Data Model + Data Mapping :

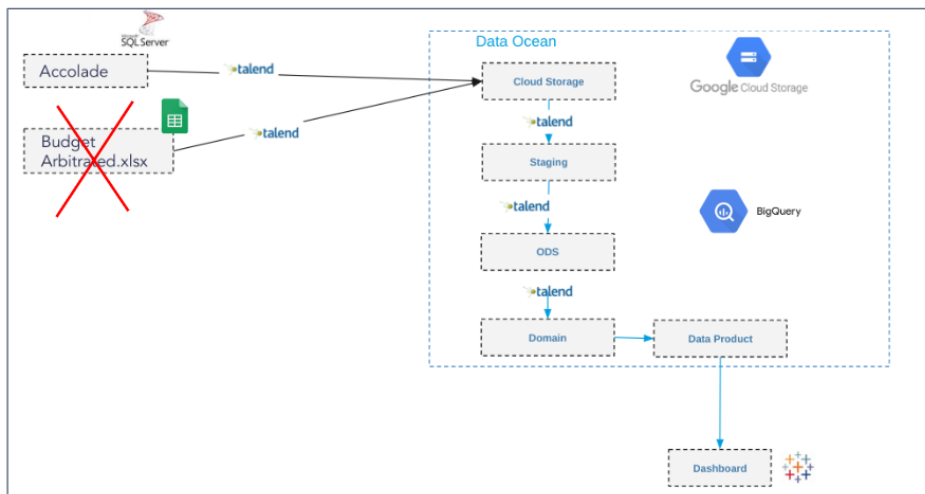
[DA&DA - Domain Mapping DT](#)

DataPrep Flow

Schema showing the different STEPS of the application flow - with the data involved at each step

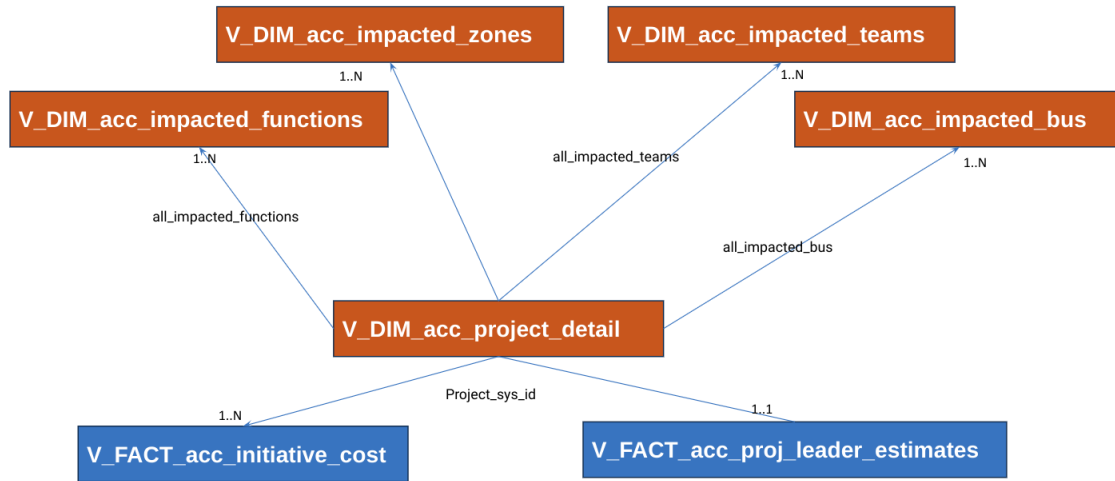


According to Data Ocean Blueprint



Accolade --> GCP Data Ocean DT (prj-data-dm-dt-[env]) --> GCP Product PMO (prj-data-pmo-dash-[env]) --> Output Tableau

Conceptual Data Model



Steps descriptions

Source table in Accolade	Flow name (Source to ODS)	Update Table in GCP (STG/ODS)	Flow name (ODS to DM)	Update Table in GCP (DM)	Flow name (Join tables)	Update Table in GCP (DM)
crv_initiative_v3 (Project Cost Budget&Cost)	F010_CRV_INITIATIVE_V3	STG.STG_ACC_0000_0000_F001_F_H_crv_initiative_v3 ODS.ODS_ACC_0000_F001_F_H_crv_initiative_v3	F211_DIM_acc_initiative F212_DIM_acc_linked_prog F213_FACT_acc_initiative_cost	DIM_acc_initiative DIM_acc_linked_prog FACT_acc_initiative_cost	F301_DIM_CRV_ZGetProjectDetails	DIM_acc_project_detail
CRV_ZGetProjectDetails (Main Project Data from standard fields)	F002_CRV_ZGetProjectDetails	STG.STG_ACC_0000_0000_F001_F_H_crv_zgetprojectdetails ODS.ODS_ACC_0000_F001_F_H_crv_zgetprojectdetails	F202_DIM_CRV_ZGetProjectDetails	DIM_acc_proj_detail		
Execute procedure MetricsForGroupID 62 (Status Indicators)	F003_SGM_PROJECT_METRICS_15	STG.STG_ACC_0000_0000_F001_I_H_sgm_projectmetrics_15 ODS.ODS_ACC_0000_F001_I_H_sgm_projectmetrics_15	F203_DIM_acc_status_indicators	DIM_acc_status_indicators		
Execute procedure MetricsForGroupID 62 Project Leader Estimates	F004_SGM_PROJECT_METRICS_62	STG.STG_ACC_0000_0000_F001_F_H_rsp_qv_getprojectmetricsforgroupid_62 ODS.ODS_ACC_0000_F001_F_H_rsp_qv_getprojectmetricsforgroupid_62	F209_DIM_acc_impacted_teams	DIM_acc_impacted_teams		
Execute procedure MetricsForGroupID 94 (Risk Analysis)	F005_SGM_PROJECT_METRICS_94	STG.STG_ACC_0000_0000_F001_F_H_rsp_qv_getprojectmetricsforgroupid_94 ODS.ODS_ACC_0000_F001_F_H_rsp_qv_getprojectmetricsforgroupid_94	F204_FACT_acc_proj_leader_estimates F205_DIM_acc_mv_risk_analysis	FACT_acc_proj_leader_estimates DIM_acc_mv_risk_analysis		
Execute procedure MetricsForGroupID 103 (Identification & Classification)	F006_SGM_PROJECT_METRICS_103	STG.STG_ACC_0000_0000_F001_F_H_rsp_qv_getprojectmetricsforgroupid_103 ODS.ODS_ACC_0000_F001_F_H_rsp_qv_getprojectmetricsforgroupid_103	F206_DIM_acc_ident_classification F207_DIM_acc_impacted_bus F208_DIM_acc_impacted_zones F210_DIM_acc_impacted_functions	DIM_acc_ident_classification DIM_acc_impacted_bus DIM_acc_impacted_zones DIM_acc_impacted_functions		

Note: Table on project prj-data-dm-dt-[env]

DataSource 1 - Accolade DB

Description

Accolade is the application where users (mainly project managers) share data (status, forecast as examples) related to a project.

Tools

Talend to collect the data from the source and store on GCP/Google Big Query.

Access rights

Only authorized users can connect into Accolade application.

Only DA&AI DataEng's and Data Architects can access data on GCP/Google Big Query.

Source

Location

Accolade DB via Talend generic account created for this purpose.

Test = cew1twegodb01.nonprod.aws.cloud.solvay.com

Prod = cew1pwegodb01.prod.aws.cloud.solvay.com.

Format

MSSQL direct table and run procedure.

Destination

Location

Extracted data will be stored on GCP/Cloud storage and Google Big Query.

Format

The format of the data saved in the databank

Sizing

Expected data volume for :

- full process from source to staging (as of 6 Dec 2023)

job_name	meta_start_date	meta_run_id	meta_source_system	meta_step	meta_status	meta_num_lines	meta_error
F010_CRV_INITIATIVE_V3	2023-12-06 02:02:45.507921 U...	4e312eaf-ec8b-40fa-8057-bbc...	ACC_DT_0000_0000_F001_20 231206020243_0000_F_H_crv _initiative_v3.csv	Bucket to Staging	OK	53480	
F002_CRV_ZGetProjectDetails	2023-12-06 02:02:32.066018 U...	fb07a89d-83b5-4ef0-b6ab-7b8...	ACC_DT_0000_0000_F001_20 231206020229_0000_F_H_crv _zgetprojectdetails.csv	Bucket to Staging	OK	7678	
F003_SGM_PROJECT_METRICS_15	2023-12-06 02:02:29.872505 U...	81cc06a6-2083-4bec-8e05-f7c...	ACC_DT_0000_0000_F001_20 231206020226_0000_F_H_rsp _qv_getprojectmetricsforgrou pid_15.csv	Bucket to Staging	OK	9728	
F004_SGM_PROJECT_METRICS_62	2023-12-06 02:02:24.334172 U...	da7e5d4d-6f61-42fb-96a9-b66...	ACC_DT_0000_0000_F001_20 231206020221_0000_F_H_rsp _qv_getprojectmetricsforgrou pid_62.csv	Bucket to Staging	OK	7646	
F005_SGM_PROJECT_METRICS_94	2023-12-06 02:02:23.207017 U...	5dda7a0e-8485-43fb-81be-ca9...	ACC_DT_0000_0000_F001_20 231206020220_0000_F_H_rsp _qv_getprojectmetricsforgrou pid_94.csv	Bucket to Staging	OK	7646	
F006_SGM_PROJECT_METRICS_103	2023-12-06 02:02:21.220125 U...	131cd10f-17d2-45a2-ac4e-3c2...	ACC_DT_0000_0000_F001_20 231206020218_0000_F_H_rsp _qv_getprojectmetricsforgrou pid_103.csv	Bucket to Staging	OK	11945	

- incremental process from ODS to DM (as of 6 Dec 2023)

job_name	meta_start_date	meta_run_id	meta_source_system	meta_step	meta_status	meta_num_lines	meta_error
F202_DIM_CRV_ZGetProjectDe...	2023-12-06 02:03:56.966059 U...	cdbaa754-53d2-4b36-b386-2a...	ACC	ODS to DM	OK	7688	
F203_DIM_acc_status_indicato...	2023-12-06 02:03:45.690826 U...	29b79613-4c92-4d8c-a089-f82...	ACC	ODS to DM	OK	7653	
F204_FACT_acc_proj_leader_e...	2023-12-06 02:04:44.182762 U...	4e90f004-fc6c-4d2c-bdfa-1ff9c...	ACC	ODS to DM	OK	7655	
F205_DIM_risk_analysis	2023-12-06 02:04:42.055351 U...	342ac317-bedf-478c-bd16-fce...	ACC	ODS to DM	OK	7646	
F206_DIM_acc_ident_classifica...	2023-12-06 02:04:44.855062 U...	5a313026-8c58-4d8b-80c9-b0...	ACC	ODS to DM	OK	7661	
F207_DIM_acc_impacted_bus	2023-12-06 02:03:44.792958 U...	59e78d4f-6922-4721-b32a-809...	ACC	ODS to DM	OK	954	
F208_DIM_acc_impacted_zones	2023-12-06 02:03:47.269723 U...	cfc3c081-595e-4154-8f96-3c1...	ACC	ODS to DM	OK	80	
F209_DIM_acc_impacted_teams	2023-12-06 02:03:39.124825 U...	85ba8413-3338-4943-b1ba-53f...	ACC	ODS to DM	OK	1730	
F210_DIM_acc_impacted_funct...	2023-12-06 02:03:48.307720 U...	66200c3f-1a73-441e-bdb3-958...	ACC	ODS to DM	OK	638	
F211_DIM_acc_initiative	2023-12-06 02:03:41.037510 U...	d5cd95cf-ffa1-46e5-bf73-2b3f...	ACC	ODS to DM	OK	2213	
F212_DIM_acc_linked_prog	2023-12-06 02:03:43.158901 U...	5f209da0-139a-41cd-8df1-2edf...	ACC	ODS to DM	OK	1116	
F213_FACT_acc_initiative_cost	2023-12-06 02:04:43.071811 U...	eb0543c1-8fed-454c-8634-b48...	ACC	ODS to DM	OK	53293	

Assessment

Check the log tables in GCP on table log_tables and run_jobs to check that there is no error loading from source to staging/ods

Check the surrogate key must be unique in the data mart layer

Scheduling

Is there an automatic schedule ? Yes

At what frequency ? to collect data 4 times/day (every 6hour).

What is the trigger ? TMC

Timing

The average time expected for :

- 4 times/day (working days: from Monday to Friday) : to be scheduled. 2:00, 8:00, 14:00, 20:00 CET (monitor by DataOps only at 8:00 and 14:00)
- full process (source to ODS)
- incremental process (ODS to DM)

Criticality

~~High~~ / Medium / Low

Logging

Table table log_tables, run_jobs, log_files, and reject_files in `prj-data-dm-dt-[environment].STG.[table]`

Draft (Notes) for DataOps Understanding:

prj-data-dm-dt :

Source GCP Buckets --> prj-data-dm-dt.STG. --> prj-data-dm-dt.ODS --> prj-data-dm-dt.DM . --> prj-data-dm-dt.ds_pmo_dashboard. |
||
V

Reporting will be done from PRJ-DATA-PMO-DASH-DEV project on DPL dataset (prj-data-pmo-dash-dev.DPL).

Tableau --> prj-data-pmo-dash-dev.DPL --> Prj-data-pmo-dash-dev.DataOcean --> prj-data-dm-dt.ds_pmo_dashboard.