Data Architecture Practices

As Data architect when you start on the project DA needs to participate in below activities :

1. Functional Analysis Review
2. "Discussion with Dataviz if not covered under Functional Analysis"
3. Source System Analysis
    a. **Description** : In this activity DA needs to understand the source that has been requested by business, document the source tables etc
    b. **Tools** : If the source is BW then use Xtract to asses the SAP BW, if source is anyother system for ex SAP, salesforce, request access of it for analysis of tables.
    c. **Deliverable** : Documenting the sources in mapping document
4. Create Xtract configuration for SAPBW Description :
    a. Description : If the source for the project is the SAP BW then use use Xtract to configure and retrive data.
    b. **Step by Step guide to connect to Xtract and get the URL for Talend**
    c. Once you get the URL , insert it the mapping document
5. Tables / Views conventions
    a. Tables conventions:
        i. STG
            1. Naming:
                a. Tables
                - The name must begin with the prefix STG_
                - Staging can only contain lowercase letters for the file name description, numeric characters, underscores (_).
                - Spaces are not allowed.
                - The name cannot be a reserved word in Google BigQuery such as WHERE or VIEW.
                - The name cannot be the same as another Google BigQuery object that has the same type.
                - When you create a table in BigQuery, the table name must be unique per dataset.
                - Only use approved acronyms which are known in the organization.
                - The table name cannot exceed the 80 characters.
                - Include the system name
                - Include the domain name
                - Include the Site name
                - Include the System Reference
                - Include the File Code
                - Include a sequential number
                - Include an extraction type
                - Include the frequency

Name convention: [001]_[002]_[003]_[004]_[005]_[006]_[007]_[008]

Where:

#	Category	Description
001	Area	Start with the prefix **STG** to identify 3 characters for this category
002	System Name	System the data come from. Ex: HLX ; ELN ; MES 3 Characters for this category
003	Site Name	Site that data come from Ex: 0000 4 Characters for this category
004	System Reference	System the data come from. Ex: HLX ; ELN ; MES Ex:0000 4 Characters for this category

005	File Code	For each file code we will have a dedicated staging table Ex: F001; F002 4 Characters for this category
006	Extraction Type	If it's a full or incremental extraction (F= Full; I = Incremental) 1 character for this category
007	Frequency	If it's Monthly, weekly, daily, quarterly etc… Ex: M = monthly, W = weekly 1 character for this category
008	File name	Identify the content of the table. Ex: Cannot exceed 63 characters

Example : STG_HLX_0000_0000_F001_F_W_stellar_escalation_follow_up

b. Columns:

- Can only contain lowercase letters.
- Spaces are not allowed.
- names must start with a letter and finish with a letter or number.
- The name cannot contain special characters. (only "_" is allowed)
- The name cannot be a reserved word such as WHERE or VIEW.
- A primary key column should usually have only 1 column serving as a primary key. It would be best to simply name this column "id".
- For dates, it's good to describe what the date represents. Names like start_date and end_date are pretty descriptive. If you want, you can describe them even more precisely, using names like call_start_date and call_end_date.

c. Views

- The name must begin with the prefix "vw_"
- Views follow many of the same rules that apply to naming tables convention.

2. Keys
3. Mandatory Columns

a.

#	Field Name	Description	Type	Example
001	meta_run_id	Id of the run, this information come from STG_runs	Integer	E.g. 12408
002	meta_md5_hsh	file check sum	String	E.g. 69095ac6258ec1c9fd151cec979ace71
003	meta_file_name	Bucket File name	String	E.g. 1291619-3d223530ac3c11ecae3500000ad02795-Emulsion Polymerization.json
004	meta_file_path	Path of the bucket	String	E.g. gs://cs-ew1-prj-dashb-rational-dev-staging/STG_SEA.csv
005	meta_execution_id	System Execution id Prefix of the tool used to populate the field + execution id that match tool log.	String	E.g. Talend_6Oyhb6
006	meta_bucket_load_date	Date the bucket file was loaded (UTC)	timestamp	E.g. 2022-12-12 17:00:21 UTC
007	meta_business_date	This date comes from the business (UTC), it's when the data was refreshed from the business side.	timestamp	E.g. 2022-12-12 17:00:21 UTC
008	meta_stg_insert_date	When the data was inserted in the staging	timestamp	E.g. 2022-12-12 18:00:42 UTC
009	meta_source_system	Identification of the source where the data comes from.	String	E.g SAP; BW; ORA

ii. ODS

1. Naming
    a. Tables
        - adopt convention names from the source, it will facilitates debugging or finding data
        - The name must begin with the prefix ODS_
        - ODS can only contain lowercase letters for the file name description, numeric characters, underscores (_).
        - Spaces are not allowed.
        - The name cannot be a reserved word in Google BigQuery such as WHERE or VIEW.
        - The name cannot be the same as another Google BigQuery object that has the same type.
        - When you create a table in BigQuery, the table name must be unique per dataset.
        - Only use approved acronyms which are known in the organization.
        - The table name cannot exceed the 80 characters.
        - Include the domain name
        - Include the Site name
        - Include the File Code
        - Include a sequential number
        - Include an extraction type
        - Include the frequency

        Name convention: [001]_[002]_[003]_[004]_[005]_[006]_[007]

        Where:

#	Category	Description
001	Area	Start with the prefix **ODS** to identify 3 characters for this category
002	System Name	System the data come from. Ex: HLX ; ELN ; MES 3 Characters for this category
003	Site Name	Site that data come from Ex: 0000 4 Characters for this category
004	File Code	For each file code we will have a dedicated staging table Ex: F001; F002 4 Characters for this category
005	Extraction Type	If it's a full or incremental extraction (F= Full; I = Incremental) 1 character for this category
006	Frequency	If it's Monthly, weekly, daily, quarterly etc… Ex: M = monthly, W = weekly 1 character for this category
007	File name	Identify the content of the table. Ex: Cannot exceed 63 characters

        Example of the staging naming convention:

        ODS_0000_F001_F_W_stellar_escalation_follow_up

    b. Columns
        - adopt convention names from the source, it will facilitates debugging or finding data
        - Can only contain lowercase letters.
        - Spaces are not allowed.
        - names must start with a letter and finish with a letter or number.
        - The name cannot contain special characters. (only "_" is allowed)
        - The name cannot be a reserved word such as WHERE or VIEW.
        - A primary key column should usually have only 1 column serving as a primary key. It would be best to simply name this column "id".

- For dates, it's good to describe what the date represents. Names like start_date and end_date are pretty descriptive. If you want, you can describe them even more precisely, using names like call_start_date and call_end_date.

2. Keys
3. Mandatory Columns

i. #	Field Name	Description	Type	Example
001	meta_run_id	Id of the run, this information come from STG_runs	Integer	E.g. 12408
002	meta_md5_hsh	file checksum	String	E.g. 69095ac6258ec1c9fd151cec979ace71
003	meta_table_name	Name of the table where the data comes from	String	E.g.
004	meta_business_date	This date comes from the business (UTC), it's when the data was refreshed from the business side.	timestamp	E.g. 2022-12-12 17:00:21 UTC
005	meta_execution_id	System Execution id Prefix of the tool used to populate the field + execution id that match tool log.	String	E.g. Talend_6Oyhb6
006	meta_ods_insert_date	When the data was inserted in the staging	timestamp	E.g. 2022-12-12 18:00:42 UTC
007	meta_source_system	Identification of the source where the data comes from.	String	E.g SAP; BW; ORA

- DM

a. Naming:
   i. Table naming should be all lowercase;
   ii. Table naming should always be in the plural (i.e. customers, not customer, invoices, not invoice);
   iii. Table columns naming should be all lowercase;
   iv. Table name should not exceed the 80 characters;
   v. Table name should respect the following convention: type_entityname, where:
      1. type:
         a. dim
         b. fact
         c. agg
      2. entity name - the business entity name or meaning (i.e. customers or invoices)
   vi. Table name cannot contain any character other than the following:
      1. Letters: a to z
      2. Numbers;
      3. The special character "_";
   vii. Table constraints (PK, FK, UK, indexes) naming should be all lowercase;
   viii. Foreign keys naming should respect the following rule: table name || "_fk" || sequence number. Example: fact_invoices_fk1;
   ix. Unique keys naming should respect the following rule: table name || "_uk" || sequence number. Example: fact_invoices_uk1;
   x. Indexes naming should respect the following rule: table name || "_idx" || sequence number. Example: fact_invoices_idx1;
   xi. Facts/events tables should have the prefix "fact_";
   xii. Dimensions/LOV tables should have the prefix "dim_";
   xiii. Aggregation tables should have the prefix "agg_";
   xiv. Relationship tables created to avoid many to many relationships should have the sufix "_rel";
   xv. Temporary tables should have the sufix "_tmp";
   xvi. Entity names should be respected, as far as possible, across domains

b. Keys:
   i. A table must always have a Primary Key defined in GCP;
   ii. Foreign keys should be created in GCP whenever applicable;
   iii. Indexes should be created only when usable by GCP (in GCP indexes are automatically disabled for small tables);
   iv. Primary Key should always be an MD5 hash based in the following rules (exceptions may be applicable if needed):
      1. Dimensions - MD5(source system key || date), where date is one of the following:
         a. Ingestion date of the master file feeding the DM table
         b. Insert date of the master ODS table feeding the DM table
         c. GCP current_date
      2. Facts - MD5(business functional key)

c. Mandatory Columns
   i. Dimensions
      1. SCD Type 1 and Type 2

#	Field Name	Description	Type	Example
001	business_id	Primary key of source table. Could be a single attribute or combination of the attributes	String	E.g 0001; "ABC"

002	<Table_name>_key	Generated SGK. As stated before, must be unique. The procedure should be implemented as a Hash function based on the Business Id, concatenate with the Extraction Date to reinforce the key uniqueness	String	Eg: gbu_key
003	meta_run_id	Id of the run, this information come from STG_runs	Integer	E.g. 12408
004	meta_execution_id	System Execution id Prefix of the tool used to populate the field + execution id that match tool log.	String	
005	inserted_date	Datetime when the record was originally inserted into this table (NOW) and will never change.	timestamp	E.g. 2022-12-12 18:00:42 UTC
006	updated_date	Datetime and set to the same date as Inserted Date when the record is inserted for the first time. Change every time the record is updated.	timestamp	E.g. 2022-12-12 18:00:42 UTC

2. SCD Type 2 Only

#	Field Name	Description	Type	Example
001	start_date	Date when this version of the Business Id is considered to be effective. Implement the rules previously defined	Timestamp	Populate with 1900-01-01 00:00:00 for the first version
002	end_date	Date when that record is no longer active Implement the rules previously defined	Timestamp	Populate with 9999-12-31 00:00:00 for all active records
003	current_flag	Identify the active record, the latest version of the record Synonym for End_Date = 9999-12-31T00:00:00	Boolean	The Active record will have the value set to True (all others should be set to False)

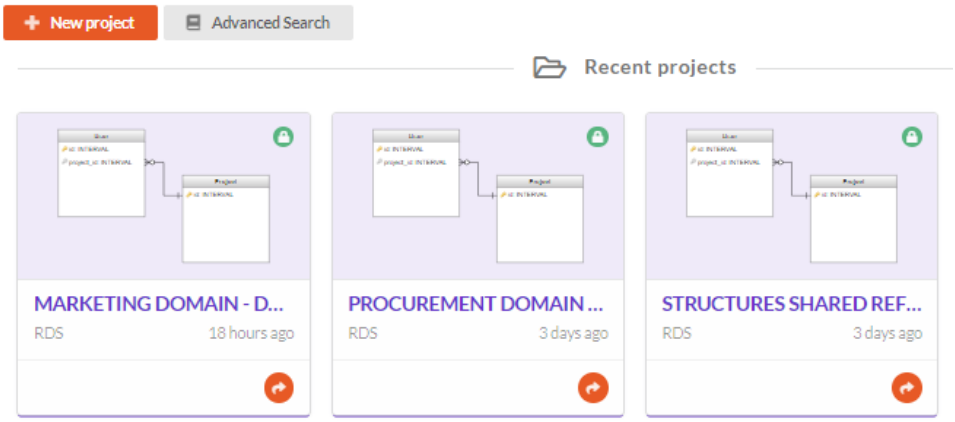
ii. Facts

1.

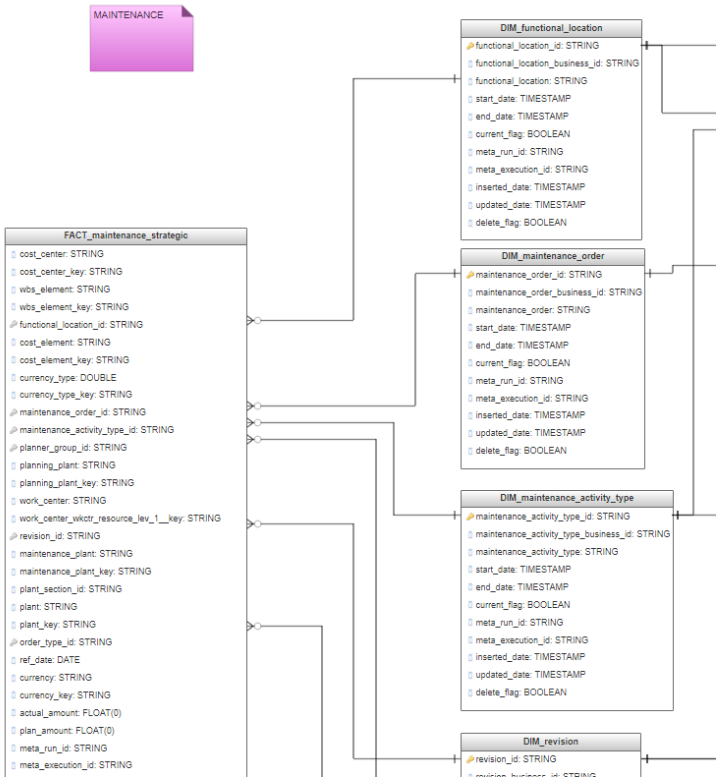
#	Field Name	Description	Type	Example
001	<Table_name>_key	Generated SGK. As stated before, must be unique. The procedure should be implemented as a Hash function based on the Business Id, concatenate with the Extraction Date to reinforce the key uniqueness	String	Eg: fact_invoices_key
002	meta_run_id	Id of the run, this information come from STG_runs	Integer	Compulsory E.g. 12408
003	meta_execution_id	System Execution id Prefix of the tool used to populate the field + execution id that match tool log.	String	Compulsory
004	inserted_date	Should be a datetime at least at the second level. Date relative to the moment when the record was inserted in this table and will never change	timestamp	E.g. CURRENT_TIMESTAMP()

- Data Modeling
    1. Data architect are in charge to design the Data Model, preferably applying Dimensional Modeling Concepts and Techniques described by Ralph Kimball.
    2. After aligning the data objects allocation to a specific business domain, the Data Architect designs the Data Model, using genmymodel tool. https://app.genmymodel.com/

3. Domain-oriented data models are stored in a dedicated Genmymodel project named: "*DOMAIN name* DOMAIN - DATA MODEL"



4. Inside a domain-oriented project, the data models are organized in a group hierarchy
    a. Level 1: domain-oriented, this group is named with the GCP project id (such as: prj-data-dm-marketing-dev)
    b. Level 2: three sub-groups: STG, ODS, DM. The Data Architect focuses on the DM, the level were the data model is stored
5. Inside the DM group, the Data Model is described by the Data Architect by FACT (transactional data) and Dimensions tables (static or almost static data)



6. The Data Architect completes the list of fields of each table and add the relationships between tables.
7. Once the design is complete, the Data Architect submits the work to peer-review with the Data Architecture team.

- DPL - Modeling
- Specify Maps to ODS
- Data Transformation Rules to Data model
    1. Data transformation will cover required transformation rules to transform the data from source to target
    2. Below are the high level details on the document
        a. Source and Target table name
        b. Type of the target table (Master table (SCD1 or SCD2) or Transaction table)
        c. Transformation rules to be applied
        d. Mandatory and primary key attributes
        e. Attributes to be used part of SCD2 to track the changes
    3. Example template:
- DTR to DPL
- Updating Data Catalog Document

- Quality inspection assurance
- Data Flows
- Manager Security roles on GCP

1.