Data Architect - Naming Convention

Control Table

João Fonseca	0.1	Initial Version	26 Dec 2022
Fernando Girante	1.1	Revision	22 May 2023

Introduction

Purpose of the document

This document describes the naming for all the objects to be used in a project or in domain for GCP and Talend technologies.This will help to normalize all the names and facilitate the way of work.

Intended Audience

This document is intended for the Data Architectures and the Data Engineering team.
It will be used as reference for any project or domain for the developments of the models.

Bucket Naming Convention

Description

A Bucket is anything you want to store in the GCS in order to perform any operations on them.
Naming your buckets is a great way to make it easy for you to find your data in the Cloud Storage service.

Buckets

Name conventions for buckets are:

- Bucket names can only contain lowercase letters, numeric characters, underscores (_), and dots (.). Spaces are not allowed. Names containing dots require verification .
- Bucket names must start and end with a number or letter.
- Bucket names must contain 3-63 characters. Names containing dots can contain up to 222 characters, but each dot-separated component can be no longer than 63 characters.
- Bucket names cannot be represented as an IP address in dotted-decimal notation (for example, 192.168.5.4).
- Bucket names cannot begin with the "goog" prefix.
- Bucket names cannot contain "google" or close misspellings, such as "g00gle".
- Include the system name
- Include the domain name
- Include the Site name
- Include the System Reference
- Include the File Code
- Include Business date
- Include a sequential number
- Include an extraction type
- Include the frequency

To generate the name of a bucket use this work document for all the conventioning names, filling the 11 first fields in the google sheet:
{+} https://drive.google.com/file/d/1jxbuJApqk20nCmbocq49_cvGHtVg3r85hX_ILrR4pfM/view+ blocked URL

0 01	System Name	System the data come from. Ex: HLX; ELN; MES 3 Characters for this category
0 02	Domain	Domain that data comes from. Ex: HR; RI; etc… If we don't know the Domain, use the generic one: IT 2 Characters for this category
0 03	Site Name	Site is the same system existing in a different location. This information comes from the source. Ex: 0000 4 Characters for this category
0 04	System Reference	Reference of the system. Ex:0000 4 Characters for this category
0 05	File Code	For each system and project we can receive more than 1 file and we need to identify each of them. This will help to understand how many files that project receives for a specific system, if one of them is missing the process will abort. Ex: F001; F002 4 Characters for this category
0 06	Business Date	When data was extracted from the business side. Format: YYYYMMDDHHMISS Ex:20220812000000 14 characters for this category
0 07	Sequential Number	If we have more than one file with exactly the same name. This can happen when we extract big volumes of data and we need to split the file in 2 or 3 Ex: 0000; 0001 4 characters for this category
0 08	Extraction Type	If it's a full or incremental extraction (F= Full; I = Incremental) 1 character for this category
0 09	Frequency	If it's Monthly, weekly, daily, quarterly etc… Ex: M = monthly, W = weekly 1 character for this category
0 10	File name	Identify the content of the table. Need to be the same name used by the table source. Ex: Cannot exceed 60 characters

Example of the bucket naming convention:
HLX_IT_0000_0000_F001_20220812000000_0000_F_W_Stellar_escalation_follow_up

Fields

For the buckets the naming of the fields usually comes from the google sheet source, to guarantee the naming convention, this needs to be discussed with the source to send the files with the fields names exactly as we need, in case it is not possible we need to normalize the fields in GCP side.
Name conventions for fields are:

- Adopt convention names from the source, it will facilitates debugging or finding data
- Spaces are not allowed.
- names must start with a letter and finish with a letter or number. (can start with "_")
- The name cannot contain special characters. (only "_" character 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". (The definition of the PK at this level needs to be defined by the source).
- 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. (The definition of the PK at this level needs to be defined by the source).

Staging Naming Convention

Description

A staging area, or landing zone, is an intermediate storage area used for data processing during the extract, transform and load (ETL) process. The data staging area sits between the data source(s) and the data target(s), which are often data warehouses , data marts , or other data repositories.
We will use this staging for merging bucket files (in case we have more than one file for the same table).
Datasets name always needs to be in CAPITAL letter ,

Dataset

The name of the Dataset is STG

Tables

Name conventions for tables are:

- The name must begin with the prefix STG_
- Staging must be 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

To generate the name of the staging table use this work document for all the conventioning names, filling the 11 first fields in the google sheet:
{+} https://drive.google.com/file/d/1jxbuJApqk20nCmbocq49_cvGHtVg3r85hX_ILrR4pfM/view+ blocked URL

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; MESEx: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 of the staging naming convention:
STG_HLX_0000_0000_F001_F_W_stellar_escalation_follow_up

Fields

Name conventions for fields are:

- Adopt convention names from the source, it will facilitates debugging or finding data
- Spaces are not allowed.
- names must start with a letter and finish with a letter or number (can start with "_")
- 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.

Views

At this level the Views are only used for security or to reduce the number of field purposes, we will not deliver data from the staging area to any report or extraction layer.
Name conventions for views are:

- The name must begin with the prefix V_AUT (authorization view)
- Views follow many of the same rules that apply to naming tables convention.
- Some views are simply tabular representations of one or more tables with a filter applied or because of security procedures (users given permissions on views instead of the underlying table(s) in some cases). Some views are used to generate report data with more specific values in the WHERE clause. Naming your views should be different depending on the type or purpose of the view. For simple views that just join one or more tables with no selection criteria, combine the names of the tables joined. For example, joining the "Customer" and "StateAndProvince" table to create a view of Customers and their respective geographical data should be given a name like "VW_customer_state_and_province". Views created expressly for a report should have an additional prefix of Report applied to them, e.g. VW_sec_sales_for2008.

Routines

A routine usually runs a set of actions and returns a dataset.
Name conventions for routine are:

- The name must begin with the prefix RT_
- If the routine is using only one table, I'll name it RT_<table_name>_<action_name>. E.g., RT_customer_insert inserts a new row in the table customer; RT_customer_delete deletes a row.
- If the routine uses more than 1 table, I would use a descriptive name for the procedure. E.g., if we want all customers with 5 or more calls, I would call this procedure similar to this – RT_customer_with_5_or_more_calls
- If the routine returns a scalar value, or performs an operation like validation, you should use the verb and noun combination. For example, "RT_validate_login".

Note: All the ETL will be performed by the Talend tool.

WDL Naming Convention

Description

Working data layer, it's an area for temporary tables or for tables we need to work on to deliver a final.

Dataset

The name of the Dataset is WDL

Tables

Name conventions for tables are:

- The name must begin with the prefix TMP_ (temporary table) or WRK_ (working table)
- 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.

Fields

Name conventions for the fields are:

- Spaces are not allowed.
- 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.

Views

Routines

ODS Naming Convention

Description

An operational data store (ODS) is used for operational reporting and as a source of data for the enterprise data warehouse (EDW). It is a complementary element to an EDW in a decision support environment, and is used for operational reporting, controls, and decision making, as opposed to the EDW, which is used for tactical and strategic decision support.

Dataset

The name of the Dataset is ODS

Tables

Name conventions for tables are:

- The name must begin with the prefix ODS_
- ODS must be 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

To generate the name of the staging table use this work document for all the conventioning names, filling the 11 first fields in the google sheet:
{+} https://drive.google.com/file/d/1jxbuJApqk20nCmbocq49_cvGHtVg3r85hX_ILrR4pfM/view+ blocked URL

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	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
Note: ODS dont need systemref because if have several system on the same site and domain in staging, they will merge on the same table in the ODS and the table on the ODS will have a column with the systemref to be identified

Fields

Name conventions for the fields are:

- adopt convention names from the source, it will facilitates debugging or finding data
- Spaces are not allowed.
- names must start with a letter and finish with a letter or number. (can start with "_")
- 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.

Views

The views at this level can be used for:

- ETL purposes
- Security purpose
- Generate report data with more specific values
- etc...

Name conventions for views are:

- The name must begin with the prefix V_
- Views follow many of the same rules that apply to naming tables convention.
- Some views are simply tabular representations of one or more tables with a filter applied or because of security procedures (users given permissions on views instead of the underlying table(s) in some cases). Some views are used to generate report data with more specific values in the WHERE clause. Naming your views should be different depending on the type or purpose of the view. For simple views that just join one or more tables with no selection criteria, combine the names of the tables joined. For example, joining the "Customer" and "StateAndProvince" table to create a view of Customers and their respective geographical data should be given a name like "VW_customer_state_and_province". Views created expressly for a report should have an additional prefix of Report applied to them, e.g. VW_report_division_000_sales_for2008.

Routines

A routine usually runs a set of actions and returns a dataset.
Name conventions for routine are:

- The name must begin with the prefix RT_
- If the routine is using only one table, I'll name it RT_<table_name>_<action_name>. E.g., RT_customer_insert inserts a new row in the table customer; RT_customer_delete deletes a row.
- If the routine uses more than 1 table, I would use a descriptive name for the procedure. E.g., if we want all customers with 5 or more calls, I would call this procedure similar to this – RT_customer_with_5_or_more_calls
- If the routine returns a scalar value, or performs an operation like validation, you should use the verb and noun combination. For example, "RT_validate_login".

Note: All the ETL will be performed by the Talend tool.

Models Naming Conventions

Description

This area will be used to describe naming conventions models for the Data warehouse and Data Marts.
A data warehouse is a large collection of business data used to help an organization make decisions.
A data mart is a subset of a data warehouse focused on a particular line of business, department, or subject area.

Dataset

The name of the Dataset is DM

Tables

Name conventions for tables are:

- For factual data the name must begin with the prefix FACT_
- For dimensional data the name must begin with the prefix DIM_
- For Bridge data the name must begin with the prefix BDG_
- For aggregation tables the name must begin with the prefix AGG_
- For snapshot tables the name must begin with the prefix SNP_
- Tables which are used for audit should be start with ADT_<table_name>
- Models can only contain lowercase letters for the 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.

Fields

Product Naming Conventions

Description

This area will be used to describe naming conventions for the product naming conventions.

Dataset

The name of the Dataset is DS_namingoftheproduct

ex: DS_PricingDataLake, DS_PMO_dashboard.

Views

V_AUT
V_ETL
V_REP for reporting purpose

Routines

A routine usually runs a set of actions and returns a dataset.
Name conventions for routine are:

- The name must begin with the prefix RT_
- If the routine is using only one table, I'll name it RT_<table_name>_<action_name>. E.g., RT_customer_insert inserts a new row in the table customer; RT_customer_delete deletes a row.
- If the routine uses more than 1 table, I would use a descriptive name for the procedure. E.g., if we want all customers with 5 or more calls, I would call this procedure similar to this – RT_customer_with_5_or_more_calls

- If the routine returns a scalar value, or performs an operation like validation, you should use the verb and noun combination. For example, "RT_validate_login".

Note: All the ETL will be performed by the Talend tool.