GCP Naming Standards

Overview

Organization Folder
- Project
  - Storage Bucket
    - Folders
      - Files
        - Field Names
  - Datasets
    - Tables
      - Field Names
- Project
  - Storage Bucket
    - Folders
      - Files
        - Field Names
  - Datasets
    - Tables
      - Field Names

*Labels applied here
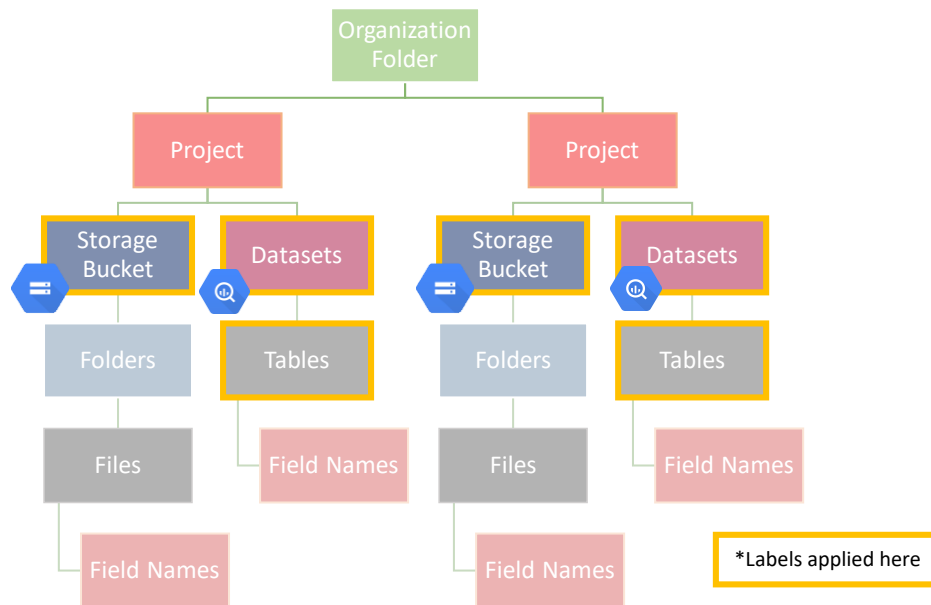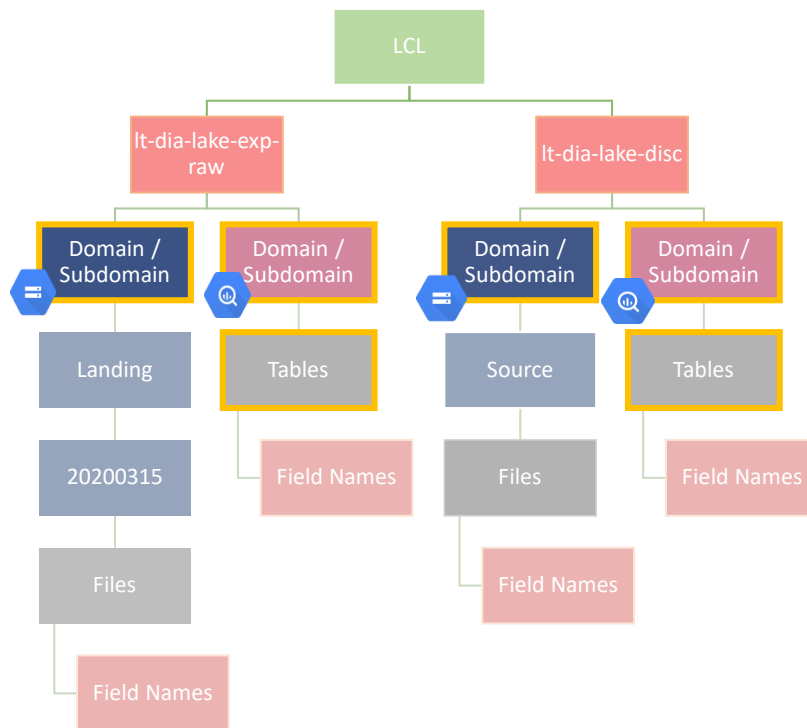
Project

organization-initiative-environment-zone

Project names are used to denote the environment. Please use the abbreviations below when denoting the environment and zone. Project names are not used to denote subject area, business domain, or data type. Please see the examples for exploration below:

Org	Initiative	Environment	Zone	AKA
lt	dia-lake	exp	consume	lt-dia-lake-exp-consume
lt	dia-lake	exp	techdata	lt-dia-lake-exp-techdata
lt	dia-lake	exp	raw	lt-dia-lake-exp-raw

Datalake-environment – current naming

Storage Buckets

sb-domain-subdomain

Bucket names must conform both the LCL Internal Standards and to Google Cloud Naming Standards. Within the datalake, Storage Buckets are at the subdomain level or the physical implementation layer (refer to appendix for subdomain structure). Subdomains that share the same name as the domain (e.g. Product) should only be listed once. Names must be all in lowercase and only contain alpha-numeric characters. The only special characters allowed in the names are hyphens separating each component. Names should be abbreviated to the values provided below to keep names short and consistent. Names should be separated by hyphens in the following format *domain-subdomain*. Please see the examples below.

Project	Resource	Domain	Subdomain	Storage Bucket Name
lt-dia-lake-exp-consume	sb	Product	Product	lt-dia-lake-exp-consume-sb-prdt
lt-dia-lake-exp-techdata	sb	Financial	Profitability	lt-dia-lake-exp-techdata-sb-fin-pnl
lt-dia-lake-exp-raw	sb	Entity	Vendor	sb-entity-vendor
lt-dia-lake-prd-consume	sb	Logistics	Warehouse Management	lt-dia-lake-prd-consume-sb-lgx-wm

*Please note that this is not an exhaustive list of all possible storage bucket names

Folders

Datasets

ds-domain-subdomain

Datasets are a collection of tables within BigQuery and are at the same lateral level as Storage Buckets within the hierarchical structure of GCP (refer to Diagram 1 in Overview). The names should be identical to their corollary Storage Bucket, except for the resource name. Datasets should be named with ds prefix (ds = dataset), while Storage Bucket are named with sb. The fully qualified name when querying a dataset and table is {projectID}.{dataset_name}.{table_name}

Names must be all in lowercase and only contain alpha-numeric characters. The only special characters allowed in the names are hyphens separating each component. Names should be abbreviated keep names short and consistent. See Appendix for abbreviation list. Names should be separated by hyphens in the following format *ds-domain-subdomain*. Please see the examples below:

Resource	Domain	Subdomain	Dataset Name
ds	Product	Product	ds-prdt
ds	Financial	Profitability	ds-fin-pnl
ds	Entity	Vendor	ds-entity-vendor
ds	Logistics	Warehouse Management	ds-lgx-wm

Tables

SOURCESYSTEM_TABLENAME
SubjectArea_Qualifiers_Granularity_Timeframe_View

Table naming convention is dependent on the environment and zone. Tables landing in raw with no transfomations should retain their name from the source, with the prefix of Source System. Source system is the system of origin of the data, also known as the source application.

Qualifiers, granularity, and timeframe should be added when it is essential to the understanding of the table. The qualifier portion denotes the most critical or distinctive elements of table that distinguish it from other tables. Some examples of common qualifiers are SSRS, Trans, Outbound, or RX. Granularity is the lowest level of detail that the table includes. Examples include site, article, vendor, or patient ID. For example, in the case of multiple sales tables at different levels of granularity (MCH1, MCH2, Article), the granularity is essential to the understanding of the table. Timeframe is the specified time period in which the data occurs. If a table only contains data for a specific timeframe, i.e. it is always prior day data, that timeframe must be called out in the name. Tables meant for consumption should have two guiding principles: Understandability and Usability. They should be readable at a glance, they should be "queryable", meaning they should use Enterprise standard acronyms and abbreviations wherever possible. Timeframes, systems, and acronyms that would normally be in uppercase should remain in uppercase. Any other term should be in Camel Case/Title Case. This means that the First letter of each important word is capitalized. Phrases should be kept together, separated by an underscore from different terms. I.e. Product_ArticleNumber; Vendor_Address; Site_PostalCode Table names should use abbreviations wherever possible. Please refer to Appendix 1 for the list of abbreviations. Please see the examples below.

Zone	Source	Subject Area	Qualifier(s)	Granularity	Time frame	View	Table Name
RAW	SAP	n/a	MARA	n/a	PW	W	SAP_MARA_W
RAW	CPMS	n/a	n/a	SKUS	n/a	n/a	CPMS_SKUS
CONSUME	n/a	Wages	Contract	Colleague	PY	n/a	Wages_Contract_Colleague_PY
CONSUME	n/a	Shipments	Outbound	MCH0	n/a	V	Shipments_Outbound_MCH0_V
CONSUME	n/a	Vendor	Address	Postal Code			Vendor_Address_PostalCode
CONSUME	n/a	Product	Seasonal	ArticleNumber			Product_Seasonal_ArticleNum

*Please note that this is not an exhaustive list of all table names
*Domain and subdomain names TBD

Labels

The purpose of labels is to classify the data types for easy querying and metadata management. Labels must be key:value pairs and be written in all lower-case characters. See the examples below:

Key	Possible Values	Usage	Level
Source Application	SAP EE HW	Monitor and organize data coming from a source system, reduce duplication of data	Dataset Tables

Sensitivity	Unrestricted Restricted PII PHI PCI	Group and see a holistic view of sensitive data, restrict access	Table Dataset View Storage Bucket
Data Elements	Customer Site Article	Group data according to data elements	Table Dataset Storage Bucket
Status	Active Inactive Decommissioned	Track the status of data and ensure access is removed from decommissioned data. Ensure retention policies are observed.	Table Dataset Storage Bucket

*Please note that this is not an exhaustive list of all possible labels.

TBD

Files

Naming convention TBD

- file naming convention should dependent on the file type but this will be teased out during implementation

  - i.e Avro or Parquet files will be chunked into small files so should fall some sort of pattern based on partition

Field Names

See standardization guidelines from BI Reporting Standards and MBID for existing naming convention

To ensure that metric names are constructed in a standardized manner, metric names should follow the "Parent-Child" formatting and the following order should be applied:

An example of a metric name adhering to this naming convention is **Sales Grth - Comp PY (%).**

Metric	Metric Type	Qualifier	Timeframe	Symbol
Sales	Grth	Comp	PY	(%)

Note 1: There is a dash between the Parent and Child. E.g. Sales Grth – Comp PY (%)
Note 2: In cases where special characters are used, such as a $ or %, there will be a space before the special character and it will be placed in parenthesis.
Note 3: Grth refers to Growth.

See the *MBID* Information Cube in *MSTR > BI Reports and Self-Service > Governance Folder* for a list of metrics and their descriptions.

See Appendix for common abbreviations and acronym list.

Cloud Functions
TBD

Data Flow
ETL Naming, coding standards