Kosh Data Model-Data Dictionary

Table of Contents

[1. Document Overview 4](#_Toc98952651)

[2. Context 4](#_Toc98952652)

[3. Intent 4](#_Toc98952653)

[4. Kosh data model entities 5](#_Toc98952654)

[4.1. Dataplace 5](#_Toc98952655)

[4.2. Dataplace component 7](#_Toc98952656)

[4.3. Dataplace component mapping 7](#_Toc98952657)

[4.4. Dataplace file system component 9](#_Toc98952658)

[4.5. Dataplace table metadata 10](#_Toc98952659)

[4.6. Table schema drift 12](#_Toc98952660)

[5. Data Type Tables 13](#_Toc98952661)

[5.1. Datatype mapping 13](#_Toc98952662)

[5.2. User-defined trans for unsupported datatype 14](#_Toc98952663)

[5.3. Inconsistent datatype options 15](#_Toc98952664)

[6. Data movement Tables: 16](#_Toc98952665)

[6.1. Data movement details 16](#_Toc98952666)

[6.2. Data movement engine lookup 17](#_Toc98952667)

[6.3. Data movement errors 18](#_Toc98952668)

[6.4. Data movement physical 19](#_Toc98952669)

[7. Column Profiling Tables 21](#_Toc98952670)

[7.1. Crawl xref 21](#_Toc98952671)

[7.2. Crawler errors 22](#_Toc98952672)

[7.3. Crawling status 22](#_Toc98952673)

[7.4. Crawling task status 23](#_Toc98952674)

[8. Engine Tables 24](#_Toc98952675)

[8.1. Engine mapping 24](#_Toc98952676)

[8.2. Engine sub type lookup 24](#_Toc98952677)

[8.3. Workflow engine 25](#_Toc98952678)

[8.4. Workflow engine bots mapping 25](#_Toc98952679)

[8.5. Compute engine 26](#_Toc98952680)

[9. Bot Tables 28](#_Toc98952681)

[9.1. Bot configuration workflow 28](#_Toc98952682)

[9.2. Bot factory 29](#_Toc98952683)

[9.3. Bot instance 30](#_Toc98952684)

[9.4. Bot status 31](#_Toc98952685)

[9.5. Bot type 32](#_Toc98952686)

[9.6. Flow status 33](#_Toc98952687)

[9.7. Flow task status 34](#_Toc98952688)

[9.8. Job schedule details 35](#_Toc98952689)

[10. Messaging protocol 36](#_Toc98952690)

[10.1. Kafka 36](#_Toc98952691)

[10.2. Kafka topics 37](#_Toc98952692)

[11. Schedules, Configuration, and Status Tables: 37](#_Toc98952693)

[11.1. Process id table map 37](#_Toc98952694)

[11.2. spark\_job\_result 38](#_Toc98952695)

[11.3. Status code 38](#_Toc98952696)

[11.4. Template group dependency 39](#_Toc98952697)

[11.5. Template group info 40](#_Toc98952698)

[11.6. Vault details 40](#_Toc98952699)

[11.7. Checkpoint status 42](#_Toc98952700)

[11.8. Credential info 44](#_Toc98952701)

[11.9. Credential type 45](#_Toc98952702)

[12. Advanced options tables 46](#_Toc98952703)

[12.1. Advanced options datatype mappings 46](#_Toc98952704)

[12.2. Advanced options 47](#_Toc98952705)

[12.3. Advanced options table ingestion 48](#_Toc98952706)

[12.4. Partition column info 49](#_Toc98952707)

[12.5. Partitions info 50](#_Toc98952708)

[12.6. Nabu Roles 52](#_Toc98952709)

[12.7. Data Access Roles 52](#_Toc98952710)

[12.8. Role scope lookup 53](#_Toc98952711)

[12.9. UI access type lookup 53](#_Toc98952712)

[12.10. UI policy lookup 54](#_Toc98952713)

[12.11. Roles info 54](#_Toc98952714)

[12.12. Roles UI policy 55](#_Toc98952715)

[12.13. Environment group role 55](#_Toc98952716)

[12.14. Environment group type lookup 56](#_Toc98952717)

# Document Overview

Kosh is currently a repository for technical metadata on sources ingested into the Platform. It also contains other data such as status information on the data flowing into the Platform through the compute engine as well as verification data gathered after those ingestion processes finish.

The desire is to also control all data movement and workflows from the Kosh repository so that information on the complete data life cycle is collected. Kosh will also provide technical metadata on the integrated layer of the platform from data stored in the Kosh repository.

# Context

This document defines the detailed representation of the technical requirements at a component level and captures the dataflow based on specific functions being performed rather than at the source level since the Kosh functions will be used across all required sources and will apply to both the acquisition of sources and the curation of those required sources. The primary goal is to capture the high-level, low-level design and implementation, including the workflows for all components and processes.

# Intent

The platform requires the ability to perform a variety of workflows in the environment, from data crawling, metadata collection to complex data movements workloads. There shall be a repository that holds all the information related to all these tasks. Nabu™ provides various modules in its ecosystem to support all these processes by storing the information in a relational data model named Kosh.

The technical metadata component of Kosh will record data about the actual data such as sources, tables, columns, files, and their counts. This gives search capabilities for source analysis and reference. Profile statistics for both structured and unstructured data sources will be collected and stored. The data can be accessed by other applications for verification and validation (as in a data movement process)

# Kosh data model entities

The contents from the next chapter will start explaining the different entities in the Kosh data model.

## Dataplace

It is a repository that contains the collection of metadata from all the sources which are extracted at the request of the user. Actual data is extracted from sources by using this metadata. This table helps in understanding the orchestration of other components related to it.

Table name: dataplace

|  |  |  |  |
| --- | --- | --- | --- |
| **Column name** | **Datatype** | **Nulls** | **Description** |
| dataplace\_id | INTEGER | NO | an identifier for dataplace referenced from dataplace table |
| dataplace\_name | TEXT | NO | a unique name for data place |
| dataplace\_component\_type\_id | SMALLINT | NO | identifier for component type |
| dataplace\_sub\_component\_id | SMALLINT | NO | An identifier for dataplace sub-component referenced from sub-component table |
| dataplace\_info | JSON | YES | contains additional information regarding dataplace |
| is\_active | BOOLEAN | YES | states if active or not |
| is\_sensitive | BOOLEAN | YES | states if dataplace is sensitive or not to crawler |
| is\_ingestible | BOOLEAN | YES | states if dataplace is ingestible or not to crawler |
| credential\_id | INTEGER | NO | identifier for credential |
| config\_id | INTEGER | YES | identifier for configuration |
| crawl\_frequency | CHARACTER VARYING | YES | refers frequency for crawling metadata |
| contact\_info | JSON | YES | contact information of the owner |
| filter\_rules | JSON | YES | this is used for future reference for filtering sources for custom types like only crawling CSV files from a file share |
| valid\_from\_ts | TIMESTAMP WITHOUT TIME ZONE | YES | the time from which the status record for this table metadata is valid |
| valid\_to\_ts | TIMESTAMP WITHOUT TIME ZONE | YES | the time to which the status record is valid. the current record should have a value of 31-dec-9999 |
| crt\_by | CHARACTER VARYING | YES | refers to the owner who created |
| crt\_ts | TIMESTAMP WITHOUT TIME ZONE | YES | refers to the time of the creation |
| mod\_by | CHARACTER VARYING | YES | displays the owner name who has modified it |
| mod\_ts | TIMESTAMP WITHOUT TIME ZONE | YES | refers the time of modification |

## Dataplace component

Dataplace component table consists of the table information of a component i.e. the table is Relational, Semi-Structured, or Un – Structured. These are mainly static tables where no changes are implemented in the later stages of the development.

Table name: dataplace\_component\_lookup

|  |  |  |  |
| --- | --- | --- | --- |
| **Column name** | **Datatype** | **Nulls** | **Description** |
| dataplace\_component\_type\_id | SMALLINT | NO | identifier for component type |
| dataplace\_component\_type | CHARACTER VARYING | NO | this stores corresponding component types -relational, unstructured, semi-structured |

## Dataplace component mapping

Dataplace component mapping table consists of the mapping between Dataplace and Dataplace sub information. i.e. It defines the dataset as structured and its corresponding relational type mapping.

Table name: dataplace\_component\_mapping\_lookup

|  |  |  |  |
| --- | --- | --- | --- |
| **Column name** | **Datatype** | **Nulls** | **Description** |
| dataplace\_component\_type\_id | SMALLINT | NO | identifier for component type |
| dataplace\_sub\_component\_id | SMALLINT | NO | An identifier for dataplace sub-component referenced from sub-component table |
| metadata\_category | TEXT | YES | refers to metadata\_category, collection, Documentum, file, relational, salesforce, SharePoint |
| additional\_info | JSON | YES | additional information for datastore |

|  |  |  |  |
| --- | --- | --- | --- |
| **Column name** | **Datatype** | **Nulls** | **Description** |
| dataplace\_id | INTEGER | NO | an identifier for dataplace referenced from dataplace table |
| dataplace\_component\_type\_id | SMALLINT | NO | identifier for component type |
| directory\_id | BIGINT | NO | identifier for directory |
| file\_id | BIGINT | NO | identifier for file |
| filename | TEXT | YES | name of the file |
| file\_relative\_path | TEXT | YES | states the relative path for the file |
| bucket\_name | TEXT | YES | reference for the bucket |
| file\_created\_ts | TIMESTAMP WITHOUT TIME ZONE | YES | states the time at which the file was created |
| file\_modified\_ts | TIMESTAMP WITHOUT TIME ZONE | YES | states the time at which the file was modified |
| file\_size | BIGINT | YES | states the size of the file |
| owner\_id | CHARACTER VARYING | YES | identifier referring to the owner |
| metatags | JSON | YES | details related to batch, process, and other details captured in JSON |
| additional\_info | JSON | YES | additional information for datastore |
| valid\_from\_ts | TIMESTAMP WITHOUT TIME ZONE | YES | the time from which the status record for this table metadata is valid |
| valid\_to\_ts | TIMESTAMP WITHOUT TIME ZONE | YES | the time to which the status record is valid. the current record should have a value of 31-dec-9999 |
| crawl\_id | UUID | YES | An identifier for crawling refers to the crawl Xref table |
| prev\_crawl\_id | UUID | YES | refers to the previous crawl\_id of crawl\_xref table |

## Dataplace file system component

This contains directory level information for both Semi-Structured and Un – Structured data.

Table name: dataplace\_file\_system\_component

|  |  |  |  |
| --- | --- | --- | --- |
| **Column name** | **Datatype** | **Nulls** | **Description** |
| dataplace\_id | INTEGER | NO | an identifier for dataplace referenced from dataplace table |
| dataplace\_component\_type\_id | SMALLINT | NO | identifier for component type |
| directory\_id | BIGINT | NO | identifier for directory |
| directory\_name | TEXT | YES | name of directory |
| root\_location\_path | TEXT | YES | / |
| is\_active | BOOLEAN | YES | states if active or not |
| is\_sensitive | BOOLEAN | YES | states if sensitive or not |
| directory\_path | TEXT | YES | refers the path of the directory |
| recursion\_depth | SMALLINT | YES | recursion for subdirectories |
| valid\_from\_ts | TIMESTAMP WITHOUT TIME ZONE | YES | the time from which the status record for this table metadata is valid |
| valid\_to\_ts | TIMESTAMP WITHOUT TIME ZONE | YES | the time to which the status record is valid. the current record should have a value of 31-dec-9999 |

## Dataplace table metadata

This table contains the attributes of the table metadata. After the ingestion is completed, destination table information is populated in this table.

Table name: dataplace\_table\_metadata\_physical

|  |  |  |  |
| --- | --- | --- | --- |
| **Column name** | **Datatype** | **Nulls** | **Description** |
| dataplace\_sub\_component\_id | SMALLINT | NO | unique identifier for dataplace subcomponent |
| dataplace\_sub\_component\_type | CHARACTER VARYING | NO | this stores corresponding component types - PostgreSQL, Oracle, MySQL, s3, SharePoint, Unix |
| additional\_info | JSON | YES | additional information for datastore |
| dataplace\_sub\_component\_label | CHARACTER VARYING | YES | The label defines if the component type is an oracle, SQL server, or cloud which will be used in UI, JDBC connection |
| dataplace\_id | INTEGER | NO | An identifier for dataplace referenced from dataplace table |
| dataplace\_component\_type\_id | SMALLINT | NO | identifier for component type |
| schema\_id | BIGINT | NO | identifier for schema |
| schema\_name | TEXT | NO | name of schema |
| table\_id | BIGINT | NO | unique identifier for the table |
| table\_name | TEXT | NO | name of table |
| table\_owner | TEXT | YES | owner of the table |
| status | CHARACTER VARYING | YES | status of the task started, stopped, failed, interrupted used for UI |
| estimated\_rows | BIGINT | YES | number of rows in the table |
| inserts | BIGINT | YES | displays the number of rows inserted |
| updates | BIGINT | YES | displays the number of rows updated |
| deletes | BIGINT | YES | displays the number of rows deleted |
| create\_table\_ts | TIMESTAMP WITHOUT TIME ZONE | YES | stores the time of table creation |
| modify\_table\_ts | TIMESTAMP WITHOUT TIME ZONE | YES | stores the time of table modification |
| table\_type | TEXT | YES | stores the reference for table type |
| table\_type\_owner | TEXT | YES | stores the name of the owner |
| table\_size | BIGINT | YES | stores the size of the table |
| num\_of\_columns | BIGINT | YES | stores the actual number of columns in a table |
| table\_layer | TEXT | YES | let’s know in which layer is the data |
| location\_path | TEXT | YES | path |
| tablespace\_name | TEXT | YES | stores the name of the tablespace |
| table\_format | TEXT | YES | states the format of the table |
| is\_encrypted | BOOLEAN | YES | states true/false |
| is\_compressed | BOOLEAN | YES | state of partition is compressed or not |
| additional\_info | JSON | YES | additional information for datastore |
| valid\_from\_ts | TIMESTAMP WITHOUT TIME ZONE | YES | the time from which the status record for this table metadata is valid |
| valid\_to\_ts | TIMESTAMP WITHOUT TIME ZONE | YES | the time to which the status record is valid. the current record should have a value of 31-dec-9999 |
| lineage\_uUId | UUID | YES | An identifier for the lineage |
| crawl\_id | UUID | YES | An identifier for crawling refers to the crawl Xref table |
| prev\_crawl\_id | UUID | YES | refers to the previous crawl\_id of crawl\_xref table |

## Table schema drift

While crawling if there is any change in the structure of the source table, the new table will be created in the destination. There will be an entry in the table\_schema\_drift if there is any change in the schema in the source table.

Table name: table\_schema\_drift

|  |  |  |  |
| --- | --- | --- | --- |
| **Column name** | **Datatype** | **Nulls** | **Description** |
| dataplace\_id | INTEGER | NO | An identifier for dataplace referenced from dataplace table |
| dataplace\_component\_type\_id | SMALLINT | NO | identifier for component type |
| schema\_id | BIGINT | NO | identifier for schema |
| schema\_name | TEXT | NO | name of schema |
| table\_id | BIGINT | NO | An identifier for table reference to table metadata |
| table\_name | TEXT | NO | name of table |
| table\_owner | TEXT | YES | owner of the table |
| schema\_change\_flag | BOOLEAN | YES | states if table schema is modified |
| temp\_table\_insert\_flag | BOOLEAN | YES | states if data in the table is updated |
| updated\_date | DATE | YES | states date on which table is updated |
| cru\_by | CHARACTER VARYING | YES | displays the owner name who has created/modified it |
| cru\_ts | TIMESTAMP WITHOUT TIME ZONE | YES | refers the time of creation/modification |

# Data Type Tables

## Datatype mapping

This table is used when ingestion is performed using SDC. This table contains information of source\_type, destination\_type, source\_datatype\_name and destination\_datatype\_name etc. We use this table to apply transformation like changing datatype of the source from int to number etc

Table name: datatype\_mapping

|  |  |  |  |
| --- | --- | --- | --- |
| **Column name** | **Datatype** | **Nulls** | **Description** |
| source\_type | CHARACTER VARYING | YES | type of source |
| intermediate\_type | CHARACTER VARYING | YES | type of intermediate stage |
| destination\_type | CHARACTER VARYING | YES | type of destination |
| source\_datatype\_name | CHARACTER VARYING | YES | datatype |
| intermediate\_datatype\_name | CHARACTER VARYING | YES | name of mapping |
| destination\_datatype\_name | CHARACTER VARYING | YES | datatype |
| source\_cast\_type | TEXT | YES | type of cast |
| intermediate\_cast\_type | TEXT | YES | type of mapping |
| source\_stg\_function | CHARACTER VARYING | YES | The transformation needed for inconsistent types from source to intermediate |
| intermediate\_stg\_function | CHARACTER VARYING | YES | the transformation needed for inconsistent types from intermediate to final |
| datamovement\_engine\_type | CHARACTER VARYING | YES | type of data movement engine |
| additional\_info | JSON | YES | additional information for datastore |
| engine\_id | SMALLINT | YES | identifier for engine |

## User-defined trans for unsupported datatype

This table contains user-defined datatypes which are mapped and are handled by this mapping table

Table name: user\_defined\_trans\_for\_unsupported\_datatype

|  |  |  |  |
| --- | --- | --- | --- |
| **Column name** | **Datatype** | **Nulls** | **Description** |
| data\_movement\_id | INTEGER | NO | An identifier for data movement referenced from the data movement table |
| source\_type | CHARACTER VARYING | NO | type of source |
| intermediate\_type | CHARACTER VARYING | NO | type of intermediate stage |
| destination\_type | CHARACTER VARYING | NO | type of destination |
| engine\_id | SMALLINT | NO | identifier for engine |
| source\_datatype\_name | CHARACTER VARYING | NO | datatype |
| advanced\_options\_sub\_type\_id | SMALLINT | NO | advanced options for any transformation |
| additional\_info | JSON | YES | additional information for user-defined types |
| valid\_from | TIMESTAMP WITHOUT TIME ZONE | YES | the time from which the status record for this table metadata is valid |
| valid\_to\_ts | TIMESTAMP WITHOUT TIME ZONE | YES | the time to which the status record is valid. the current record should have a value of 31-dec-9999 |
| cru\_by | CHARACTER VARYING | YES | displays the owner name who has created/modified it |
| cru\_ts | TIMESTAMP WITHOUT TIME ZONE | YES | refers the time of creation/modification |
|  |  |  |  |
|  |  |  |  |

## Inconsistent datatype options

This table contains inconsistent datatype

Table name: inconsistent\_datatype\_options

|  |  |  |  |
| --- | --- | --- | --- |
| **Column name** | **Datatype** | **Nulls** | **Description** |
| source\_type | CHARACTER VARYING | NO | type of source |
| intermediate\_type | CHARACTER VARYING | NO | type of intermediate stage |
| destination\_type | CHARACTER VARYING | NO | type of destination |
| datamovement\_engine\_type | CHARACTER VARYING | NO | type of data movement engine |
| source\_type\_name | CHARACTER VARYING | NO | type of source |
| inconsistent\_datatype | CHARACTER VARYING | NO | inconsistent datatype |
| description | TEXT | YES | description |
| mapping\_options | JSON | YES | additional options |
| engine\_id | SMALLINT | YES | identifier for engine |

# Data movement Tables:

## Data movement details

This table stores information of source and destination

Table name: data\_movement\_details

|  |  |  |  |
| --- | --- | --- | --- |
| **Column name** | **Datatype** | **Nulls** | **Description** |
| data\_movement\_id | INTEGER | YES | unique identifier of a data movement |
| destination\_dataplace\_id | INTEGER | YES | an identifier for dataplace referenced from dataplace for destination |
| filter\_rule | JSON | YES | the filter value is given by the user |
| filter\_type\_id | INTEGER | YES | an identifier for filter type |
| ingest\_all\_tables\_views | JSON | YES | whether to ingest only all views within the selected schema |
| ingestion\_table\_format | TEXT | YES | whether to ingest only the tables/views/both within the selected schema |
| priority\_order | SMALLINT | YES | states priority |
| source\_schema\_id | BIGINT | YES | identifier for schema |
| destination\_schema\_id | BIGINT | YES | refers to the schema of the destination |
| source\_dataplace\_id | INTEGER | YES | unique id of the dataplace which is selected as a source |
| valid\_from\_ts | TIMESTAMP WITHOUT TIME ZONE | YES | the time from which the status record for this table metadata is valid |
| valid\_to\_ts | TIMESTAMP WITHOUT TIME ZONE | YES | the time to which the status record is valid. the current record should have a value of 31-dec-9999 |
| crt\_by | CHARACTER VARYING | YES | refers to the owner who created |
| crt\_ts | TIMESTAMP WITHOUT TIME ZONE | YES | refers to the time of the creation |

## Data movement engine lookup

This table contains details of engines related to data movement

Table name: data\_movement\_engine\_lookup

|  |  |  |  |
| --- | --- | --- | --- |
| **Column name** | **Datatype** | **Nulls** | **Description** |
| data\_movement\_engine\_id | SMALLINT | NO | identifier of data movement engine |
| data\_movement\_engine\_type | CHARACTER VARYING | NO | type of movement engine |

## Data movement errors

This table stores errors captured during data movement, if there is any error while performing the task, it's gets populated in the datamovement\_errors table

Table name: data\_movement\_errors

|  |  |  |  |
| --- | --- | --- | --- |
| **Column name** | **Datatype** | **Nulls** | **Description** |
| process\_id | BIGINT | NO | An identifier for subtask |
| bot\_uUId | UUID | NO | referenced from bot |
| process\_context | CHARACTER VARYING | NO | process context |
| error\_json | TEXT | YES | JSON of movement |
| created\_ts | TIMESTAMP WITHOUT TIME ZONE | YES | time of creation |
| task\_map | TEXT | YES | map of movement |
| retry\_attempt | INTEGER | YES | number of retries attempted |
| bot\_type | CHARACTER VARYING | YES | type of bot |

## Data movement physical

This table contains pipeline information of each data movement

Table name: Data\_movement\_physical

|  |  |  |  |
| --- | --- | --- | --- |
| **Column name** | **Datatype** | **Nulls** | **Description** |
| data\_movement\_id | INTEGER | YES | an identifier for data movement referenced from the data movement table |
| data\_movement\_name | TEXT | YES | name of the data movement |
| contact\_info | JSON | YES | contact information of the owner |
| data\_movement\_type | TEXT | YES | type of data movement |
| refresh\_freq | CHARACTER VARYING | YES | refresh frequency ( daily, monthly, weekly) |
| connection\_profile\_id | SMALLINT | YES | an identifier for column profile type |
| partition\_type\_id | INTEGER | YES | an identifier for partition type |
| compute\_engine\_id | SMALLINT | YES | an identifier for compute engine |
| data\_movement\_engine\_id | SMALLINT | YES | identifier of data movement engine |
| workflow\_engine\_id | INTEGER | YES | an identifier for workflow engine referenced from the workflow engine |
| workflow\_id | INTEGER | YES | identifier for workflow |
| priority\_order | SMALLINT | YES | states priority |
| data\_movement\_additional\_info | JSON | YES | contains additional required info for the file |
| valid\_to\_ts | TIMESTAMP WITHOUT TIME ZONE | YES | the time to which the status record is valid. the current record should have a value of 31-dec-9999 |
| valid\_from\_ts | TIMESTAMP WITHOUT TIME ZONE | YES | the time from which the status record for this table metadata is valid |
| crt\_by | CHARACTER VARYING | YES | refers to the owner who created |
| crt\_ts | TIMESTAMP WITHOUT TIME ZONE | YES | refers to the time of the creation |
| mod\_by | CHARACTER VARYING | YES | displays the owner name who has modified it |
| mod\_ts | TIMESTAMP WITHOUT TIME ZONE | YES | refers the time of modification |

# Column Profiling Tables

## Crawl Xref

This table stores information on crawling

Table name: crawl\_xref

|  |  |  |  |
| --- | --- | --- | --- |
| **Column name** | **Datatype** | **Nulls** | **Description** |
| crawl\_end\_ts | TIMESTAMP WITHOUT TIME ZONE | YES | the time at which the crawl process has stopped |
| crawl\_id | UUID | YES | identifier for crawl |
| crawl\_start\_ts | TIMESTAMP WITHOUT TIME ZONE | YES | the time at which the crawl process has started |
| dataplace\_id | INTEGER | YES | An identifier for dataplace referenced from dataplace table |
| prev\_crawl\_id | UUID | YES | refers to the previous crawl\_id of crawl\_xref table |
| directory\_id | BIGINT | YES | identifier for directory |
| datalayer\_component\_id | SMALLINT | YES | identifier of a data component |
| schema\_id | BIGINT | YES | identifier for schema |
| valid\_from\_ts | TIMESTAMP WITHOUT TIME ZONE | YES | the time from which the status record for this table metadata is valid |
| valid\_to\_ts | TIMESTAMP WITHOUT TIME ZONE | YES | the time to which the status record is valid. the current record should have a value of 31-dec-9999 |
| crawl\_instance\_id | BIGINT | YES | an identifier for crawl instance |

## Crawler errors

This table contains error information that is captured while crawling

Table name: crawler\_errors

|  |  |  |  |
| --- | --- | --- | --- |
| **Column name** | **Datatype** | **Nulls** | **Description** |
| crawl\_id | UUID | YES | An identifier for crawling refers to the crawl Xref table |
| Hostname | CHARACTER VARYING | YES | hostname |
| error\_message | TEXT | YES | a status message of errors |
| errors\_recorded\_ts | TIMESTAMP WITHOUT TIME ZONE | YES | timestamp when an error is observed |
| crawl\_instance\_id | BIGINT | YES | an identifier for crawl instance |

## Crawling status

This table captures status information for crawling

Table name: crawling\_status

|  |  |  |  |
| --- | --- | --- | --- |
| **Column name** | **Datatype** | **Nulls** | **Description** |
| dataplace\_id | INTEGER | NO | an identifier for dataplace referenced from dataplace table |
| crawl\_instance\_id | BIGINT | NO | an identifier for crawl instance |
| status\_code\_id | SMALLINT | YES | identifier for status |
| start\_ts | TIMESTAMP WITHOUT TIME ZONE | YES | start time |
| end\_ts | TIMESTAMP WITHOUT TIME ZONE | YES | End-time |
| valid\_from\_ts | TIMESTAMP WITHOUT TIME ZONE | YES | the time from which the status record for this table metadata is valid |
| valid\_to\_ts | TIMESTAMP WITHOUT TIME ZONE | YES | the time to which the status record is valid. the current record should have a value of 31-dec-9999 |

## Crawling task status

This table stores status of the crawling

Table name: crawling\_task\_status

|  |  |  |  |
| --- | --- | --- | --- |
| **Column name** | **Datatype** | **Nulls** | **Description** |
| dataplace\_id | INTEGER | NO | an identifier for dataplace referenced from dataplace table |
| crawl\_instance\_id | BIGINT | NO | an identifier for crawl instance |
| process\_context | CHARACTER VARYING | YES | process context |
| status\_code\_id | SMALLINT | YES | identifier for status |
| task\_ts | TIMESTAMP WITHOUT TIME ZONE | YES | timestamp when the task is created |

# Engine Tables

## Engine mapping

This table contains information about engine mapping

Table name: engine\_mapping\_lookup

|  |  |  |  |
| --- | --- | --- | --- |
| **Column name** | **Datatype** | **Nulls** | **Description** |
| engine\_mapping\_id | INTEGER | NO | An identifier for engine mapping |
| engine\_id | SMALLINT | NO | identifier for engine |
| engine\_sub\_type\_id | SMALLINT | NO | an identifier for an engine subtype |
| additional\_info | JSON | YES | additional information for datastore |
| valid\_from\_ts | TIMESTAMP WITHOUT TIME ZONE | YES | the time from which the status record for this table metadata is valid |
| valid\_to\_ts | TIMESTAMP WITHOUT TIME ZONE | YES | the time to which the status record is valid. the current record should have a value of 31-dec-9999 |

## Engine sub type lookup

This table contains different types of engines

Table name: engine\_sub\_type\_lookup

|  |  |  |  |
| --- | --- | --- | --- |
| **Column name** | **Datatype** | **Nulls** | **Description** |
| engine\_sub\_type\_id | SMALLINT | NO | An identifier for an engine subtype |
| engine\_sub\_type | CHARACTER VARYING | NO | a subtype of engine |
| additional\_properties | JSON | YES | additional information for the workflow engine |
| config\_type\_id | INTEGER | YES | an identifier for configuration type |

## Workflow engine

This table stores details of the workflow engine

Table name: workflow\_engine

|  |  |  |  |
| --- | --- | --- | --- |
| **Column name** | **Datatype** | **Nulls** | **Description** |
| workflow\_engine\_id | INTEGER | NO | an identifier for the workflow engine |
| workflow\_engine\_name | TEXT | NO | name of workflow engine |
| additional\_info | JSON | YES | additional information for the workflow engine |
| kafka\_id | INTEGER | NO | identifier for Kafka referenced from Kafka table |
| valid\_to\_ts | TIMESTAMP WITHOUT TIME ZONE | YES | the time to which the status record is valid. the current record should have a value of 31-dec-9999 |
| valid\_from\_ts | TIMESTAMP WITHOUT TIME ZONE | YES | the time from which the status record for this table metadata is valid |
| cru\_by | CHARACTER VARYING | YES | displays the owner name who has created/modified it |
| cru\_ts | TIMESTAMP WITHOUT TIME ZONE | YES | refers the time of creation/modification |

## Workflow engine bots mapping

This table stores details of relation for workflow and bots

Table name: workflow\_engine\_bots\_mapping

|  |  |  |  |
| --- | --- | --- | --- |
| **Column name** | **Datatype** | **Nulls** | **Description** |
| workflow\_engine\_id | INTEGER | NO | refers to workflow engine table |
| bot\_factory\_id | INTEGER | NO | refers to bot factory table |
| bot\_type\_id | SMALLINT | NO | type of bot identifier |
| config\_id | INTEGER | YES | identifier for configuration |
| dataplace\_id | INTEGER | YES | An identifier for dataplace referenced from dataplace table |
| compute\_engine\_id | SMALLINT | YES | an identifier for compute engine |
| bot\_primary\_topic | TEXT | NO | topic details for bot |
| bot\_priority | SMALLINT | YES | priority of bot |
| bot\_topic | TEXT | YES | topic details |
| group\_name | CHARACTER VARYING | YES | group details for the topic |
| min\_bots | SMALLINT | YES | minimum number of bots |
| max\_bots | SMALLINT | YES | maximum number of bots |
| additional\_info | JSON | YES | additional information for the workflow engine |
| cru\_by | CHARACTER VARYING | YES | displays the owner name who has created/modified it |
| cru\_ts | TIMESTAMP WITHOUT TIME ZONE | YES | refers the time of creation/modification |

## Compute engine

This table contains details of compute engine

Table name: compute\_engine

|  |  |  |  |
| --- | --- | --- | --- |
| **Column name** | **Datatype** | **Nulls** | **Description** |
| compute\_engine\_id | SMALLINT | NO | an identifier for compute engine |
| compute\_engine\_name | TEXT | NO | name of engine |
| compute\_engine\_config\_id | INTEGER | YES | an identifier for compute engine for configuration |
| compute\_engine\_info | JSON | YES | information of engine |
| compute\_engine\_host\_name | CHARACTER VARYING | YES | name of engine |
| additional\_info | JSON | YES | additional information for datastore |
| valid\_from\_ts | TIMESTAMP WITHOUT TIME ZONE | YES | the time from which the status record for this table metadata is valid |
| valid\_to\_ts | TIMESTAMP WITHOUT TIME ZONE | YES | the time to which the status record is valid. the current record should have a value of 31-dec-9999 |
| cru\_by | CHARACTER VARYING | YES | displays the owner name who has created/modified it |
| cru\_ts | TIMESTAMP WITHOUT TIME ZONE | YES | refers the time of creation/modification |
| engine\_mapping\_id | SMALLINT | YES | an identifier for engine mapping |

# Bot Tables

## Bot configuration workflow

This table stores configuration of Bots that are associated with the workflow.

Table name: bot\_configuration\_workflow

|  |  |  |  |
| --- | --- | --- | --- |
| **Column name** | **Datatype** | **Nulls** | **Description** |
| workflow\_id | INTEGER | NO | identifier for workflow |
| workflow\_type | TEXT | NO | type of configuration |
| workflow\_name | TEXT | NO | name of configuration |
| data\_movement\_engine\_id | SMALLINT | YES | identifier of data movement engine |
| data\_movement\_engine\_type | CHARACTER VARYING | YES | type of movement\_engine |
| source\_type | CHARACTER VARYING | YES | type of source |
| destination\_type | CHARACTER VARYING | YES | type of destination |
| template\_info | JSON | YES | information of template configuration |
| engine\_mapping\_id | SMALLINT | YES | An identifier for engine mapping |

## Bot factory

Bot factory consumes the message and scales up the bots. Bot\_factory table consists of columns like workflow\_engine\_id and bot\_factory\_id.

Table name: bot\_factory

|  |  |  |  |
| --- | --- | --- | --- |
| **Column name** | **Datatype** | **Nulls** | **Description** |
| bot\_factory\_id | INTEGER | NO | an identifier for bot factory |
| workflow\_engine\_id | INTEGER | NO | An identifier for workflow engine referenced from the workflow engine |
| bot\_factory\_topic | TEXT | NO | details of topic bot factory |
| bot\_factory\_group\_name | CHARACTER VARYING | NO | name of the bot factory group |
| bot\_constraints | JSON | YES | constraints/configs for bots |
| env\_id | INTEGER | NO | identifier for environment |
| cru\_by | CHARACTER VARYING | YES | displays the owner name who has created/modified it |
| cru\_ts | TIMESTAMP WITHOUT TIME ZONE | YES | refers the time of creation/modification |

## Bot instance

This table stores details of the bot, whenever the bot gets scaled up, the bot\_uUId, bot\_type\_id, and other information are populated in the bot\_instance table.

Table name: bot\_instance

|  |  |  |  |
| --- | --- | --- | --- |
| **Column name** | **Datatype** | **Nulls** | **Description** |
| bot\_uUId | UUID | YES | referenced from bot |
| bot\_type\_id | SMALLINT | NO | type of bot identifier |
| bot\_factory\_message\_process\_id | BIGINT | NO | identifier for task |
| bot\_factory\_uUId | UUID | NO | identifier for botfactory |
| bot\_factory\_id | INTEGER | NO | an identifier for bot factory reference to bot factory |
| workflow\_engine\_id | INTEGER | NO | An identifier for workflow engine referenced from the workflow engine |
| bot\_topic | TEXT | YES | details of topic bot factory |
| status\_code\_id | SMALLINT | YES | identifier for status |
| cru\_by | CHARACTER VARYING | YES | displays the owner name who has created/modified it |
| cru\_ts | TIMESTAMP WITHOUT TIME ZONE | YES | refers the time of creation/modification |

## Bot status

This table consists of information related to the status of each bot.

Table name: bot\_status

|  |  |  |  |
| --- | --- | --- | --- |
| **Column name** | **Datatype** | **Nulls** | **Description** |
| bot\_uUId | UUID | YES | referenced from bot |
| bot\_type\_id | SMALLINT | NO | type of bot identifier |
| status\_code\_id | SMALLINT | YES | identifier for status |
| additional\_info | JSON | YES | additional information for datastore |
| cru\_ts | TIMESTAMP WITHOUT TIME ZONE | YES | refers the time of creation/modification |
| batch\_id | BIGINT | YES | an identifier for the task, unique id created for each job |
| process\_id | BIGINT | YES | An identifier for subtask |
| task\_name | CHARACTER VARYING | YES | name of task |

## Bot type

This table stores different types of Bots

Table name: bot\_type

|  |  |  |  |
| --- | --- | --- | --- |
| **Column name** | **Datatype** | **Nulls** | **Description** |
| bot\_type\_id | SMALLINT | NO | type of bot identifier |
| bot\_type | CHARACTER VARYING | NO | type of bot |
| bot\_classname | TEXT | YES | name of bot class |
| bot\_message\_bus\_class\_name | TEXT | YES | name of bot class message |
| control\_message\_bus\_class\_name | TEXT | YES | name of type |
| cru\_by | CHARACTER VARYING | YES | displays the owner name who has created/modified it |
| cru\_ts | TIMESTAMP WITHOUT TIME ZONE | YES | refers the time of creation/modification |

## Flow status

This table contains the status of the workflow

Table name: flow\_status

|  |  |  |  |
| --- | --- | --- | --- |
| **Column name** | **Datatype** | **Nulls** | **Description** |
| flow\_id | BIGINT | NO | identifier for flow |
| flow\_name | TEXT | NO | name of flow |
| batch\_id | BIGINT | NO | an identifier for the task, unique id created for each job |
| status\_code\_id | SMALLINT | NO | identifier for status |
| number\_of\_flow\_tasks | INTEGER | NO | tasks of status |
| flow\_ts | TIMESTAMP WITHOUT TIME ZONE | YES | timestamp of flow |
| job\_type\_id | SMALLINT | YES | an identifier for job type referenced from job type table (i.e profiling, ingestion, crawling ) |
| data\_movement\_id | INTEGER | YES | An identifier for data movement referenced from the data movement table |
| datastore\_id | INTEGER | YES | an identifier for datastore referenced from datastore table |
| job\_schedule\_id | BIGINT | YES | an identifier for job schedule |

## Flow task status

This table contains details of the task

Table name: flow\_task\_status

|  |  |  |  |
| --- | --- | --- | --- |
| **Column name** | **Datatype** | **Nulls** | **Description** |
| data\_movement\_id | INTEGER | YES | An identifier for data movement referenced from the data movement table |
| datastore\_id | INTEGER | YES | An identifier for datastore referenced from the datastore table |
| batch\_id | BIGINT | NO | An identifier for the task, unique id created for each job |
| process\_id | BIGINT | NO | An identifier for subtask |
| object\_id | BIGINT | NO | an identifier for table referenced from table metadata or file metadata |
| job\_type\_id | SMALLINT | NO | an identifier for job type referenced from job type table (i.e profiling, ingestion, crawling ) |
| status\_code\_id | SMALLINT | NO | identifier for status |
| retry\_count | SMALLINT | YES | number of retries |
| flow\_task\_ts | TIMESTAMP WITHOUT TIME ZONE | NO | timestamp of flow task |
| flow\_task\_date | DATE | NO | The date on which task is created |

## Job schedule details

This table contains details of the schedule of Job

Table name: job\_schedule\_details

|  |  |  |  |
| --- | --- | --- | --- |
| **Column name** | **Datatype** | **Nulls** | **Description** |
| job\_type\_id | SMALLINT | NO | an identifier for job type referenced from job type table (i.e profiling, ingestion, crawling ) |
| dataplace\_id | INTEGER | YES | An identifier for dataplace referenced from dataplace table |
| datastore\_id | INTEGER | YES | An identifier for datastore referenced from the datastore table |
| data\_movement\_id | INTEGER | YES | An identifier for data movement referenced from the data movement table |
| object\_id | JSON | YES | An identifier for table referenced from table metadata or file metadata |
| refresh\_frequency | CHARACTER VARYING | YES | refresh frequency ( daily, monthly, weekly) |
| additional\_properties | JSON | YES | additional information to be stored for job |
| valid\_from\_ts | TIMESTAMP WITHOUT TIME ZONE | YES | the time from which the status record for this table metadata is valid |
| valid\_to\_ts | TIMESTAMP WITHOUT TIME ZONE | YES | the time to which the status record is valid. the current record should have a value of 31-dec-9999 |
| cru\_by | CHARACTER VARYING | YES | displays the owner name who has created/modified it |
| cru\_ts | TIMESTAMP WITHOUT TIME ZONE | YES | refers the time of creation/modification |
| job\_schedule\_id | BIGINT | NO | An identifier for job schedule created incrementally |

# Messaging protocol

## Kafka

This table contains details of message brokers details of Kafka hosts, consists of information related to Kafka brokers, zookeeper URLs, producer config, consumer config, etc.

Table name: kafka

|  |  |  |  |
| --- | --- | --- | --- |
| **Column name** | **Datatype** | **Nulls** | **Description** |
| kafka\_id | INTEGER | NO | identifier for kafka referenced from kafka table |
| kafka\_brokers | TEXT | NO | host details |
| zookeeper\_urls | TEXT | NO | URL for zookeeper |
| private\_key\_path | TEXT | YES | private key |
| public\_key\_path | TEXT | YES | public key |
| private\_key\_value | TEXT | YES | salt applied on the private key |
| public\_key\_value | TEXT | YES | salt applied on public key |
| producer\_config | JSON | YES | producer details |
| consumer\_config | JSON | YES | consumer details |
| max\_message\_size | INTEGER | YES | maximum size of the message |
| additional\_properties | JSON | YES | additional information to be stored for Kafka |
| valid\_to\_ts | TIMESTAMP WITHOUT TIME ZONE | YES | the time to which the status record is valid. the current record should have a value of 31-dec-9999 |
| valid\_from\_ts | TIMESTAMP WITHOUT TIME ZONE | YES | the time from which the status record for this table metadata is valid |
| cru\_by | CHARACTER VARYING | YES | displays the owner name who has created/modified it |
| cru\_ts | TIMESTAMP WITHOUT TIME ZONE | YES | refers the time of creation/modification |

## Kafka topics

This table contains details related to Kafka topics

Table name: kafka\_topics

|  |  |  |  |
| --- | --- | --- | --- |
| **Column name** | **Datatype** | **Nulls** | **Description** |
| kafka\_id | INTEGER | NO | identifier for kafka referenced from kafka table |
| kafka\_topics | JSON | YES | topics created in kafka |

# Schedules, Configuration, and Status Tables:

## Process id table map

This table contains the information of process\_id links with specific object\_id.

Table name: process\_id\_table\_map

|  |  |  |  |
| --- | --- | --- | --- |
| **Column name** | **Datatype** | **Nulls** | **Description** |
| data\_movement\_id | INTEGER | YES | An identifier for data movement referenced from the data movement table |
| process\_id | BIGINT | NO | an identifier for subtask |
| object\_id | BIGINT | YES | an identifier for table referenced from table metadata or file metadata |
| process\_started\_ts | TIMESTAMP WITHOUT TIME ZONE | NO | details of when the process is started |
| batch\_id | BIGINT | NO | An identifier for the task, unique id created for each job |
| batch\_name | CHARACTER VARYING | YES | name of batch/job |
| Environment | CHARACTER VARYING | YES | environment |

## spark\_job\_result

This table contains the information on the status of the task for each spark bot.

Table name: spark\_job\_result

|  |  |  |  |
| --- | --- | --- | --- |
| **Column name** | **Datatype** | **Nulls** | **Description** |
| process\_id | BIGINT | NO | An identifier for subtask |
| application\_id | CHARACTER VARYING | YES | An identifier for job application |
| error\_details | TEXT | YES | error |
| Status | BOOLEAN | YES | status of the task started, stopped, failed, interrupted used for UI |
| exit\_status | SMALLINT | YES | interrupted status |
| additional\_info | JSON | YES | additional information of spark job |

## Status code

This table stores the status code which is used for identifying if the task is started stopped, failed, etc.

Table name: status\_code\_lookup

|  |  |  |  |
| --- | --- | --- | --- |
| **Column name** | **Datatype** | **Nulls** | **Description** |
| status\_code\_type | CHARACTER VARYING | NO | type of status related to flow, task, dependency |
| status\_code | CHARACTER VARYING | NO | status of the task started, stopped, failed, interrupted |
| status\_code\_id | SMALLINT | NO | identifier for status |
| status\_description | TEXT | YES | description |
| status\_code\_label | CHARACTER VARYING | YES | status of the task started, stopped, failed, interrupted used for UI |
| status\_code\_color | CHARACTER VARYING | YES | color to be represented in UI green for success, red for failure, yellow for running |
| additional\_info | JSON | YES | additional information of the status |

## Template group dependency

This table stores details of dependency of templates

Table name: template\_group\_dependency

|  |  |  |  |
| --- | --- | --- | --- |
| **Column name** | **Datatype** | **Nulls** | **Description** |
| template\_group\_id | BIGINT | NO | an identifier for template group |
| dependant\_template\_group\_id | BIGINT | NO | an identifier for dependent template group i.e ( common templates for import ) |
| cru\_by | CHARACTER VARYING | YES | displays the owner name who has created/modified it |
| cru\_ts | TIMESTAMP WITHOUT TIME ZONE | YES | refers the time of creation/modification |

## Template group info

This table stores template information

Table name: template\_group\_info

|  |  |  |  |
| --- | --- | --- | --- |
| **Column name** | **Datatype** | **Nulls** | **Description** |
| template\_group\_id | BIGINT | NO | an identifier for template group |
| template\_group\_name | TEXT | NO | name of the template group |
| description\_of\_templategroup | TEXT | NO | description of template group |
| template\_group\_value | TEXT | NO | template |
| template\_version | NUMERIC | NO | version of template |
| git\_version | CHARACTER VARYING | YES | a version of the template on git |
| classification | TEXT | NO | classification of template group i.e is it for indexing, crawling, profiling, ingestion |
| additional\_info | JSON | YES | additional information for template group |
| valid\_from\_ts | TIMESTAMP WITHOUT TIME ZONE | YES | the time from which the status record for this table metadata is valid |
| valid\_to\_ts | TIMESTAMP WITHOUT TIME ZONE | YES | the time to which the status record is valid. the current record should have a value of 31-dec-9999 |
| cru\_by | CHARACTER VARYING | YES | displays the owner name who has created/modified it |
| cru\_ts | TIMESTAMP WITHOUT TIME ZONE | YES | refers the time of creation/modification |

## Vault details

This table contains details of Vault which are used for authentication

Table name: vault\_details

|  |  |  |  |
| --- | --- | --- | --- |
| **Column name** | **Datatype** | **Nulls** | **Description** |
| vault\_id | INTEGER | NO | this column contains the id for Nabu |
| vault\_name | TEXT | NO | name of vault |
| vault\_description | TEXT | YES | refers to the description for vault |
| vault\_base\_url | TEXT | NO | refers to URL for vault |
| vault\_authentication\_type | TEXT | NO | refers to authentication type - i.e., username/password, Kerberos, API key, etc |
| vault\_authentication\_details | JSON | NO | details of encrypted details for the vault to authenticate |
| vault\_type | CHARACTER VARYING | NO | refers to vault type - harshicorp |
| vault\_default\_entry\_path | CHARACTER VARYING | YES | refers to the path wherein the vault is stored |
| additional\_info | JSON | YES | refers to any other additional pipelines |
| cru\_by | CHARACTER VARYING | YES | displays the owner name who has created/modified it |
| cru\_ts | TIMESTAMP WITHOUT TIME ZONE | YES | refers the time of creation/modification |

## Checkpoint status

This table contains details of checkpoint i.e stores details about ingestion

Table name: checkpoint\_status

|  |  |  |  |
| --- | --- | --- | --- |
| **Column name** | **Datatype** | **Nulls** | **Description** |
| dataplace\_id | INTEGER | YES | an identifier for dataplace referenced from dataplace table |
| datastore\_id | INTEGER | YES | an identifier for datastore referenced from datastore table |
| data\_movement\_id | INTEGER | YES | an identifier for data movement referenced from the data movement table |
| process\_id | BIGINT | YES | an identifier for subtask |
| table\_id | BIGINT | YES | an identifier for table reference to table metadata |
| datastore\_table\_id | BIGINT | YES | An identifier for table referenced from datastore table |
| status | CHARACTER VARYING | YES | status of the task started, stopped, failed, interrupted used for UI |
| application\_id | CHARACTER VARYING | NO | An identifier for job application |
| job\_type | CHARACTER VARYING | NO | type of job (ingestion, profiling, curation, crawling, indexing) |
| hostname | CHARACTER VARYING | NO | hostname |
| error\_msg | TEXT | YES | A status message of errors |
| start\_time | TIMESTAMP WITH TIME ZONE | YES | start time |
| end\_time | TIMESTAMP WITH TIME ZONE | YES | End-time |
| valid\_from\_ts | TIMESTAMP WITHOUT TIME ZONE | YES | the time from which the status record for this table metadata is valid |
| valid\_to\_ts | TIMESTAMP WITHOUT TIME ZONE | YES | the time to which the status record is valid. the current record should have a value of 31-dec-9999 |
| verification\_status | BOOLEAN | YES | status |

## Credential info

This table stores information about credentials

Table name: credential\_info

|  |  |  |  |
| --- | --- | --- | --- |
| **Column name** | **Datatype** | **Nulls** | **Description** |
| credential\_id | INTEGER | NO | identifier for credential |
| credential\_name | CHARACTER VARYING | YES | refers to the name of the credential |
| credential\_type\_id | SMALLINT | NO | identifier of credential mapping |
| credential\_source | CHARACTER VARYING | NO | refers if the credential is added by the user or is already present and used |
| vault\_id | INTEGER | NO | this refers to vault id from vault\_details table |
| vault\_entry\_key | TEXT | YES | vault key details are stored to be integrated with the vault on the back-end |
| vault\_entry\_details | JSON | YES | displays details of vault |
| additional\_info | JSON | YES | additional information for datastore |
| valid\_from\_ts | TIMESTAMP WITHOUT TIME ZONE | YES | the time from which the status record for this table metadata is valid |
| valid\_to\_ts | TIMESTAMP WITHOUT TIME ZONE | YES | the time to which the status record is valid. the current record should have a value of 31-dec-9999 |
| crt\_by | CHARACTER VARYING | YES | refers to the owner who created |
| crt\_ts | TIMESTAMP WITHOUT TIME ZONE | YES | refers to the time of the creation |
| mod\_by | CHARACTER VARYING | YES | displays the owner name who has modified it |
| mod\_ts | TIMESTAMP WITHOUT TIME ZONE | YES | refers the time of modification |

## Credential type

This table stores different types of credentials

Table name: credential\_type\_lookup

|  |  |  |  |
| --- | --- | --- | --- |
| **Column name** | **Datatype** | **Nulls** | **Description** |
| credential\_type\_id | SMALLINT | NO | identifier of credential mapping |
| credential\_type | CHARACTER VARYING | NO | this field refers to credential types like LDAP, plain, Kerberos, as auth, GCP auth |
| credential\_validation\_schema | JSON | NO | validates the connection and this template is used for JSON types |
| credential\_key\_names | JSON | NO | all the extra keys/policies are added to access data in UI |
| credential\_UI\_json | JSON | YES | all the extra keys/policies are added to access UI |

# Advanced options tables

## Advanced options datatype mappings

This table is used for any customization reqUIred in data

Table Name: advanced\_options\_datatype\_mappings

|  |  |  |  |
| --- | --- | --- | --- |
| **Column name** | **Type** | **Nulls** | **Description** |
| datamovement\_id | smallint | NO | An identifier for data movement referenced from the data movement table |
| datatype\_mapping\_id | bigint | NO | An identifier for datastore mapping |
| additional\_info | JSON | YES | additional information for datastore |
| valid\_from\_ts | timestamp without time zone | YES | the time from which the status record for this table metadata is valid |
| valid\_to\_ts | timestamp without time zone | YES | the time to which the status record is valid. the current record should have a value of 31-dec-9999 |
| cru\_by | character varying | YES | displays the owner name who has created/modified it |
| cru\_ts | timestamp without time zone | YES | refers the time of creation/modification |

## Advanced options

This table stores different types of advanced options

Table name: advanced\_options\_lookup

|  |  |  |  |
| --- | --- | --- | --- |
| **Column name** | **Datatype** | **Nulls** | **Description** |
| advanced\_options\_type\_id | SMALLINT | YES | identifier for advanced options |
| advanced\_options\_type | CHARACTER VARYING | YES | type of options |
| additional\_info | JSON | YES | additional information for advanced options |

## Advanced options table ingestion

Provides details of any options that are applied while ingestion

Table name: advanced\_options\_table\_ingestion

|  |  |  |  |
| --- | --- | --- | --- |
| **Column name** | **Datatype** | **Nulls** | **Description** |
| table\_advanced\_options\_id | BIGINT | NO | identifier for advanced options |
| source\_table\_id | BIGINT | NO | identifier for table |
| data\_movement\_id | INTEGER | NO | an identifier for data movement referenced from the data movement table |
| advanced\_options\_condition | JSON | YES | conditional clause |
| valid\_from\_ts | TIMESTAMP WITHOUT TIME ZONE | YES | the time from which the status record for this table metadata is valid |
| valid\_to\_ts | TIMESTAMP WITHOUT TIME ZONE | YES | the time to which the status record is valid. the current record should have a value of 31-dec-9999 |

## Partition column info

This table contains partition details for the columns which are at the source

Table name: partition\_column\_info

|  |  |  |  |
| --- | --- | --- | --- |
| **Column name** | **Datatype** | **Nulls** | **Description** |
| dataplace\_id | INTEGER | NO | An identifier for dataplace referenced from dataplace table |
| dataplace\_component\_type\_id | SMALLINT | NO | identifier for component type |
| schema\_id | BIGINT | NO | identifier for schema |
| schema\_name | TEXT | NO | name of schema |
| table\_id | BIGINT | NO | an identifier for table reference to table metadata |
| table\_name | TEXT | NO | name of table |
| column\_id | BIGINT | YES | An identifier for column referenced from column metadata |
| column\_name | TEXT | NO | name of the column |
| ordinal\_position | INTEGER | NO | position of column ( index of column) |
| additional\_info | JSON | YES | additional information to be stored for partition |
| valid\_from\_ts | TIMESTAMP WITHOUT TIME ZONE | YES | the time from which the status record for this table metadata is valid |
| valid\_to\_ts | TIMESTAMP WITHOUT TIME ZONE | YES | the time to which the status record is valid. the current record should have a value of 31-dec-9999 |
| crawl\_id | UUID | YES | an identifier for crawling refers to the crawl Xref table |
| prev\_crawl\_id | UUID | YES | refers to the previous crawl\_id of crawl\_xref table |

## Partitions info

This table contains data places wherein partition type table ingestion is triggered

Table name: partitions\_info

|  |  |  |  |
| --- | --- | --- | --- |
| **Column name** | **Datatype** | **Nulls** | **Description** |
| dataplace\_id | INTEGER | NO | an identifier for dataplace referenced from dataplace table |
| dataplace\_component\_type\_id | SMALLINT | NO | identifier for component type |
| schema\_id | BIGINT | NO | identifier for schema |
| schema\_name | TEXT | NO | name of schema |
| table\_id | BIGINT | NO | an identifier for table reference to table metadata |
| table\_name | TEXT | NO | name of table |
| partition\_id | BIGINT | YES | type of partition |
| partition\_name | TEXT | NO | name of partition |
| partition\_position | INTEGER | YES | position of partition |
| db\_partition\_position | INTEGER | YES | position of partition in the database |
| num\_of\_rows | BIGINT | YES | number of rows in the partition |
| is\_compressed | BOOLEAN | YES | state of partition is compressed or not |
| low\_value | TEXT | YES | low value |
| high\_value | TEXT | YES | high value in the column |
| additional\_info | JSON | YES | additional information to be stored for partition |
| valid\_from\_ts | TIMESTAMP WITHOUT TIME ZONE | YES | the time from which the status record for this table metadata is valid |
| valid\_to\_ts | TIMESTAMP WITHOUT TIME ZONE | YES | the time to which the status record is valid. the current record should have a value of 31-dec-9999 |
| crawl\_id | UUID | YES | an identifier for crawling refers to the crawl Xref table |
| prev\_crawl\_id | UUID | YES | refers to the previous crawl\_id of crawl\_xref table |

## Nabu Roles

Nabu roles define the functionality of the application that is accessible for each user or LDAP Group.

Roles can be created for individual users or LDAP groups, with the functionality pages and access permissions for those users/groups to access the Nabu application.

Credentials, Datastores, Synonym, Entity, Facet, Filter, Fieldstore, Data Place, Ingestion, Monitoring dashboard, Ingestion scheduling and Manage Access are the functionality pages that are in Nabu.

For each functionality page, the users have below three access permissions.

**View:** view access permission for any functionality page user can view the dashboard.

For example: with view permission to datastores functionality page, the user can only view the list of datastores on the datastore dashboard.

**Modify:** modify access permission for any functionality page user can view, modify and delete the existing items on the dashboard.

For example: with modify permission to the datastores functionality page, the user can view, modify and delete any existing datastores on the datastore dashboard.

**Create:** create access permission for any functionality page user can create new items, view, modify and delete the existing items on the dashboard.

For example: with creating permission to the datastores functionality page, the user can create a new datastore, view, modify and delete any existing datastores on the datastore dashboard.

## Data Access Roles

Data access roles define the access permission of each datastore or datastore group for each user or LDAP group. A datastore group is nothing but a group of datastores for which all we are providing access permissions together.

While creating data access roles, users can select a list of users/groups and the access permissions for those users/groups individually.

Below are the access permissions levels of datastores.

**Metadata**: with metadata access permission, the user can search for only the datastore, tables, and columns of that datastore.

**Read:** read access permission, the user will have metadata access and also can see the table data and can search with the column value.

**Reviewer**: reviewer access permission, the user will have metadata, read access and also able to create tags for tables and columns, add a description for tables, add a comment to column and add review

**Write** access permission, user can perform transformations of tables.

**Datastore\_Owner:** with datastore\_owner access permission, the user will have complete access to the datastore.

**Roles Data Model:**

## Role scope lookup

role\_scope\_lookup table stores the details of role type (example: UI, Datastore, Datastore Group)

|  |  |  |
| --- | --- | --- |
| **Column name** | **Datatype** | **Description** |
| role\_scope\_id | bigint NOT NULL | This Field refers to incremental id for the scope of the role |
| role\_scope\_name | character varying(128) | Refers to the name of role scope |

## UI access type lookup

UI\_access\_type\_lookup table stores the details of access permission types (create, modify, view)

|  |  |  |
| --- | --- | --- |
| **Column name** | **Datatype** | **Description** |
| UI\_access\_type\_id | Bigint, NOT NULL | This Field refers to incremental id for UI access type |
| UI\_access\_type\_name | character varying(128) | Refers to the name type of UI access |
| UI\_access\_hierarchy\_order | Bigint | Refers to the hierarchy order of the UI access |

## UI policy lookup

UI\_policy\_lookup table stores the details of functionality pages and access permissions with unique UI\_policy\_id.

|  |  |  |
| --- | --- | --- |
| **Column name** | **Datatype** | **Description** |
| UI\_policy\_id | Bigint, NOT NULL | This Field refers to incremental id for UI policy |
| UI\_policy\_name | character varying(128) | Refers to the name of UI policy |
| UI\_access\_type\_id | Bigint, NOT NULL | This Field refers to incremental id for UI access type |
| UI\_policy\_description\_info | character varying(128) | Refers to the description information of UI policy |

## Roles info

roles\_info table stores the details of the role whenever a new role is created. Each role has a unique role\_id.

|  |  |  |
| --- | --- | --- |
| **Column name** | **Datatype** | **Description** |
| role\_id | Bigint, NOT NULL | This Field refers to incremental id for role |
| role\_name | character varying(128) | Refers to the name of the role |
| role\_scope\_id | bigint NOT NULL | This Field refers to incremental id for role scope |
| other\_info | JSON | Contains description of the entity |
| valid\_from\_ts | timestamp (6) without time zone | States the time at which the file is valid |
| valid\_to\_ts | timestamp (6) without time zone | States the time till when the file is valid |
| crt\_by | character varying (128) | Refers to the owner who created |
| crt\_ts | timestamp (6) without time zone | Refers to the time of the creation |
| mod\_by | character varying (128) | Refers to the owner who modified |
| mod\_ts | timestamp (6) without time zone | Refers the time of modification |

## Roles UI policy

roles\_UI\_policy tables store the details of roles assigned(UI\_policy\_id) for each role\_id.

|  |  |  |
| --- | --- | --- |
| **Column name** | **Datatype** | **Description** |
| role\_id | Bigint, NOT NULL | This Field refers to incremental id for role |
| UI\_policy\_id | Bigint, NOT NULL | This Field refers to incremental id for UI policy |

## Environment group role

env\_group\_role\_info table stores the details of user/group and their roles ids

|  |  |  |
| --- | --- | --- |
| **Column name** | **Datatype** | **Description** |
| env\_group\_type\_id | bigint NOT NULL | This Field refers to incremental id for the type of the env group |
| env\_group\_name | character varying(128) | Refers to the name of env group |
| role\_id | Bigint, NOT NULL | This Field refers to incremental id for role |

## Environment group type lookup

env\_group\_type\_lookup tables state whether it is user or group

|  |  |  |
| --- | --- | --- |
| **Column name** | **Datatype** | **Description** |
| env\_group\_type\_id | bigint NOT NULL | This Field refers to incremental id for the type of the env group |
| env\_group\_name | character varying(128) | Refers to the name of env group |