Office of the Chief Digital Officer

Data Lake Naming Conventions

0.5

# Revision History

|  |  |  |  |
| --- | --- | --- | --- |
| Revision Date | Version No. | Author | Description of Change/Revision |
| 13/11/2017 | 0.6 | Selva Murugesan | Revised snapshot resction |
| 29/09/2017 | 0.5 | Robert A. Marshall | Removed ‘Active Directory Group’ |
| 29/09/2017 | 0.4 | Robert A. Marshall | Corrections |
| 28/09/2017 | 0.3 | Robert A. Marshall | Formatting and proof-reading |
|  | 0.2 | Selva Murugesan |  |
|  | 0.1 | Selva Murugesan |  |

Contents

[Revision History 2](#_Toc494460192)

[Principles 3](#_Toc494460193)

[Volumes and sub-volumes 4](#_Toc494460194)

[MapR Roles 5](#_Toc494460195)

[Access Views 6](#_Toc494460196)

[Database Name 7](#_Toc494460197)

[Table Name 8](#_Toc494460198)

[Flat File Name 9](#_Toc494460199)

[Partition Name 10](#_Toc494460200)

[Streamsets Jobs Name 11](#_Toc494460201)

[Oozie Workflow and Coordinators Name 12](#_Toc494460202)

[Field Names 14](#_Toc494460203)

[Reports in Business Intelligence Tools 15](#_Toc494460204)

[Snapshots 15](#_Toc494460205)

[Sandpits for Data Scientist 16](#_Toc494460206)

# Principles

The below principles sets the precedence for naming files in the data lake

1. Underscore in the names are allowed
2. English (Australia) is the preferred language
3. Prefixes and Suffixes are allowed
4. Spaces in the names are not allowed
5. General acronyms known to ACT Government users are allowed
6. Version numbering to be included in the name for Streamsets and Oozie jobs
7. Names should contain only alphanumeric
8. Names should not start with a number
9. Names should always start with a letter
10. Camel Case to be used
11. Names should be short as possible and must not exceed more than 64 characters

The above principles should be adhered to naming the below categories. There are special circumstances where some principles have exceptions.

1. Volumes and sub-volumes
2. Active Directory Group
3. MapR Roles
4. Access Views
5. Database name
6. Table name
7. Flat file name
8. Streamsets jobs
9. Oozie jobs
10. Reports in BI tools
11. Field names
12. Snapshots
13. Sandpits for Data Scientists

# Volumes and sub-volumes

The structure of the volumes and sub-volumes are listed below

Directorate

Business Area

Business System 1

Landing

Processed

.

.

Business System 2

Landing

Processed

.

.

Naming Conventions

* The volume name is the directorate name
* The sub-volume name is the business area; Sub-volume should be as short as possible and should describe the business area
* The sub-sub-volume name is the business system name
* Landing is the default directory for storing the raw files (Landing Zone)
* Processed is the default directory for storing the processed files (after transformation, filtering, joining, etc.)

# MapR Roles

The principles for naming MapR roles are

* MapR roles names should identify the type of role they perform
* No ambiguity in name
* Camel Case should be used
* MapR Role names can have suffix that contains tenancy name

## Examples

ClusterAdmin

DataScientist\_Health

DataAnalyst\_TCCS

DataEngineer\_Education

# Access Views

Format for naming the access views is

[View\_] {TableName / UsageViewName} [\_All]

The prefix and suffix are mandatory as denoted in []. The table name / usage view name as denoted in {} are optional suggestions, this section can be named as considered appropriate.

The list of allowed suffixes are

|  |  |  |
| --- | --- | --- |
| Case No | Suffix | Usage description |
| 1 | \_All | View contains all the columns and all records of the source table |
| 2 | \_DeIdentified | View contains deidentified columns |
| 3 | \_Filtered | View is filtered based on some criteria; This is to show the view has row level restriction |
| 4 | \_Aggregated | Records are aggregated |
| 5 | \_DeIdentifiedFiltered | View contains the deidentified columns and also filtered based on some criteria |
| 6 | \_Other descriptions based on view’s purpose/use case | If a view is created with some complex operations that cannot be categorised as Case No 1 -5, then use any brief description that suits.  Eg. View\_HealthStat\_AHIW |

The list of allowed prefix is

|  |  |
| --- | --- |
| Prefix | Usage description |
| View\_ | All the views must be prefixed with “View\_” |

## Examples

Table name: Coverages

Access View name

* View\_Coverages\_All

If the view contains all the records and columns of the table coverages

* View\_Coverages\_Aggregated

If the view is aggregated based some criteria (based on age groupings, coverage period, etc...)

* View\_Coverage\_Compliance
* If this view is created to submit some data for compliance purpose
* View\_CoveragesMay2017\_Filtered and View\_CoveragesJune2017\_Filtered

One filtered to entries in May of 2017 and the other to June of 2017.

# Database Name

The naming conventions for naming the database are

* Must be unique / mirror business system as close as possible
* Must adhere to principles
* Camel Case is allowed
* No plurals
* Landing Zone
  + Must use a suffix (\_Landing) for landing database
  + Recommended if the user is bringing in data from only one business system
* Processed Zone
  + No suffix required
* Intermediate
  + Must use a suffix (\_Intermediate) for intermediate database that was created by joining multiple sources

## Example 1

If the business system name is RiskMan

Then

|  |  |
| --- | --- |
| Landing Zone Database name | RiskMan\_Landing |
| Processed Zone Database | RiskMan |
| Intermedia Database name | NIL |

## Example 2

If you are building a database from two different business system name e.g RiskMan and AIMS

Then

|  |  |
| --- | --- |
| Landing Zone Database name | NIL |
| Processed Zone Database | RiskManAIMS |
| Intermedia Database name | RiskManAIMS\_Intermediate |

# Table Name

The naming conventions for naming the table are

* Adheres to the naming convention principles
* Landing Zone
  + Should reflect the source database table name / file name
  + No prefix / suffix is allowed
  + Use Camel Case
  + No plurals are allowed
* Processed Zone
  + No prefix / suffix is allowed
  + Use Camel Case
  + No plurals are allowed
  + Name should describe as close to data content
* Hive Table Name
  + Should reflect the source database table name / file name

<Table names for dimensional data>

## Example

If SQL Server Table name (source) is ‘Policy’ or ‘Policies’

Then

|  |  |
| --- | --- |
| Landing Zone File name | Policy |
| Processed Zone File | Policy |
| Hive table name | Policy |

# Flat File Name

This naming convention should be followed if you are ingesting data from a flat file rather than databases/tables.

* Landing
  + The flat file name must adhere to generic naming principles
  + No prefix / suffix allowed
  + Camel Case is required for the file and table name
  + The flat file name should be the same as the associated table name

For example if AFP sends a file that contains crime statistics and the file name is 123crimes.csv then

Landing

|  |  |
| --- | --- |
| MapR-FS name | Crime.csv |
| Hive database name | AFP\_Landing |
| Hive table name | Crime |

* Processed

Format for naming the processed flat file is

{TableName}[\_Business Context / \_Purpose / \_ reason]

## Example 1

If you have liquor license data that needs to be published as a part of legislative obligation, create a processed file names

LiquorLicense\_OpenData

## Example 2

If you have to submit some health statistics to AIHM, then create a processed file names

HealthStatistics\_AIHW

# Partition Name

The naming conventions for partitions in a HIVE table are

* Retain the same field name if possible
* Field name as per the [Field Name conventions](#_Field_Names)
* Use singular name
* Must be alphanumeric
* Suffix must be proceeded by an underscore
  + Suffix Options

\_D : If the filed name data type is Date

\_T : If the filed name data type is Time

\_DT : If the filed name data type is Date and Time

## Example 1

When the table is partitioned based on a date field named License Expiry

Partition name is LicenseExpiry\_D

## Example 2

When the data is partitioned based on a String field named Venue Location

Partition name is VenueLocation

## Example 3

When the table is partitioned based on multiple fields such as LicenseRegistered, LicenseExpiry, VenueLocation then

The partition names are LicenseRegistered\_DT, LicenseExpiry\_D, and VenueLocation

# Streamsets Jobs Name

The naming conventions for naming the Streamsets jobs are

* Should start with an alphabet
* Camel Case should be used

Format is

{Ingest\_ / Process\_ / Publish\_} {SourceDatabaseName + \_TargetTableName} {\_V1.0}

* Use major revision for plan change
* Use Minor revision for bug fixes

## Example 1

If the Streamsets job is ingesting data from the ‘ABC’ business system, whose table name is events then

Ingest\_ABC\_Events\_V1.0

If you have some changes to the above file to fix some bugs / add minor features, then next version of this file should be

Ingest\_ABC\_Events\_V1.1

If you have done major changes to the ingestion process, then next version should be

Ingest\_ABC\_Events\_V2.0

## Example 2

If the Streamsets job is performing any processing (transformation, etc..) then it should be named as

Process\_ABC\_Events\_V1.0

## Example 3

If the Streamsets job is publishing any data (to open data portal, etc..) then it should be named as

Publish\_ABC\_Events\_V1.0

# Oozie Workflow and Coordinators Name

## Workflow

The naming conventions for naming the Oozie workflow are

* Should start with a letter (not a number or symbol)
* Camel Case should be used

Format is

{BusinessContext} {\_V1.0}

* Use major revision for plan change
* Use minor revision for bug fixes

## Example 1

If the Oozie workflow is scheduling a few Streamsets jobs to be run followed by spark job and so on and this scheduled workflow is to automate liquor licensing reporting, then name the Oozie workflow as

LiquorLicenseReporting\_V1.0

## Example 2

If the Oozie workflow is scheduling a few Streamsets jobs to be run followed by spark job to perform some data analytics operations and this scheduled workflow is to automate a machine learning model output for the XYZ project then

MachineLearning\_XYZProject\_V1.0

## Coordinators

The naming conventions for naming the Oozie coordinators are

* Should start with a letter (not a number or symbol)
* Camel Case should be used

Format is

{BusinessContext} {\_frequency}

* The suffix \_frequency should contain the below arguments

Yearly: workflow running yearly

Monthly: workflow running monthly

Weekly: workflow running monthly

Daily: workflow running daily

Hourly: workflow running hourly

## Example 1

If the Oozie coordinator’s job is scheduling a workflow that automates a Streamsets job for liquor licensing report generation, which runs every month, then name the Oozie coordinator’s job as

LiquorLicenseReporting\_Monthly

# Field Names

The naming conventions for naming the field names are

* Landing
  + Retain the same field name if possible
  + No spaces in field name
  + Must be Alphanumeric
* Processed
  + Start with a letter
  + Camel Case should be used
  + Suffix should use underscore
  + Use singular name
* Suffix Options

\_D : If the filed name data type is Date

\_T : If the filed name data type is Time

\_DT : If the filed name data type is Date and Time

\_ID : If the filed name is the Primary Key / Foreign Key

## Example 1

If the fieldname is 123 address, then the new field name shall be Address

## Example 2

If the fieldname is 123 addresses, then the new field name shall be Address

## Example 3

If the fieldname is shop name, then the new field name shall be ShopName

## Example 4

If the fieldname is appointmentdate, then the new field name shall be AppointmentDate\_D

# Reports in Business Intelligence Tools

Format is

{BusinessContext\_}{Business use-case/database/business system} {\_V1.0}

* Use major revision for plan change
* Use minor revision for bug fixes

The list of approved prefixes for business context are

BS\_ : For denoting business system

<projectname>\_ : For denoting reports belonging to a specific project

DB\_ : For denoting database

## Example 1

If the report is built in Tableau BI tools for producing lists of liquor licenses for a project called ‘ABCProject’, then name the file

ABCProject\_liquorlicense\_V1.0

# Snapshots

Format for Snapshot schedules

{VolumeName}{\_frequency}

* The suffix \_frequency should contain the below arguments

<Num>Min : If snapshots are taken minute wise; Allowed <Num> parameter is 1,5,10,15,20,25,30,45

Hourly: snapshots are taken at hourly

Daily: If snapshots are taken daily

Weekly: snapshots taken weekly

Monthly: snapshots taken monthly

Yearly: snapshots taken yearly

YYYY-MM-DD : snapshot is taken only once

# Sandpits for Data Scientist

The structure of the sand pits for data scientist. These sandpits is where the data scientist have more flexibility to play with different datasets. These sandpits will be removed once the project finishes based on stakeholder’s approval.

Directorates

Project Sandpit

Project Name 1

Project Name 2

.

.