

Dataikuプロセス整理術

2024/10

アジェンダ

なぜ整理術を身につけることは重要か？

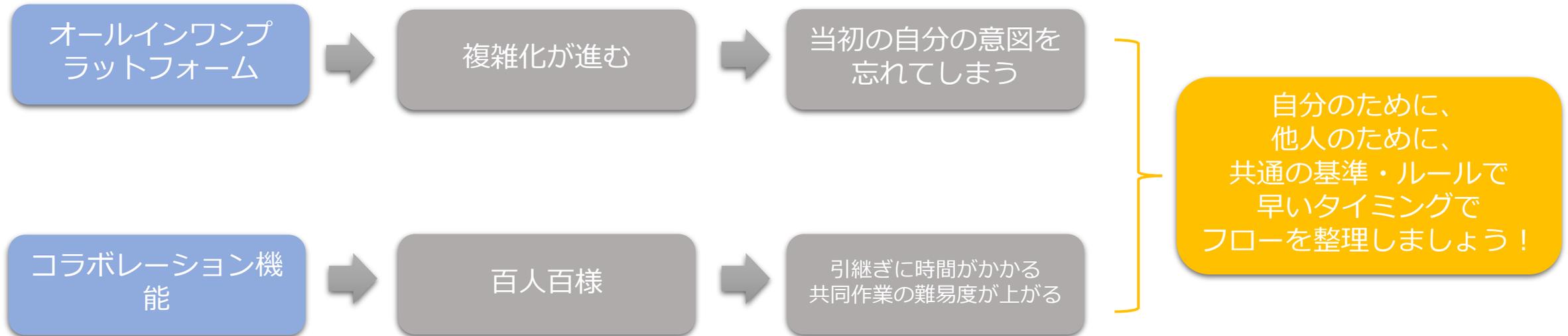
フローの整理術：その1~6

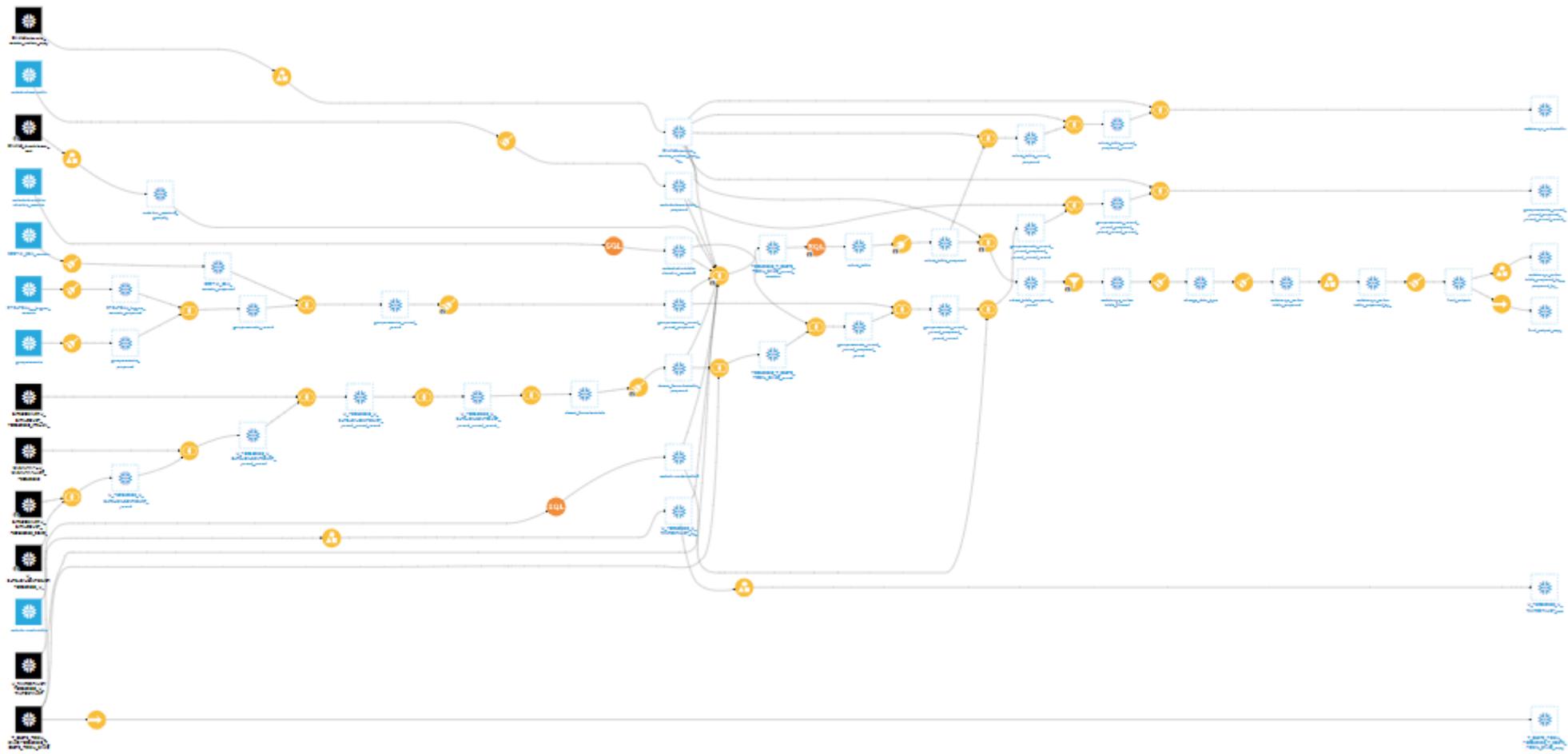
フロー整理に関連した「あるといいな」

Dataiku フロー整理術

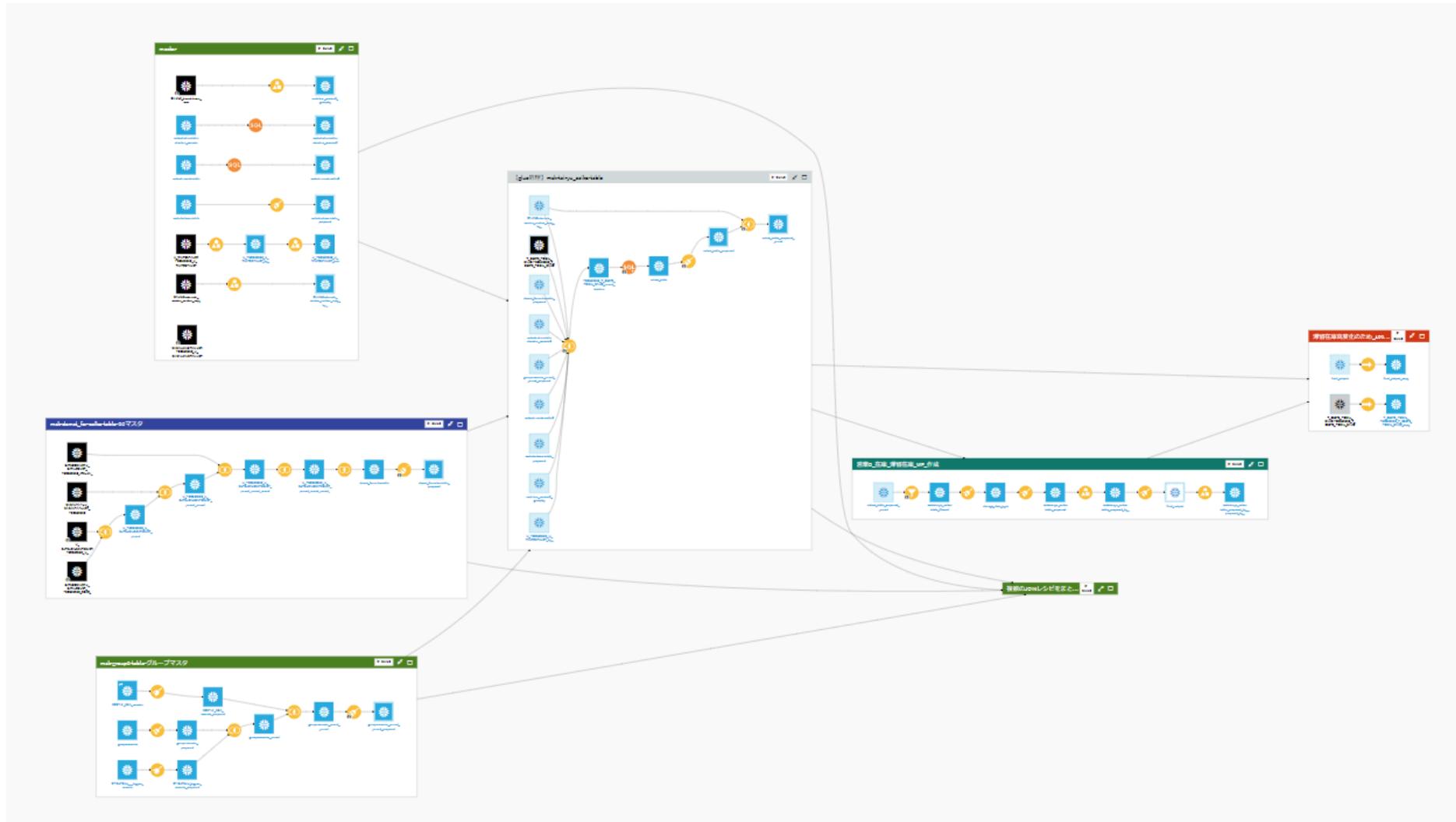
なぜ整理術を身につけることは重要か？

DATAIKUの特徴





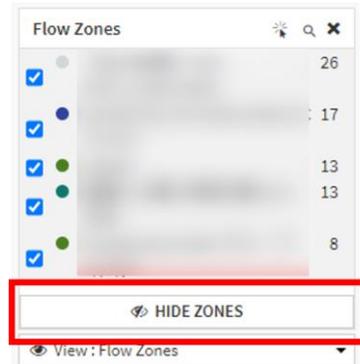
整理術その1 : ZONE



整理術その1：ZONE

ZONEを活用することで、フローの可読性を高められるのと、使用難易度を低くできる。

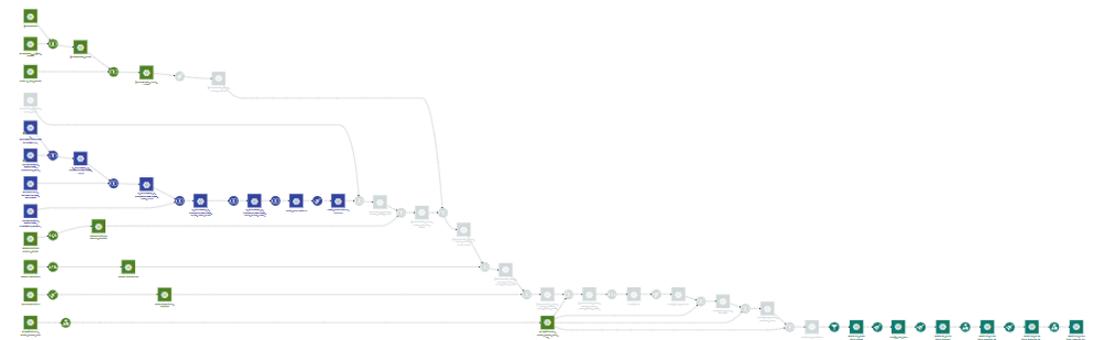
- 処理の群れをまとめる
→同じ目的を持つ処理を1つのゾーンに集約
- 特定のZONEのみ実行する
→目的によって指定されたZONEのみ実行することもできる
(シナリオ設定も)
- ZONEの色
→色に意味を持たせよう。
- ZONEビューの表示/非表示
→ZONEビューを非表示にすることもできる



ZONEを表示：



ZONEを非表示：

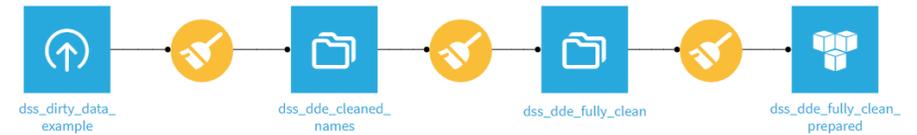


ZONE : どの粒度でZONEで分けるか

○同じ目的を持つ処理を1つのゾーンにまとめる



×細かすぎる



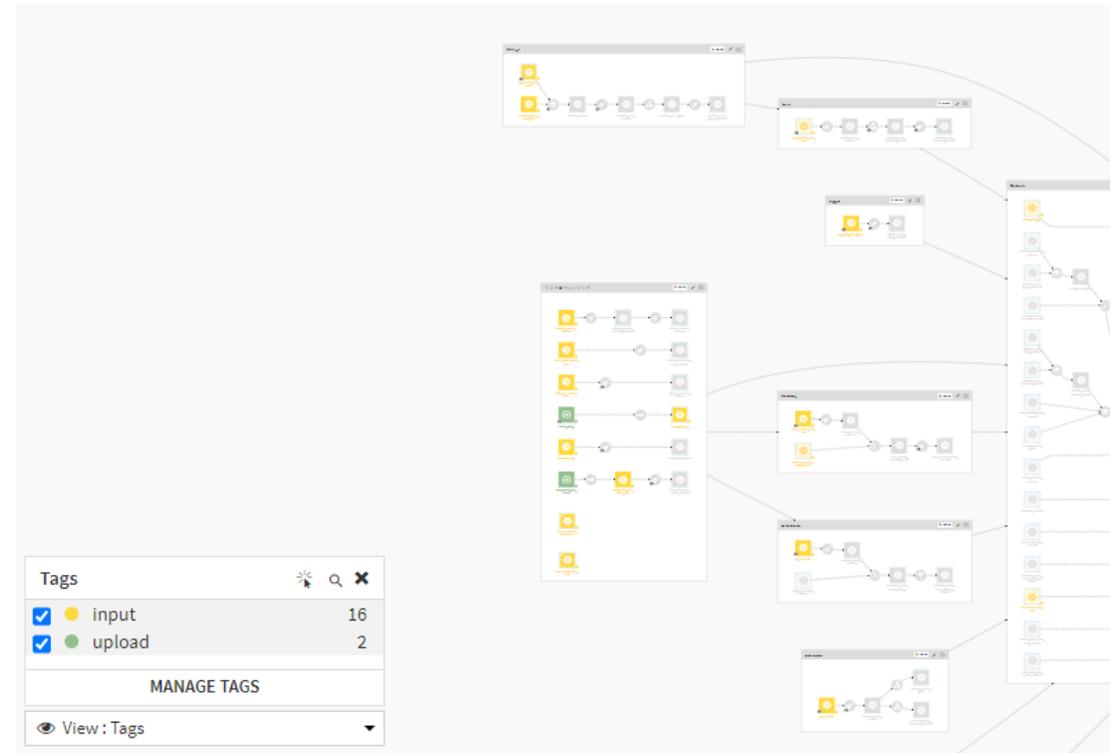
整理術その2：タグ

ZONEと組み合わせて使用することで、フローの可読性をワンランクアップ

よく使うタグ：

Input/Output/test(開発)/マスタ

- データセットやレシピを分類
→検索しやすくするため
- タグビューの表示/非表示
→タグを俯瞰



ZONEとタグの使い分け

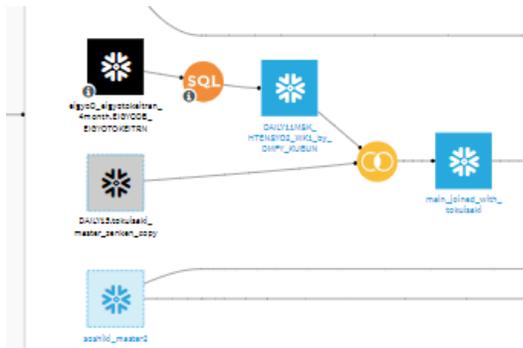
	ZONE	タグ
目的	同じ意味を持つレシピやデータセットの「まとめり」を作る	後から検索しやすくするために、特殊なレシピやデータセットをマーク
レシピやデータセットの数	多め	少なめ
使用頻度	高い	必要に応じて
その他	ZONEにタグをつけることは可能	

整理術その3 : Share Dataset

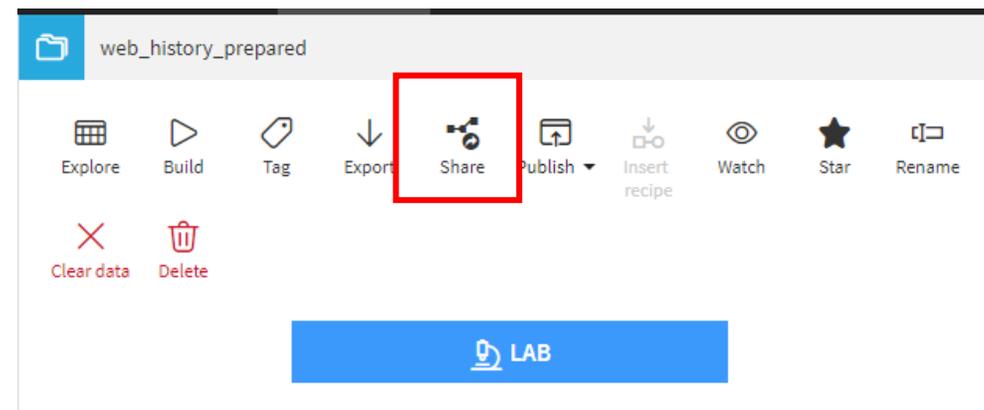
別プロジェクトで作られたデータセットを引用する場合、Share Dataset機能でデータセットをそのまま参照することがおすすめ。

- ✓ Inputのデータセットが、どのフローによって作られたか追うのにとっても便利になる。
- ✓ Data QualityやChartsなどの設定も継承されるので、データセットが管理しやすくなる。

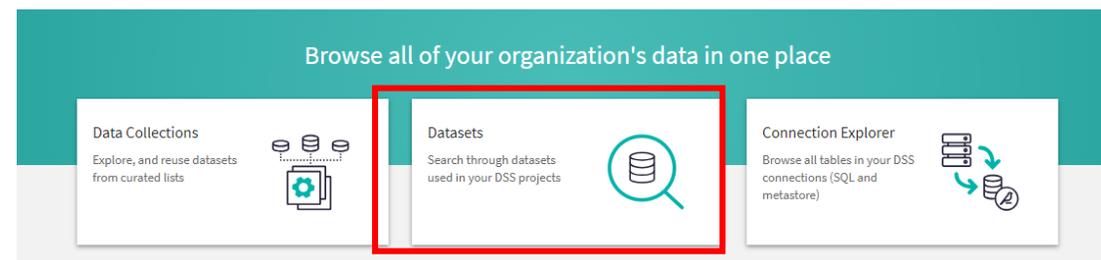
※シェアされたデータセットは下記のように黒いアイコンで表示される。



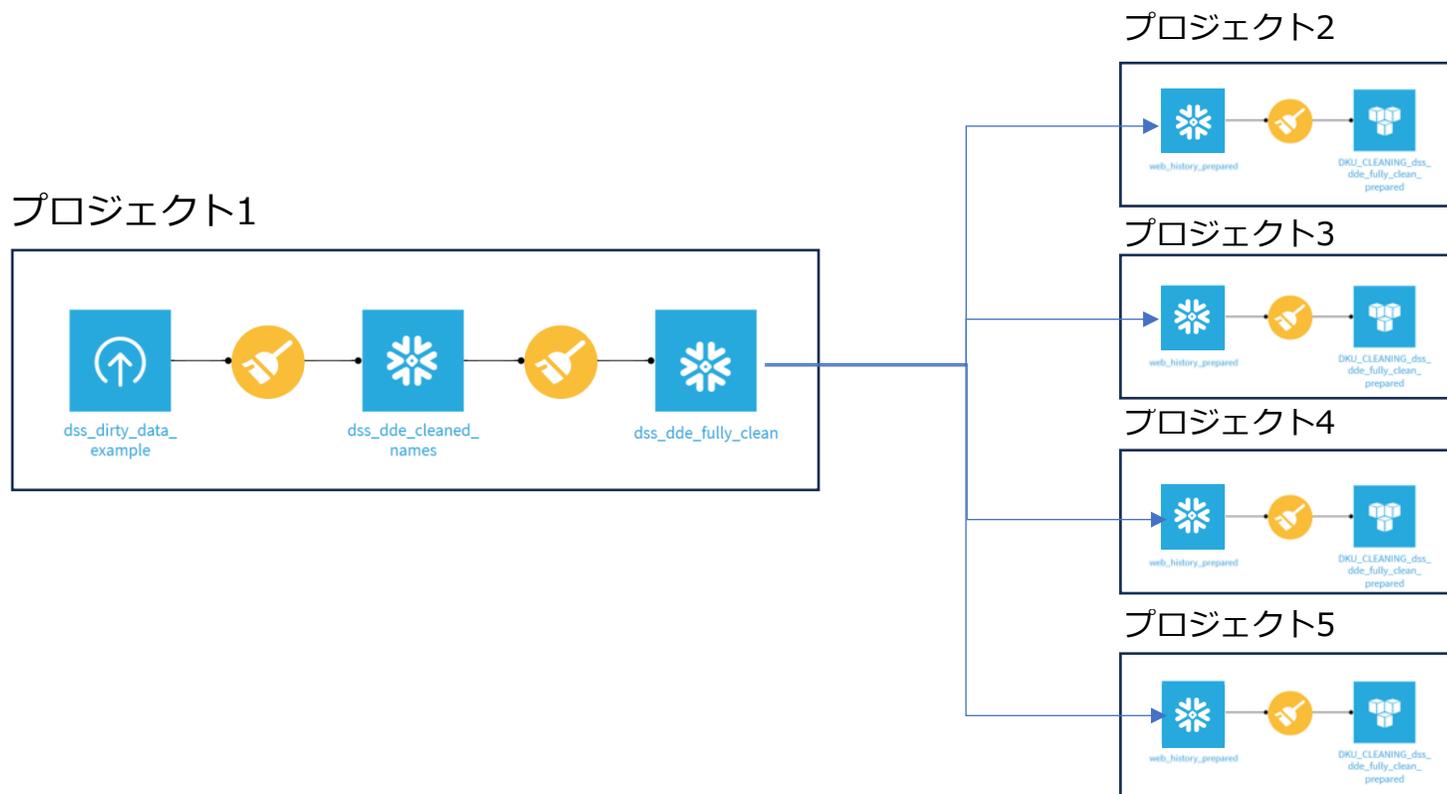
データセットをシェアする方法①



データセットをシェアする方法②



整理術その3 : Share Dataset



Share Dataset

Share Datasetを使うメリット：

- Data QualityやChartsなどの設定が継承される
- 別プロジェクトによって作られたデータであることがすぐわかる
ソースプロジェクトへの移動も便利
- サンプルの更新は一回行えばすべて更新される
- (データセットの格納先がデータベースの場合)テーブルの差し替えは楽に済む

整理術その4：注釈(フロー)

あらゆる処理に注釈をつけよう。注釈をつける理由はたくさんある。

- ・理解を助ける
- ・メンテナンスが容易になる
- ・将来の自分を助ける

など

Dataikuでは、いろいろなところに注釈をつけることができる。

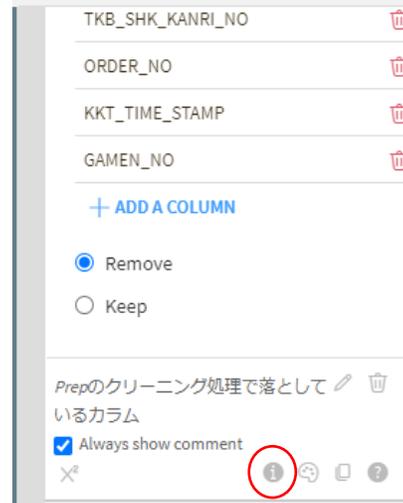
- ・データセットやレシピ
- ・ZONE
- ・PrepareレシピのStep

データセットやレシピ、ZONEは右側のサイドバーから注釈を付けられる



整理術その4：注釈(Prepareレシピ)

①をクリックして、注釈を記入することができる



Prepareレシピの補足

同じ目的を持つ処理群をグループにまとめる

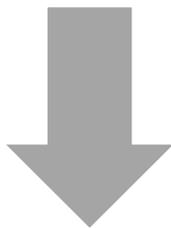
散乱した処理→



↑グループ化された処理群

整理術その4：注釈

注釈は書きたければいくらでも書ける
どのくらい注釈をつければよいか？



Detailsを見るだけでレシピの概要がわかるように

→ Main_Cleansing

Details

About EDIT

Click to add tags

①日付カラムの作成

②特徴量を作成

Creation 2 months ago by S

Last modification just now by S

Watched by 1 user

Starred by 0 users

Recipe type data preparation

Last successful build Thursday, 29 August 2024 at 15:39

Last successful build duration about 13 seconds

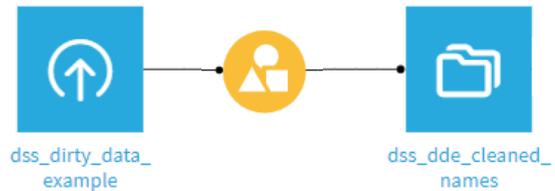
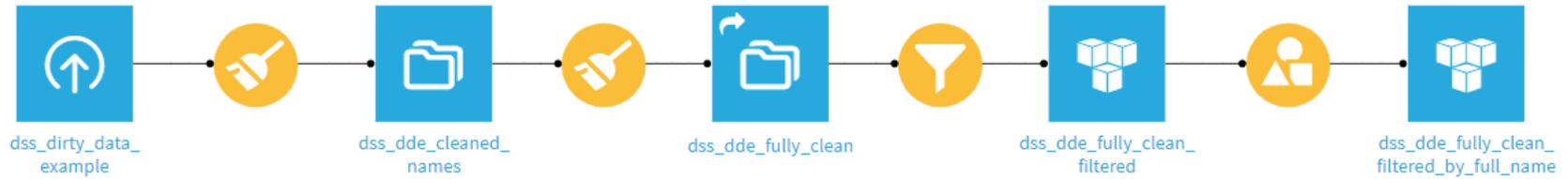
Summary

Parse date in 計上日

特徴量カラム作成 (11 steps)

- Create column Quarter with formula `datePart(val("計上日"),"quarter")`
- Create column Day Of Week with formula
- Create column Month Of Year with formula `datePart(val("計上日"),"months")`
- Create column 月初/月末flag with formula `if(val("Day Of Month")==1,1,0)`
- Create column Week Of Month with formula
- Create column Day Of Month with formula `datePart(val("計上日"),"days")`
- Create column 月初/月末/日付 with formula
- Create column Week No. with formula
- Create column 気温段階 with formula
- Create column 土日flag with formula
- Create column year with formula `datePart(val("計上日"),"year")`

整理術その5：レシピの数を減らしましょう



※prepareレシピの処理内容はカラム作成もしくはフィルターの場合はGroupレシピに代替できる

整理術その6：ネーミング



+ Add item(s) to build step

Dataset Folder Model Evaluation store Knowledge bank Flow zone

Dataset Nothing selected

Filter...

- karayouki
- karayouki_joined
- karayouki_joined_stacked
- karayouki_joined_stacked_joined
- karayouki_joined_stacked_joined_joined
- karayouki_joined_stacked_joined_joined_prepared
- karayouki_joined_stacked_joined_joined_prepared_stacked
- karayouki_joined_stacked_joined_joined_prepared_stacked_joined
- karayouki_joined_stacked_joined_joined_prepared_stacked_joined_joined

A screenshot of a software interface for adding items to a build step. The 'Dataset' category is selected, and a dropdown menu is open showing a list of dataset names. The names are variations of 'karayouki' with various suffixes like '_joined', '_stacked', and '_prepared'.

Replace recipe input

Replace dataset with: No dataset selected

Filter...

K

- karayouki
- karayouki_joined
- karayouki_joined_stacked
- karayouki_joined_stacked_joined
- karayouki_joined_stacked_joined_joined

A screenshot of a software interface for replacing a recipe input. The 'Replace dataset with:' dropdown is open, showing a list of dataset names. The names are variations of 'karayouki' with various suffixes like '_joined', '_stacked', and '_joined_joined'.

まとめ

整理術その1 : ZONE

整理術その2 : タグ

整理術その3 : Share Dataset

整理術その4 : 注釈

整理術その5 : レシピの数を減らしましょう

整理術その6 : ネイミング

フロー整理に関連した「あるといいな」

- ネストされたZONE
- アイコンの自由配置、もしくはZONEの自由配置
- データセットネーミングの日本語サポート
- データセットの選択のUIから行えるように

Home · Discussions · Product Ideas

Allow nested flow zones 🔖



info-rchitect

***** Posts: 180

September 2022

35

35 votes

In the Backlog · Last Updated September 2022

Hi,

I use flow zones a lot and appreciate the value. Why not extend the capability and allow nested flow zones, i.e. a flow zone within a flow zone?

thx

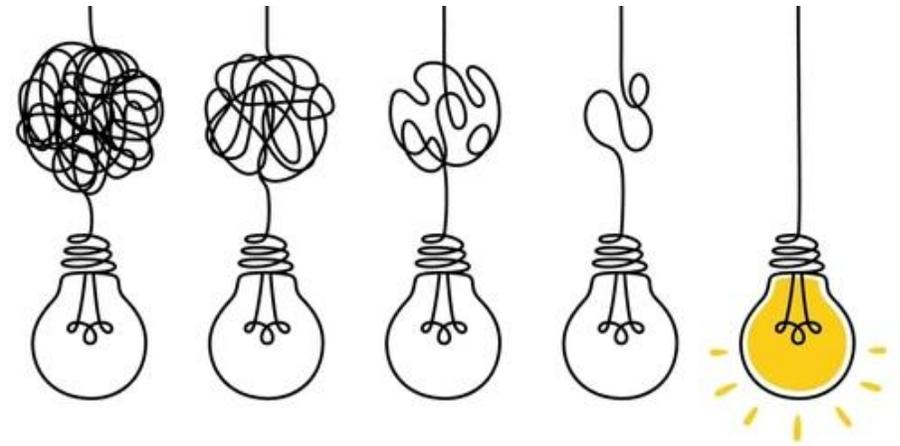
REPLY



最後に

「工、其の事を善くせんと欲すれば
必らず先ず其の器を利くす。」

---孔子の『論語』



<https://www.truestar.co.jp/>

Shibuya Center Place 8F | 1-16-3 Dogenzaka | Shibuya-ku | Tokyo 150-0043 | Japan
Tel: 03 5422 6561 | Fax: 03 5422 6562 | e-mail: info@truestar.co.jp