Dataikuフロー整理術

2024/10

アジェンダ

登壇者自己紹介・会社紹介

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」

自己紹介

名前：残間大地（ざんま　だいち）
／ d.zamma@truestar.co.jp

所属：株式会社truestar
データコンサルティング事業部
データサイエンス & アナリティクスグループ
（ディレクター）

スキル・経験値：
- テ　ー　マ：市場予測、新商品・既存品需要予測、効果測定、属性拡張、ターゲット予測、データ活用・可視化、KPI構造化、消費者セグメンテーション、クレーム分類、大規模言語モデル活用
- 手　　法：階層ベイズモデル、共分散構造分析、パス解析、時系列解析、機械学習、因果推論、LLM等
- 言語・ツール：Python, R, tableau, pytorch, Snowflake, Dataiku
- 業界 ・ 部門 ：
  - 食品、飲料・アルコール、化粧品、生保・損保、サービス、広告代理店、自動車、スポーツ用品、リテール
  - マーケティング、営業・営業企画、経営企画、調査部、広告部、人事、DXなど

直近の仕事：
需要予測、LLM活用PoC、効果測定・営業担当者の業務効率化・価格政策関連の顧客課題の企画・提案、育成・組織マネジメント

自己紹介

名前：スウ　シンテキ
／ s.zou@truestar.co.jp
所属：株式会社truestar
データコンサルティング事業部
データサイエンス & アナリティクスグループ

スキル：
DATAIKU・Alteryx・ Tableau

直近の仕事：
DATAIKUのETL機能をゴリゴリ使っております

会社概要

会社名

株式会社 truestar

所在地

東京都渋谷区道玄坂1-16-3 渋谷センタープレイス 8F

代表取締役

藤　俊久仁

従業員数

54名（2024年5月1日時点・非正規含む）

従業員平均年齢

36歳（2024年5月1日時点・非正規含む）

株主

truestar hd 株式会社（100%）

加盟団体

LBMA Japan (Location Based Marketing Association)

クライアント（一部抜粋）※

パートナー（一部抜粋）※

※ クライアント・パートナー企業の情報は、事前にご承諾いただいた場合のみ五十音順で掲示しています。

事業概要 | 主力サービス

データドリブンな意思決定を後方支援！

データコンサルティング事業

データ
アナリティクス
分析

データ
ビジュアライゼーション
可視化

データ
プレパレーション
前処理

データサイエンス

データマネジメント

データ基盤構築

パートナー連携で対応

Prepper Open Data Bank | 収集も加工も不要！すぐに使えるオープンデータ！

不毛なデータプレップ作業から貴重なデータ分析者の時間を解き放て！

無償公開
データ代は一切不要

商用利用可
ビジネス用途で安心して使える

一元管理
Snowflake Marketplace で共有

公開データ （2023年10月1日時点）

人口・世帯数
国勢調査(e-Stat)
将来推計人口(国土数値情報)

境界（ポリゴン）
国勢調査小地域(e-Stat)
行政区域(国土数値情報)

天気
過去実績・予報(気象庁)

出生数
人口動態調査(e-Stat)

医療
医療機関・医療圏(国土数値情報)
医療施設調査(e-Stat)

法人情報
法人活動情報(経産省 gBizINFO)
従業員数(厚労省 日本年金機構)
証券コード(金融庁 EDINET)

人流
1kmメッシュ人流(G空間情報)

鉄道
駅・路線(国土数値情報)

カレンダー
祝日(内閣府)

地価・土地
地価公示(国土数値情報)
住宅・土地統計調査(e-Stat)

経済・産業
経済センサス(e-Stat)
工業統計調査(e-Stat)

https://podb.truestar.co.jp/

※商用・二次利用可能なデータのみ取り扱う

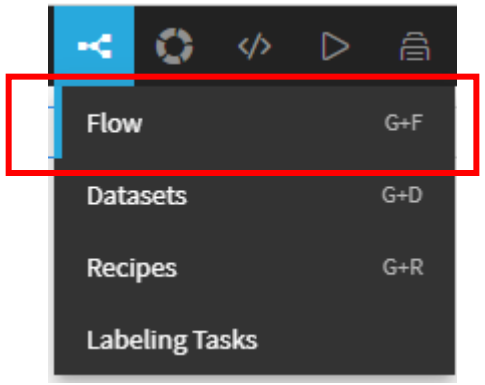
弊社の直近の取り組み(Dataiku案件)

✓ **<u>工場設備部品の交換時期予測検知</u>**
お客様の設備モニタリングデータからDataiku機械学習を使い、部品交換時期の予測検知
設備劣化の予知により工場保全制度の向上を実現

✓ **<u>自社 Prepper Open Data Bank(PODB)サービス構築</u>**
- 弊社サービスにおいて、外部の各種オープンを成形・加工し「すぐ使えるデータ」として提供
- 運用コストダウンやパフォーマンス向上、さらに今後のAI利用を鑑み、既存ETLシステム（Alteryx）からDataikuへ移行しサービス運営中

✓ **<u>Dataiku×SNOWFLAKEの環境構築</u>**
既存ETLツールで構築していたフロー群をDataiku×SNOWFLAKEの環境へ移行

直近の仕事

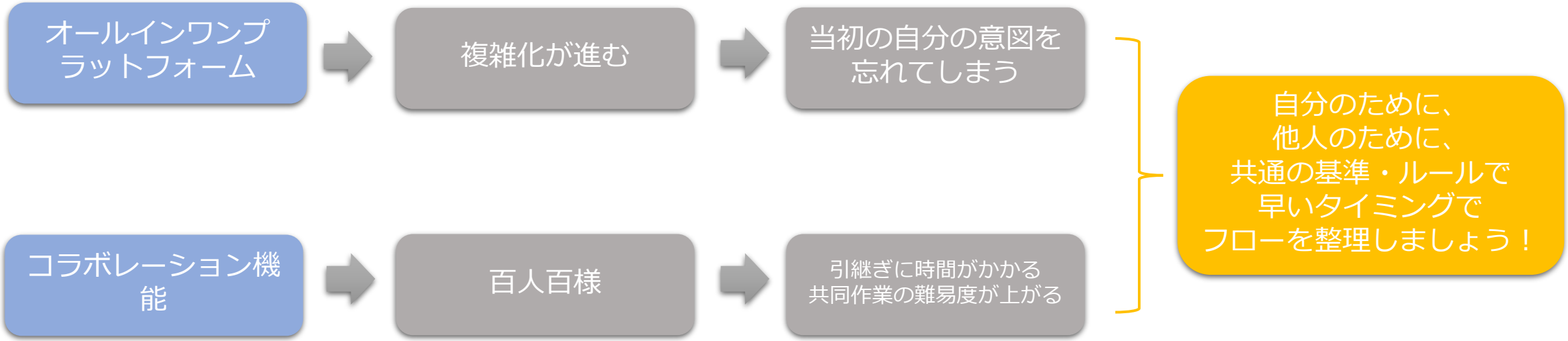


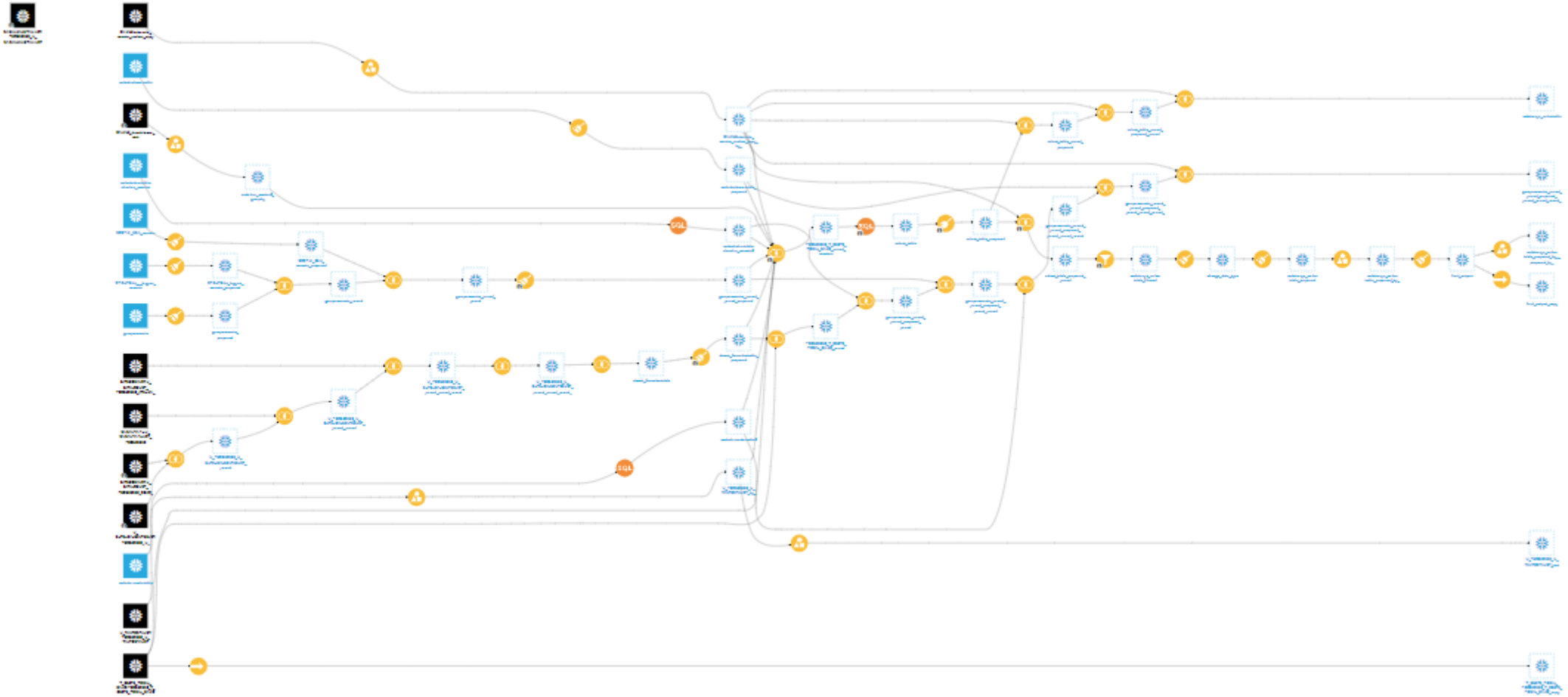
80%の作業はこちらの画面で行う

Dataikuフロー整理術

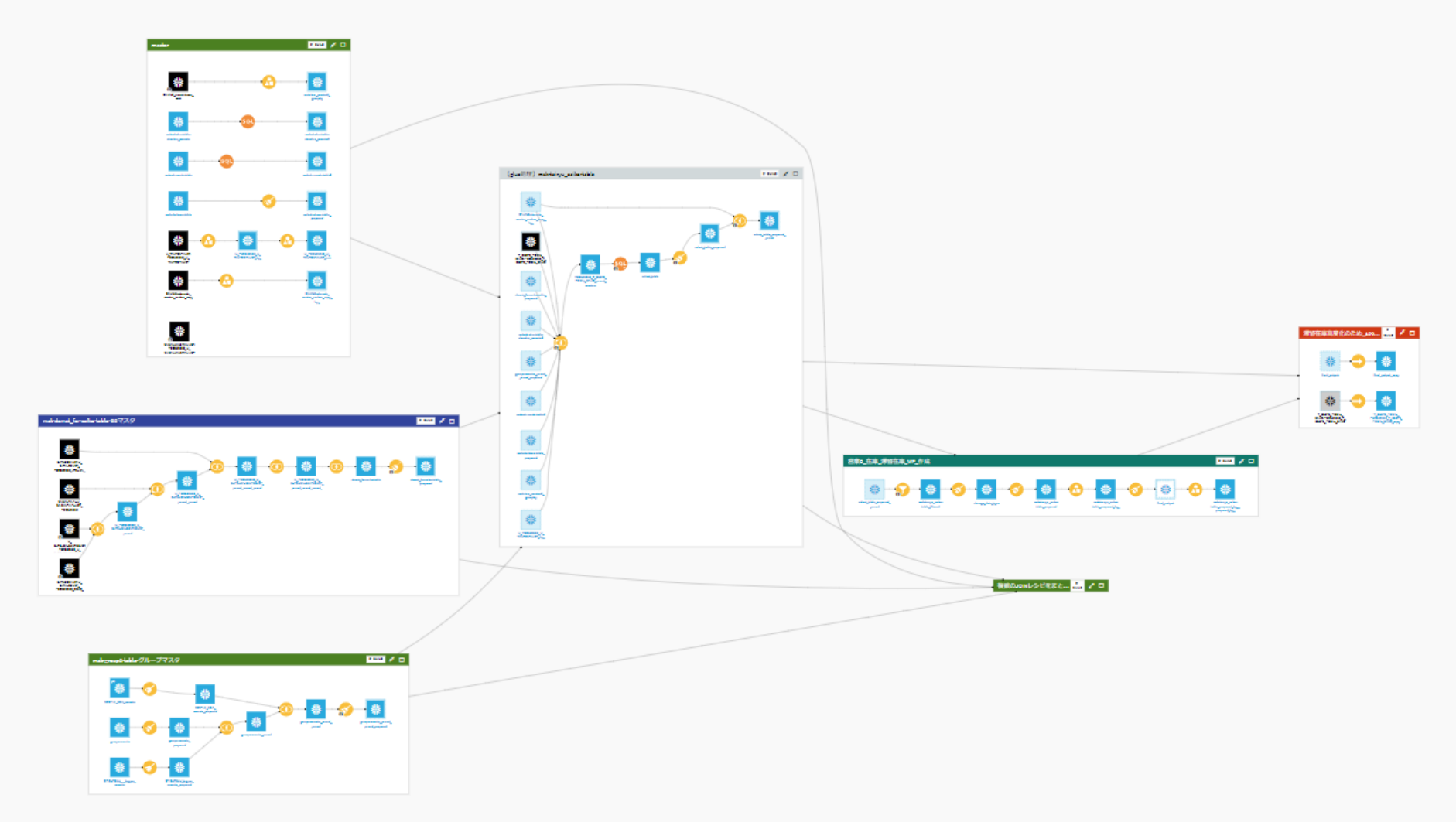
なぜ整理術を身につけることは重要か？

DATAIKUの特徴





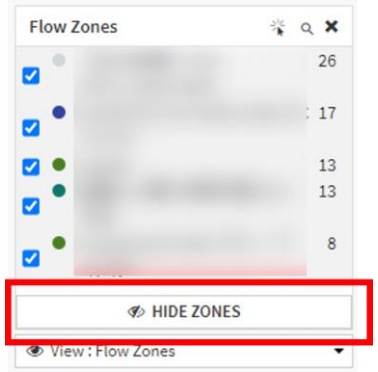
整理術その1：ZONE



整理術その1：ZONE

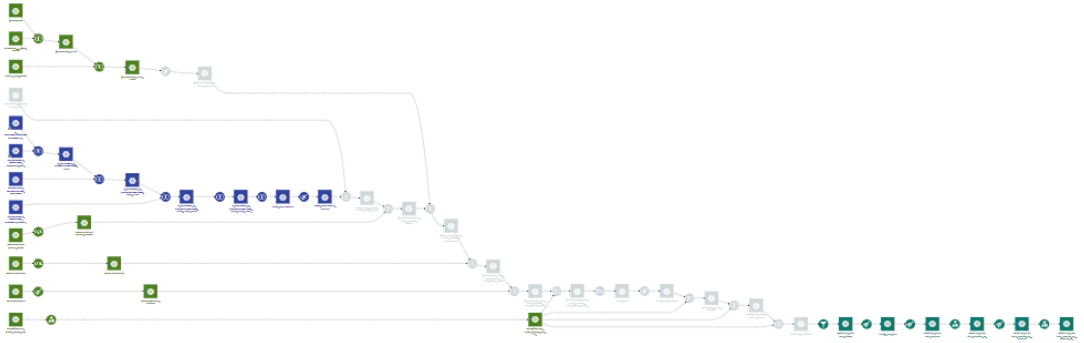
ZONEを活用することで、フローの可読性を高められるのと、使用難易度を低くできる。

・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：

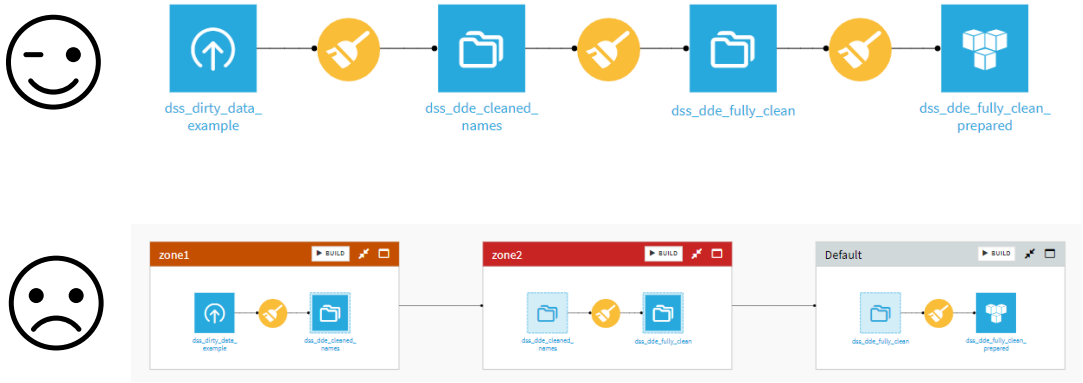
ZONEを非表示：



ZONE：どの粒度でZONEで分けるか

〇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



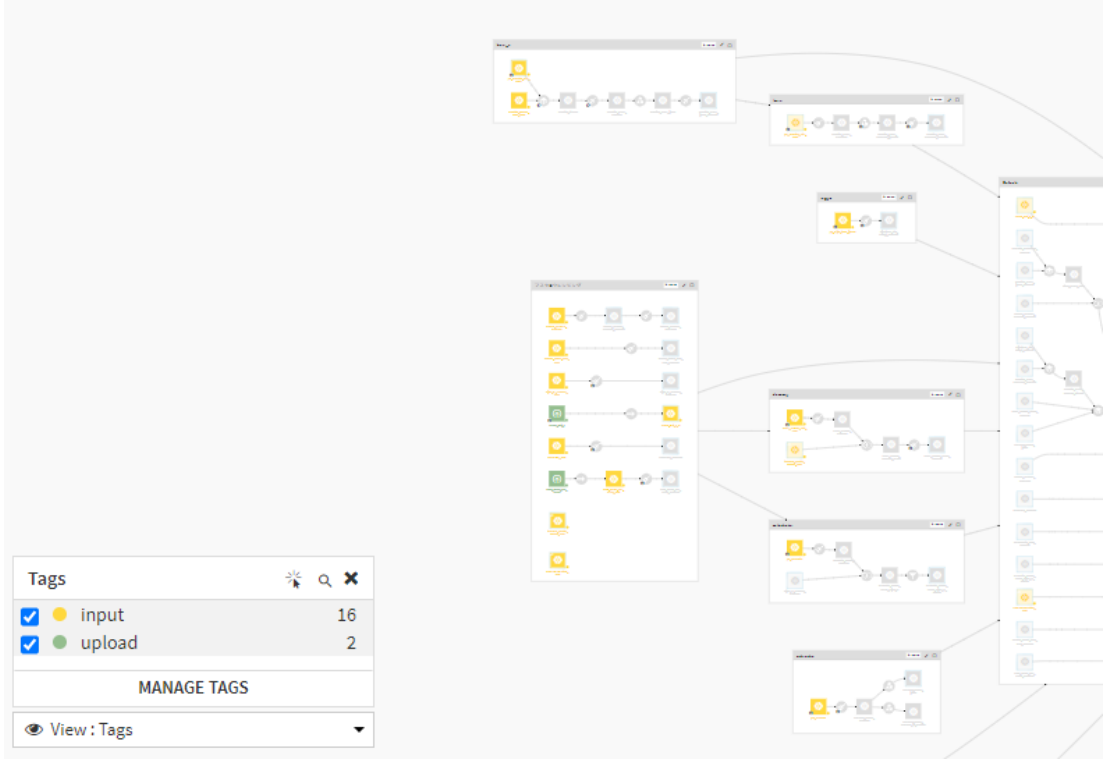
整理術その2：タグ

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

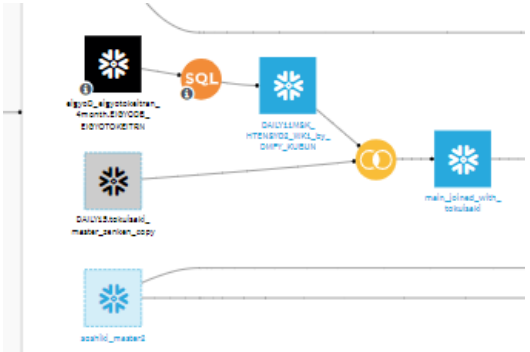
	ZONE	タグ
目的	同じ意味を持つレシピやデータセットの「まとまり」を作る	後から検索しやすくするために、特殊なレシピやデータセットをマーク
レシピやデータセットの数	多め	少なめ
使用頻度	高い	必要に応じて
その他	ZONEにタグをつけることは可能	

整理術その3： Share Dataset

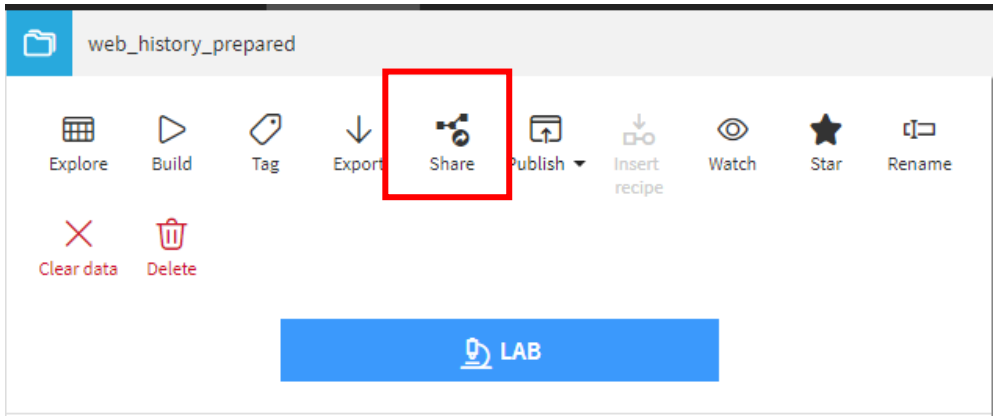
別プロジェクトで作られたデータセットを引用する場合、
Share Dataset機能でデータセットをそのまま参照することがおすすめ。

- ✓ 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。

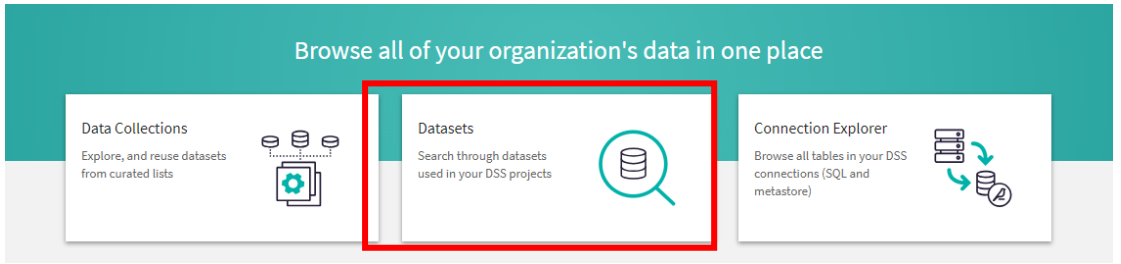
※シェアされたデータセットは下記のように黒いアイコンで表示される。



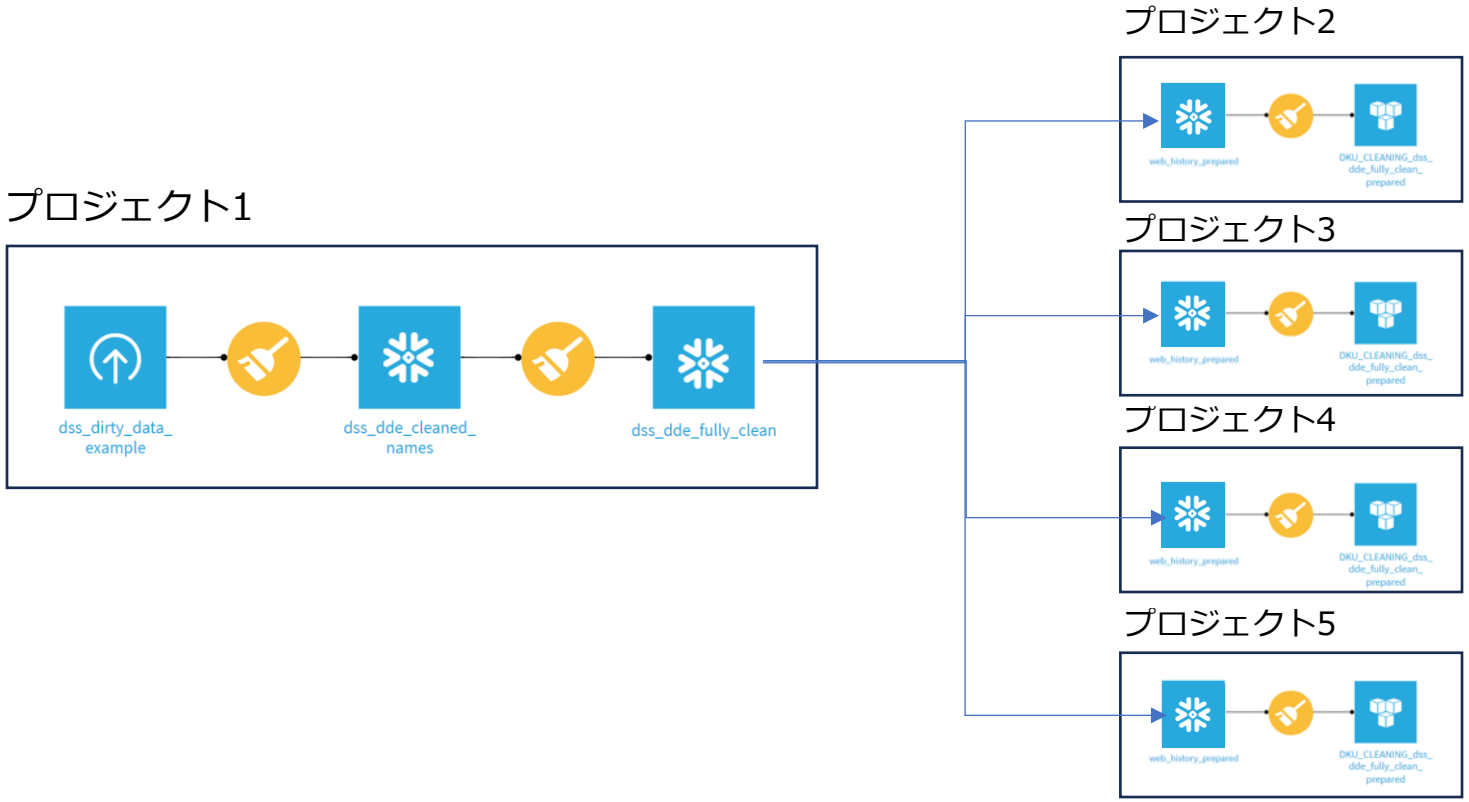
データセットをシェアする方法①



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)

ⓘ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

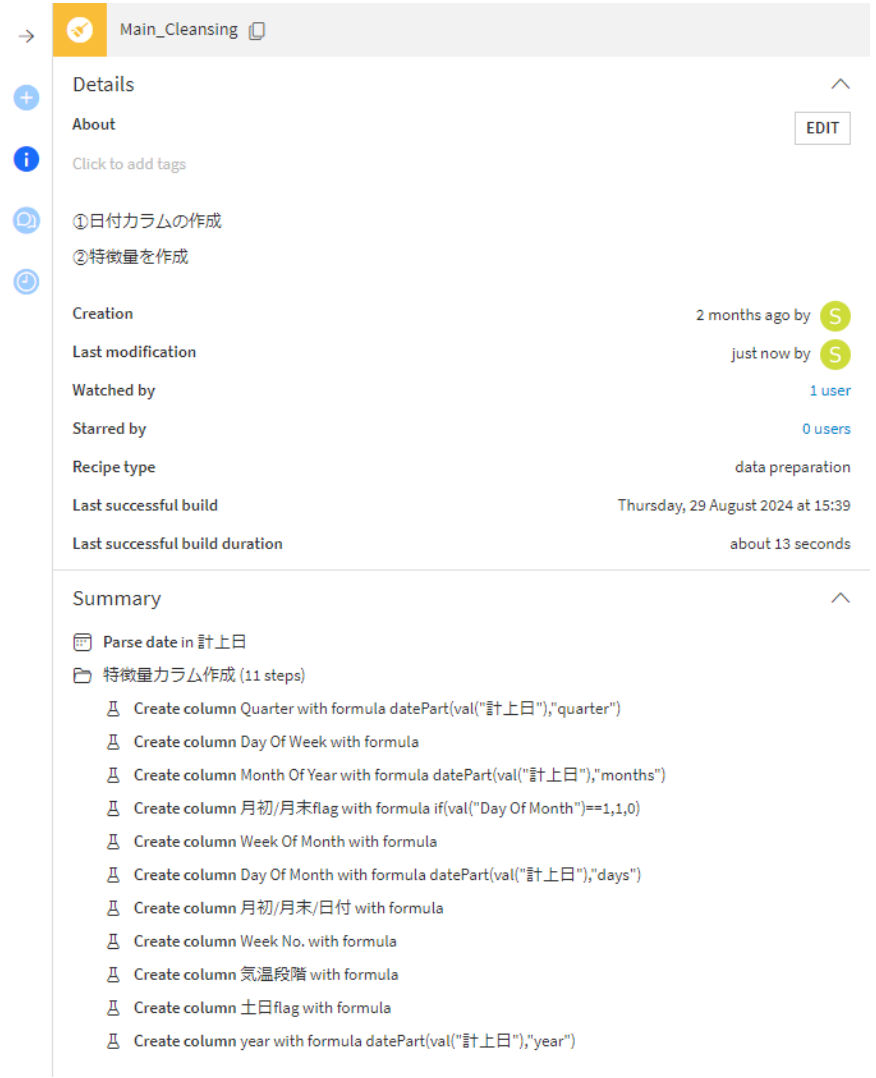
散乱した処理→

↑グループ化された処理群

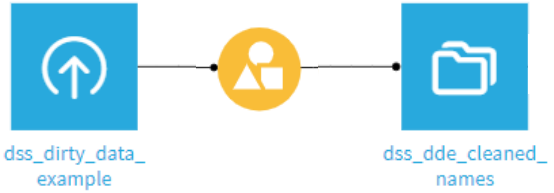
整理術その4：注釈

注釈は書きたければいくらでも書ける
どのぐ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

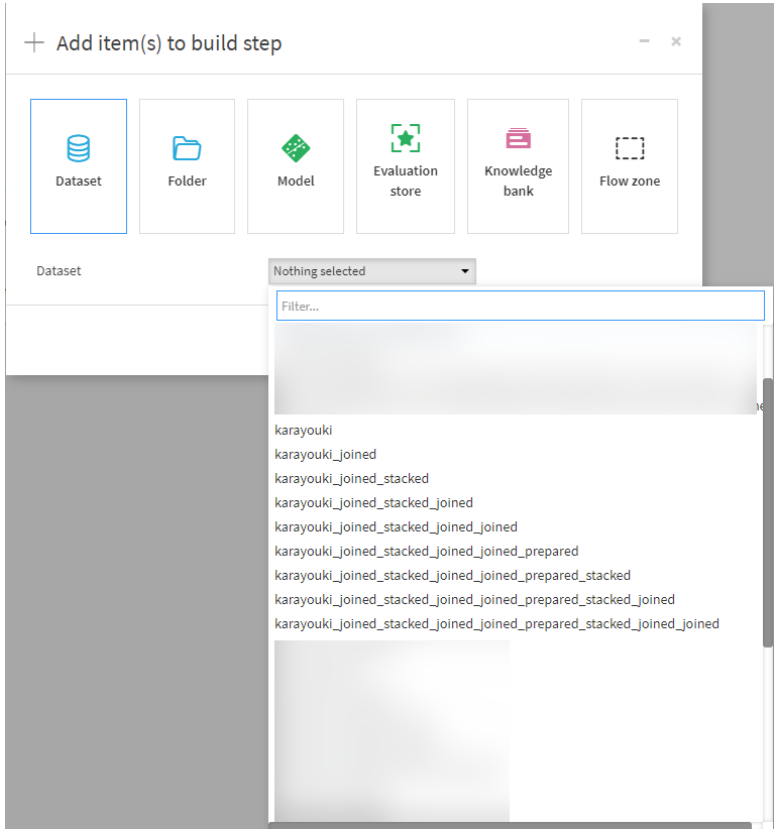


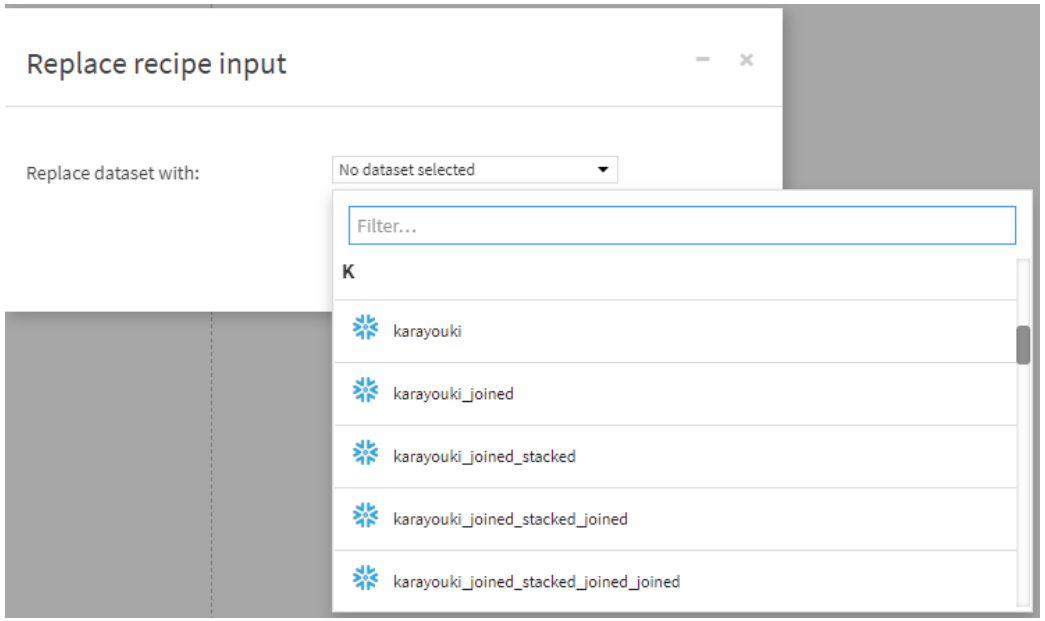
整理術その5：レシピの数を減らしましょう



※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：ネイミング





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

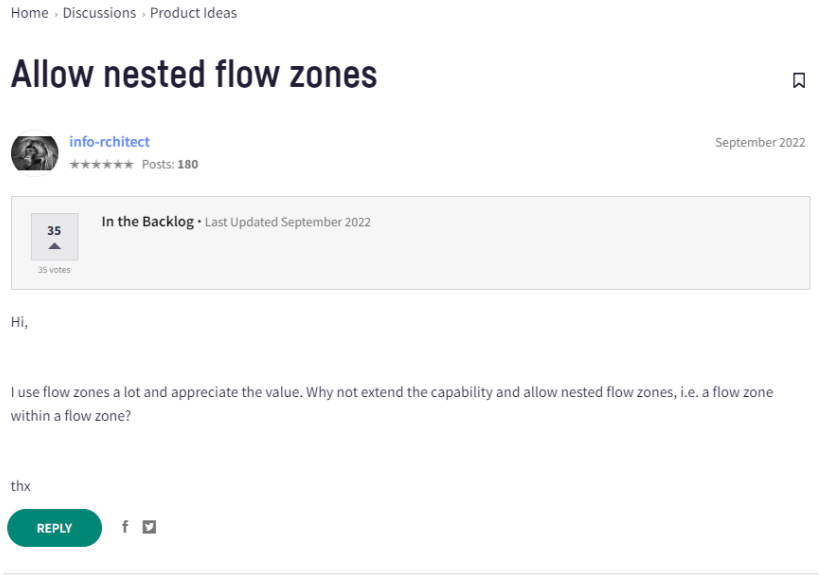
フロー整理に関連した「あるといいな」

・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

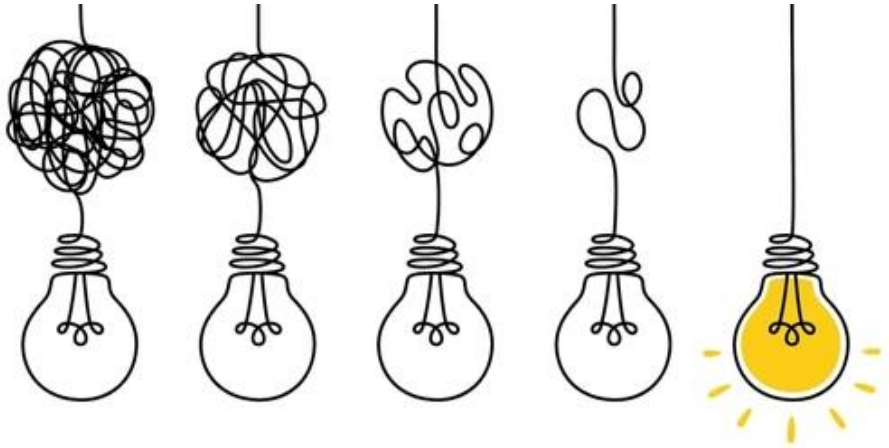
・データセットの選択のUIから行えるように



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