Un projet ML de bout-en-bout avec Snowflake & Dataiku

Octo Technology - Groupe Utilisateurs Dataiku Paris

MEHDI MOULOUDJ

05/10/2021

Sommaire

01

Pourquoi migrer ?

Introduction à Dataiku

L'entreprise

- **Entreprise française** spécialisée dans la data science
- Commercialise **DSS**, une plateforme de Data Science
- Fondée en 2013, basée à New York
- Centre de R&D à Paris

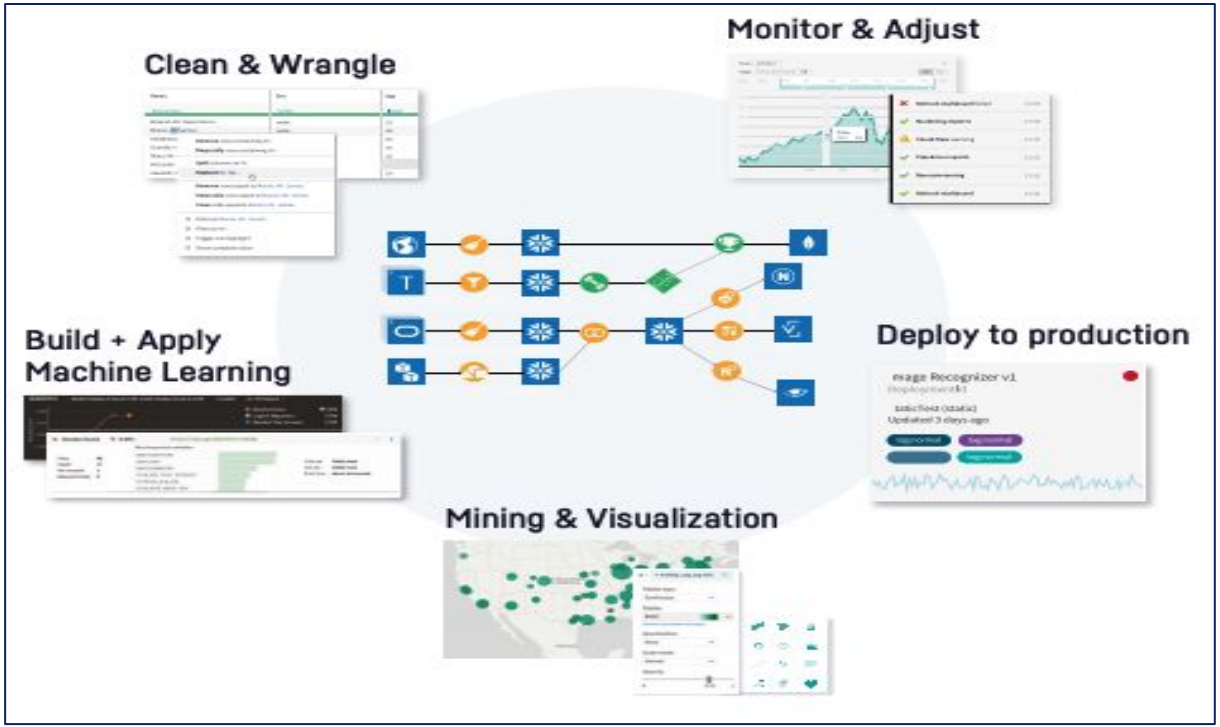
Quelques chiffres

- **400+ clients** (OTAN, Pfizer, OVH, Pricemoov')
- **650 employés** dans le monde dont 250 en France
- **Valorisée à $4.6 Mds** avec plus de **$700M** de levée de fonds

Description synthétique

- Data Science Studio (DSS), **pour toutes les phases du cycle de vie d'un projet de Machine Learning**
- Installé **"on-premise"** ou dans des **environnements cloud**
- **Mettre la data science à portée de tous**
- Accent mis sur la collaboration de **Data Engineer, Data Scientist, Citizen Data Scientist et Business Analyst**

Fonctionnalités clés de DSS



- Préparation visuelle des données
- Data visualisation
- Modélisation et Machine Learning visuel
- Analyse des modèles, reporting
- Déploiement et monitoring des modèles
- Travail collaboratif
- Délégation du “compute”

Introduction à Snowflake

L'entreprise

- Entreprise américaine spécialisée dans le **stockage et l'analyse de données dans le Cloud**
- Commercialise un entrepôt de données **(Data Warehouse) as-a-service**
- Fondée en **2012**

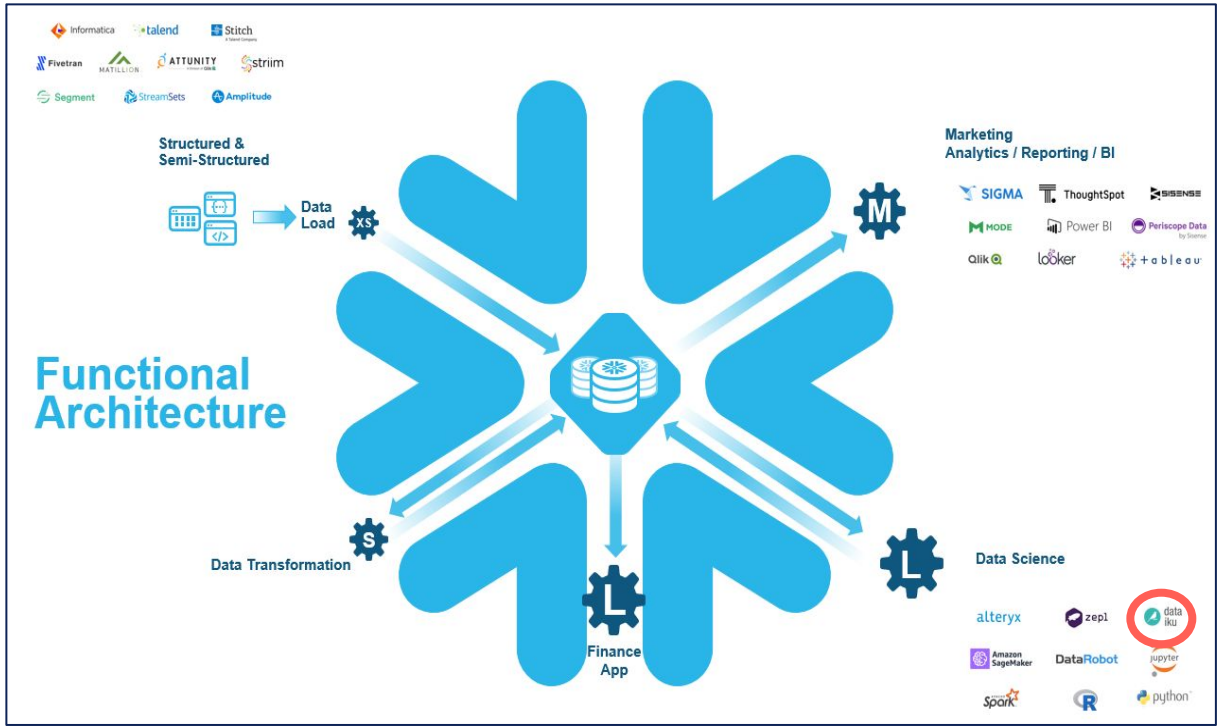
Quelques chiffres

- Plus de **4900 clients** dans le monde entier (Rakuten, Deliveroo, Emirates, F.F.F...)
- +1 Mds$ de CA, 3500 employés
- Cotée au Nasdaq depuis 2020

Description synthétique

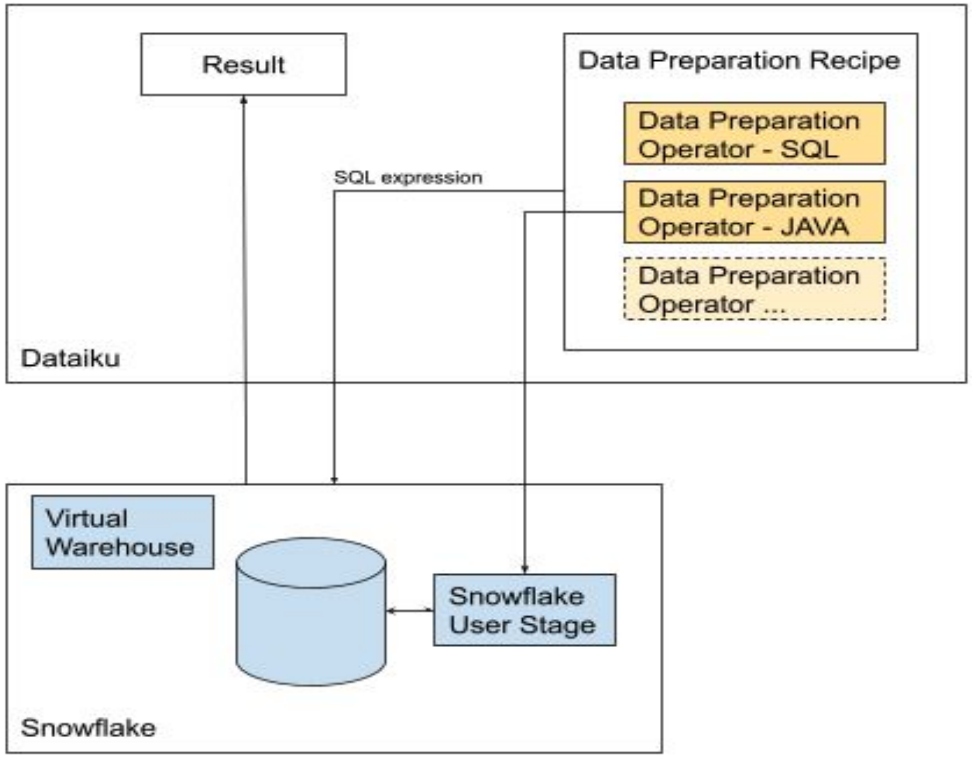
- Consolidation des données en une **source unique de vérité (single source of truth)** qui alimente des applications analytiques et de data science
- Architecture **Multi-cluster Shared-data, scalable** pour des accès aux mêmes données **indépendamment et sans conflit**
- Pour une stratégie data-driven **sans investir dans du hardware** rapidement obsolète, s**ans besoin de compétence ni de temps** pour la configuration et la maintenance et quasi-infiniment scalable

Fonctionnalités clés de Snowflake



- **Centralisation** des données
- **Scalabilité** de stockage et computationnelle
- **Élasticité dynamique des clusters** dynamique
- **"Pay what you consume"**
- **Séparation** nette entre ressources de **stockage et computationnelles**
- **Optimisation automatisée** des requêtes, des partitionnements de tables
- **Écosystème de partenaires** en amont et en aval

Snowflake x Dataiku : le mariage parfait ?



- **Un partenariat fort**
- Exécution de certaines recettes **de préparation, transformations de DSS sur** un warehouse Snowflake
- **Inférence en batch** de modèle prédictif **peut être déléguée à Snowflake** via ses JAVA UDFs
- **Scalabilité et contrôle des coûts :**
  - Dimensionner/Redimensionner instantanément des ressources de calcul
  - Ne payer que pour les secondes où les clusters sont réellement utilisés
- Néanmoins compliqué de **monitorer la consommation Snowflake depuis DSS**
- **Des limites sur les recettes qui peuvent être déléguées**
  → Qu'en sera-t-il avec **Snowpark** ?

02

Présentation du cas d'usage

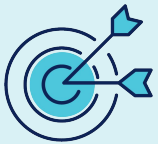
Challenge Planète-Oui : Utilisation des bornes de recharge de véhicules électriques sur Paris

273 terminaux de recharge répartis sur **94 stations**

Pour chaque station et chaque terminal une mesure par quart d'heure entre le **25/11/2019** et le **08/11/2020**

Problématique :

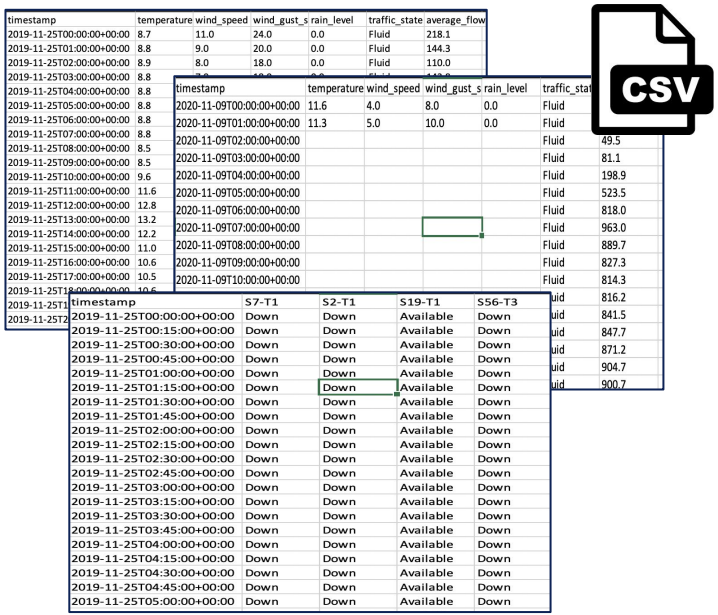
- Exploiter l'historique pour apprendre un modèle **supervisé** qui prédit l'état des bornes de recharge
- Récolter les données **fraîches** évaluer le modèle historique
- Mettre en place un **suivi de performance du modèle** et une logique de **réentrainement**



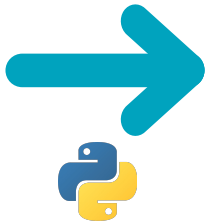
Les différents états :
Disponible
En charge
Passive
Hors Ligne
En panne

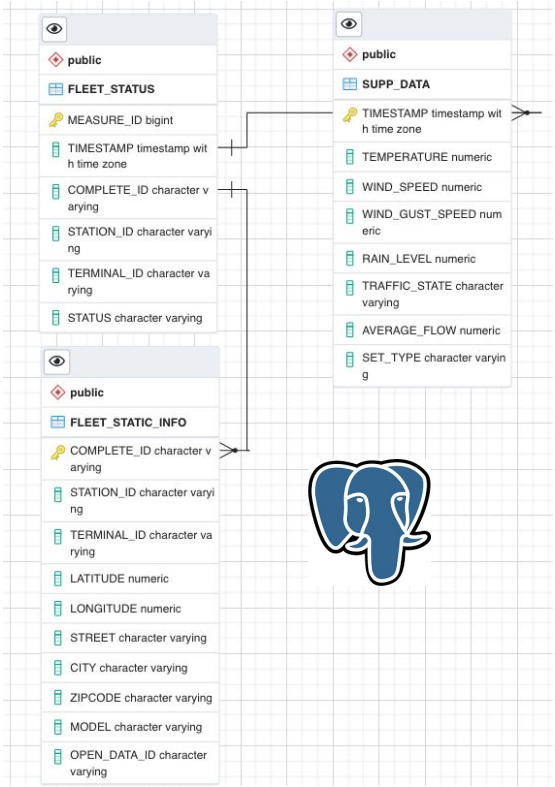
En plus des données d'**états des bornes**, de **météo, de trafic routier**, P-O fournit la **source de ces données...**

Transformation et chargement des données historiques sur PostgreSQL



Xtrain.csv, Xtest.csv,
Ytrain.csv, Static_info.csv

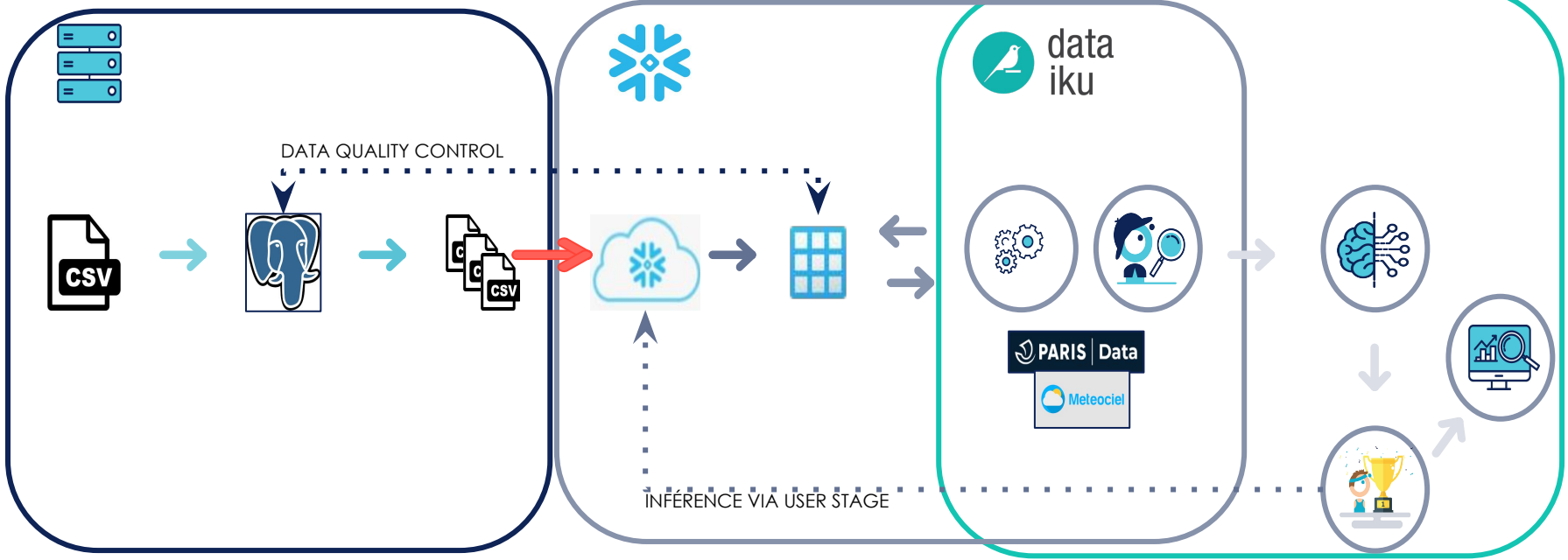




03

La migration

Architecture globale



TRANSFORMATION & IMPORT SUR POSTGRES	SPLIT ET EXPORT DEPUIS POSTGRES	STAGING	INGESTION DANS LES TABLES SNOWFLAKE DEPUIS LE STAGE	PRÉ-PROCESSING, EXPLORATION, RÉCOLTE DE NOUVELLES DONNÉES	ENTRAÎNEMENT, DÉPLOIEMENT DU MODÈLE CHAMPION, REPORTING ET MONITORING

Résultats et recommandations

Staging

- Extraire les données depuis Postgres et les stocker dans des fichiers
- Organiser **logiquement** ses fichiers (Temporel, Sous-catégories…) **10 < size < 100 Mb (Compressed)**
- Choisir le type de **STAGE**
- Nommer le stage, définir le format d'import, et configurer les options
- Installer et Configurer **SnowSQL**
- Utiliser la commande **PUT** sur **SnowSQL** pour uploader depuis le disque vers le stage créé

Loading

- Créer les tables **transitoires** et les tables **finales**
- Utiliser la commande **COPY INTO** pour alimenter les tables, définir la stratégie de gestion des erreurs, de nettoyage du stage
- Réaliser des checks sur les métadonnées
- **Réitérer** la migration jusqu'à la table **transitoire** et utiliser la commande **UPDATE** pour insérer les éventuelles lignes manquantes

	Complet & Compressed	Complet & Uncompressed	Half & Compressed	By Month & Compressed
Nb de fichiers	1	1	2	13
Taille moyenne (Mo)	50,9	456	25,5	4,0
Taille moyenne (MB)	407,2	3 654	204	32
Temps de stage 1 fichier (sec.)	3,6	46,2	1	**0,5**
Temps de stage TOTAL (sec.)	48,6	46,2	**2**	9,3
Temps de load TOTAL (sec.)	16,8	17,3	11	**7,9**

04

Préparation & exploration des données (démo)

Préparation et exploration des données

1

Analyse Données brutes

S'assurer de la qualité des données. Un constat : il manque des timestamps

2

Resampling

Technique d'ajout des timestamps manquants. Complété par "previous_value"

3

Jointures

Jointures entre les différentes tables nettoyées, complétées. Le dataset résultant sera la base de l'EDA

4

Analyse exploratoire

Analyse univariés des variables, bivariées en fonction de la target (le statut des bornes)

5

Feature engineering

Création de variables discriminantes pour la classification (cos, sin de l'heure de la journée...). On fait même du target engineering

05

Modélisation sur Dataiku (démo)

Retour sur les use cases

1 Point de vue utilisateur

L'utilisateur souhaite savoir quelle borne sera disponible à une **échéance proche** dans le futur :

> Classification multi-classes

> Classification binaire (Dispo/Pas dispo)

Approches

- Forecasting récursif ou direct
- **Un modèle par échéance**

2 Point de vue fournisseur d'énergie

Le fournisseur d'énergie souhaite savoir quelle sera la consommation de chaque borne à une **échéance plus longue** dans le futur :

> Classification binaire (Charge/Non Charge)

Approches

- Forecasting récursif ou direct

3 Point de vue gestionnaire du parc

Le gestionnaire du parc souhaite faire de la **maintenance prédictive** :

> Classification binaire (En panne/Pas en panne)

Approches

- Forecasting récursif ou direct

Approches pour le point de vue utilisateur

Modèle forecasting récursif

Données annotées

Données à prédire

timestamp	complete_id	…	status_previous	status
20/06/2020 00:00	S1T1	…	Available, Offline	Available
20/06/2020 00:00	S1T2	…	Charging, Offline	Charging
20/06/2020 00:00	S1T3	…	Available, Available	Available
20/06/2020 00:15	S1T1	…	Offline, Available	Available
20/06/2020 00:15	S1T2	…	Offline, Charging	Charging
20/06/2020 00:15	S1T3	…	Available, Available	Available
20/06/2020 00:30	S1T1	…	Available, Available	Offline
20/06/2020 00:30	S1T2	…	Available, Charging	
20/06/2020 00:30	S1T3	…	Available, Available	
20/06/2020 00:45	S1T1	…	????	
20/06/2020 00:45	S1T2	…	????	
20/06/2020 00:45	S1T3	…	????	
20/06/2020 01:00	S1T1	…		
20/06/2020 01:00	S1T2	…		
20/06/2020 01:00	S1T3	…		

Modèle supervisé H+24 (Q+96)

Données historiques

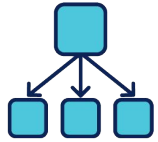
Données à prédire

timestamp	complete_id	…	status_previous	status	status+ 24h
20/06/2020 00:00	S1T1	…	Available, Offline	Available	Down
20/06/2020 00:00	S1T2	…	Charging, Offline	Charging	Available
…	…	…	…	…	…
20/06/2020 23:45	S1T2	…	Offline, Charging	Charging	Charging
20/06/2020 23:45	S1T3	…	Available, Available	Available	Available
21/06/2020 00:00	S1T1	…	Offline, Offline	Offline	
21/06/2020 00:30	S1T2	…	Available, Charging	Charging	
21/06/2020 00:30	S1T3	…	Available, Available	Offline	
21/06/2020 00:45	S1T1	…			
21/06/2020 00:45	S1T2	…			
21/06/2020 00:45	S1T3	…			
21/06/2020 01:00	S1T1	…			
21/06/2020 01:00	S1T2	…			
21/06/2020 01:00	S1T3	…			

06

Mise en production (démo)

Récoltes de données “nouvelles”



	1 État des bornes	2 État du trafic	3 Météo
Disponibilité	• Temps réel • Non historisées	• Quotidienne (publiées vers 2H30) • Historisées	• Horaire • Historisées
Qualité	• Le statut “passive” n’existe pas/plus • 1200 terminaux contre 273 dans le JDD Planète-Oui	• Un état du trafic pour toute la ville mais des centaines de points de mesure → Clé d’agrégation potentiellement différente.	• Choix du point de mesure potentiellement différent de celui de Planète-Oui
Accessibilité	API	API	Pas d’API → Web Scraping
Lien	Paris Open Data	Paris Open Data	Meteociel

There is a better way

Scenarii Dataiku, Pipeline d'inférence et de monitoring

Step 1

Tous les ¼ d'heure, lancement du script de récupération des états des bornes

Jour J
00h

Step 2

Lancement du script de récupération des données météo

Historisation des données sur les états des bornes et nettoyage de la table daily

Jour J
23h50

Step 3

Récupération des données de trafic de la veille

Fin des étapes de pré processing stacking des données PRÉ-WINDOWING

Step 4

Split labelled/unlabelled

Évaluation sur les labelled

Prédiction sur les unlabelled

Jour J +1
6h

Step 5

Création des rapports sur l' évaluation

Création des visuels sur la prédiction

Trigger de réentrainement

There is a better way

07

Bilan

Retour d'expérience sur l'utilisation des plateformes

Les +

- La simplicité, l'élasticité et les performances de Snowflake
- L'intégration très simple entre les deux plateformes
- La préparation des données avec DSS
- La bonne intégration dans DSS des outils open-source

Les -

- La gestion de la mémoire machine lors des process in-memory sur DSS
- Manque d'information sur la consommation de crédits Snowflake depuis Dataiku