codein.
Digital addict

Business Case

OPTIMISATION DES FLUX DE DONNÉES :
AUTOMATISATION ET SYNCHRONISATION DES SYSTÈMES AVEC APACHE AIRFLOW ET DBT

Business Case rédigé par

ROMAIN BONNAL

Responsable du pôle Data
Directeur Associé

codein.
Digital addict

#1

Contexte

CONTEXTE


© Crédit photo : Sandaya

Sandaya, acteur majeur dans le domaine de l'hôtellerie de plein air, a entrepris la refonte de son **Property Management System (PMS)**, un système clé pour la gestion opérationnelle et administrative de ses établissements (réservations, inventaire, gestion des clients, etc.).

Dans ce cadre, il est impératif de redéfinir et de moderniser l'ensemble des flux de données qui relient le PMS actuel aux différents systèmes du système d'information (SI) : comptabilité, business intelligence, CRM, gestion des avis clients, et bien d'autres.

Actuellement, ces flux de données sont fragmentés et gérés soit directement via le PMS actuel, soit par le biais d'extractions manuelles au format CSV. Ce mode de fonctionnement entraîne une **dispersion des processus**, un **manque de traçabilité**, et une **complexité accrue** pour la maintenance.

Pour remédier à ces problématiques, la mise en place d'un **pipeline de données centralisé et automatisé** a été décidée, afin de garantir **l'efficacité, la fiabilité et l'évolutivité des flux d'information**.

OBJECTIFS

Objectifs quantitatifs

Assurer une refonte complète et conforme des flux de données ISO, condition indispensable pour garantir la migration réussie et intégrale du PMS vers le nouveau système.

Réduction des erreurs de transfert de données à un niveau proche de zéro grâce à l'automatisation.

Objectifs qualitatifs

Visibilité et monitoring

Offrir des interfaces de suivis claires permettant de suivre en temps réel et l'état des flux.

Traçabilité

Garantir un suivi précis des données à travers leur cycle de vie, pour une meilleure gouvernance et une conformité accrue.

Centralisation des données

Réunir la gestion des flux de données dans un pipeline unique pour éliminer les approches disparates actuelles.

Maintenance simplifiée

Créer une architecture robuste et bien documentée, facilitant la gestion et les évolutions futures.

#2

Mise en œuvre

ARCHITECTURE MISE EN PLACE

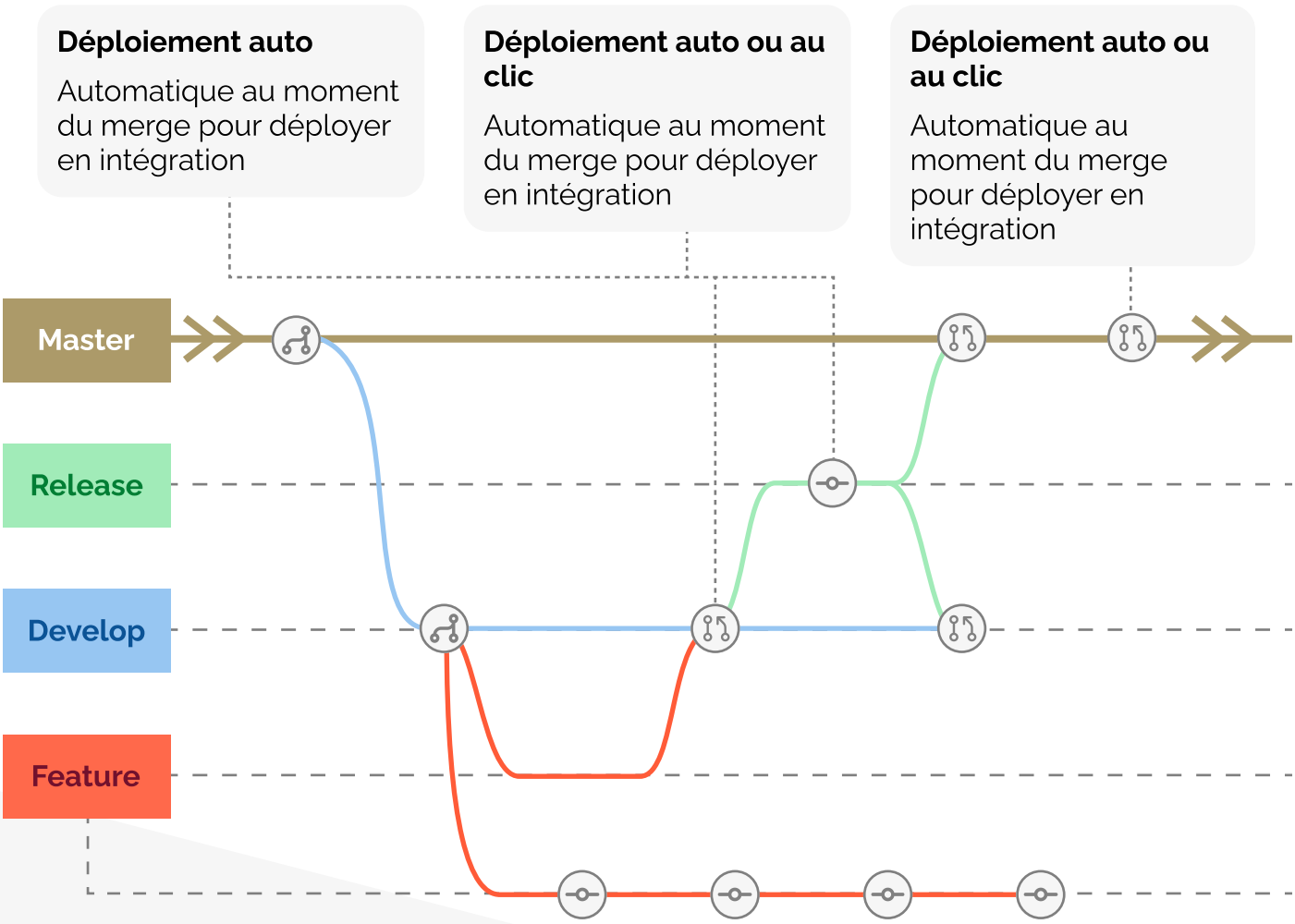
Mise en place d'un pipeline de données moderne

Basé sur Apache Airflow pour l'orchestration des tâches et dbt pour la transformation des données, garantissant une gestion automatisée et performante des flux.

Approche "Data as Code"

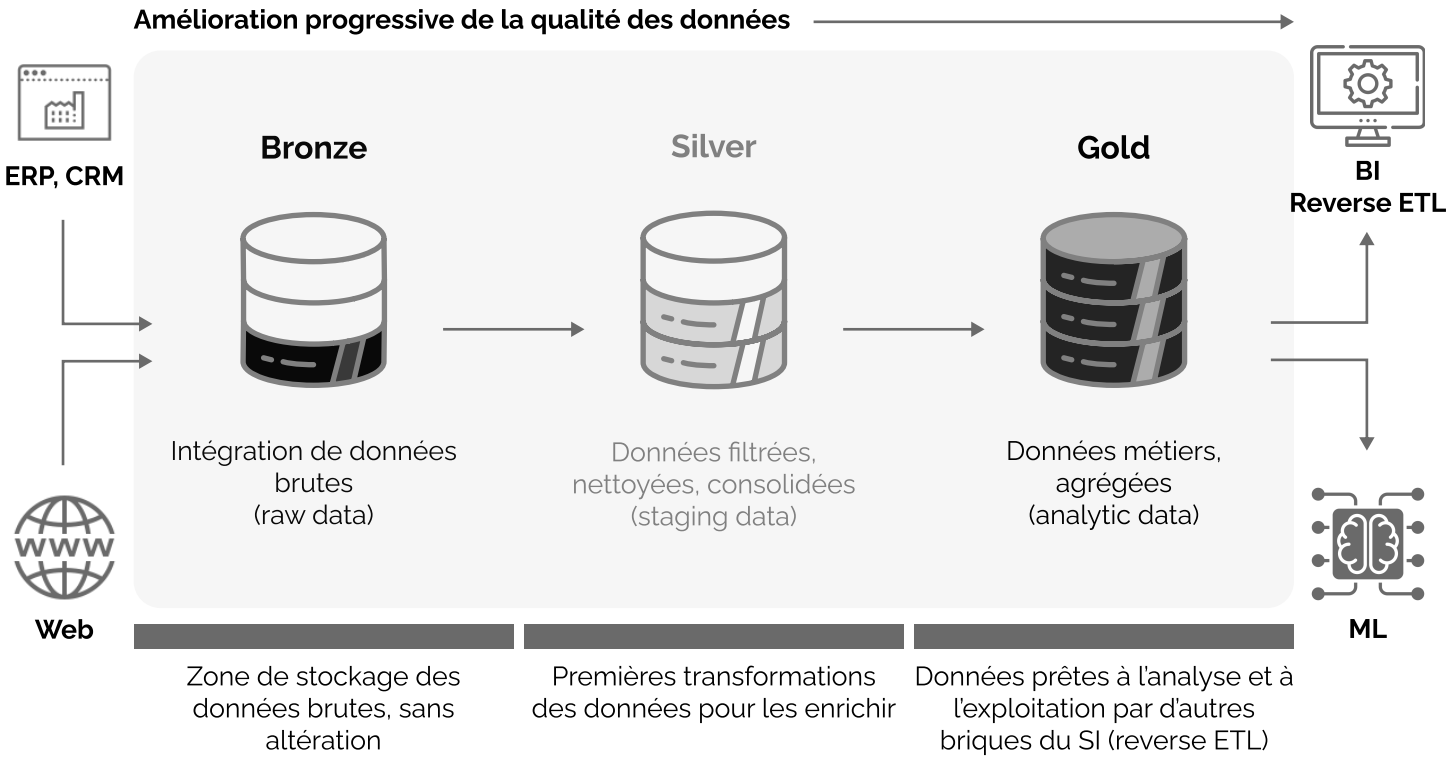
Adoption d'une méthodologie inspirée des pratiques de développement logiciel classique, avec des avantages tels que :

- Gestion du versioning avec Git.
- Code review systématique pour garantir la qualité des livrables.
- Tests automatisés et intégration continue pour des déploiements fiables et rapides.



Architecture en médaillon

Implémentation d'une architecture data conforme aux bonnes pratiques, structurée autour des couches brutes (raw), intermédiaires (staging) et finales (gold).



Amélioration de la gouvernance des données

Avec des outils et processus intégrés pour la documentation, la supervision et le respect des normes en vigueur.

BÉNÉFICES CLIENTS

Ce projet apporte des bénéfices significatifs à plusieurs niveaux pour Sandaya :

#1 | Amélioration de la fiabilité des données

Grâce à un pipeline de données centralisé et automatisé, les erreurs liées aux processus manuels et aux flux disparates sont éliminées, garantissant des données précises et cohérentes dans tous les systèmes connectés.

#2 | Efficacité opérationnelle

L'automatisation et l'orchestration des flux permettent de réduire considérablement le temps consacré à la gestion manuelle des données, libérant ainsi les équipes pour se concentrer sur des tâches à plus forte valeur ajoutée.

#3 | Visibilité et contrôle renforcés

Le monitoring en temps réel et la traçabilité des flux offrent une meilleure supervision des processus, permettant de détecter et de résoudre rapidement les éventuels problèmes.

#4 Soutien à la transformation digitale

Ce projet garantit le succès de la migration du PMS, un élément central de la stratégie digitale de Sandaya, tout en posant les bases d'une infrastructure de données moderne et scalable pour répondre aux besoins futurs.

#5 Facilité de maintenance et évolutivité

L'approche "data as code" adoptée dans ce projet simplifie la gestion des évolutions futures, grâce à une documentation automatisée, un versioning des transformations, et des pratiques modernes de développement.

Ces bénéfices positionnent ce projet comme un levier clé pour l'efficacité, l'innovation, et la compétitivité de Sandaya dans un environnement digital de plus en plus exigeant.

CHIFFRES CLÉS

35
sources de données synchronisées

chaque jour pour des flux d'information fiables et continus

10
systèmes stratégiques connectés

et alimentés quotidiennement pour une efficacité maximale

3000+
appels API

quotidiens pour maintenir les données toujours à jour

50.000.000
lignes de données transformées

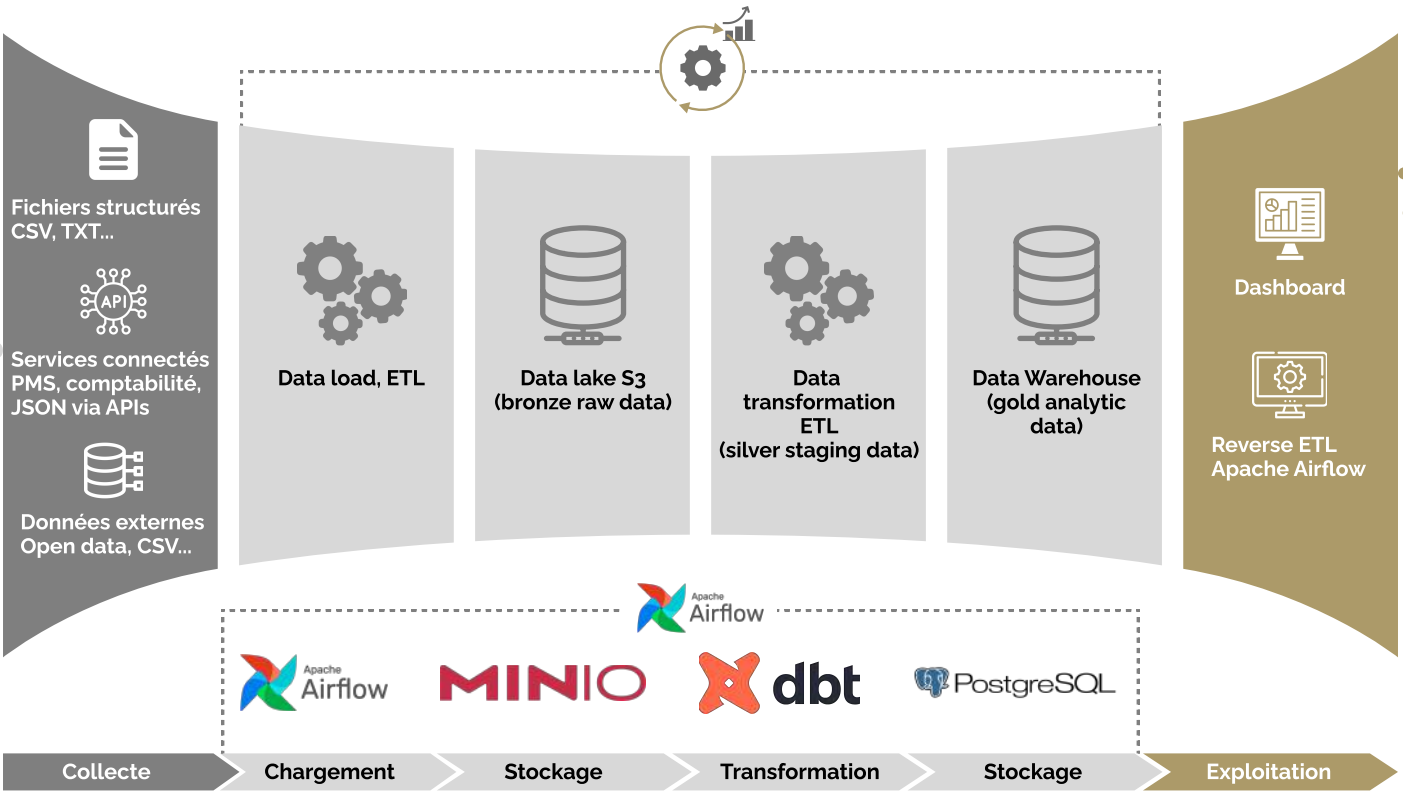
chaque année grâce à des mises à jour automatisées

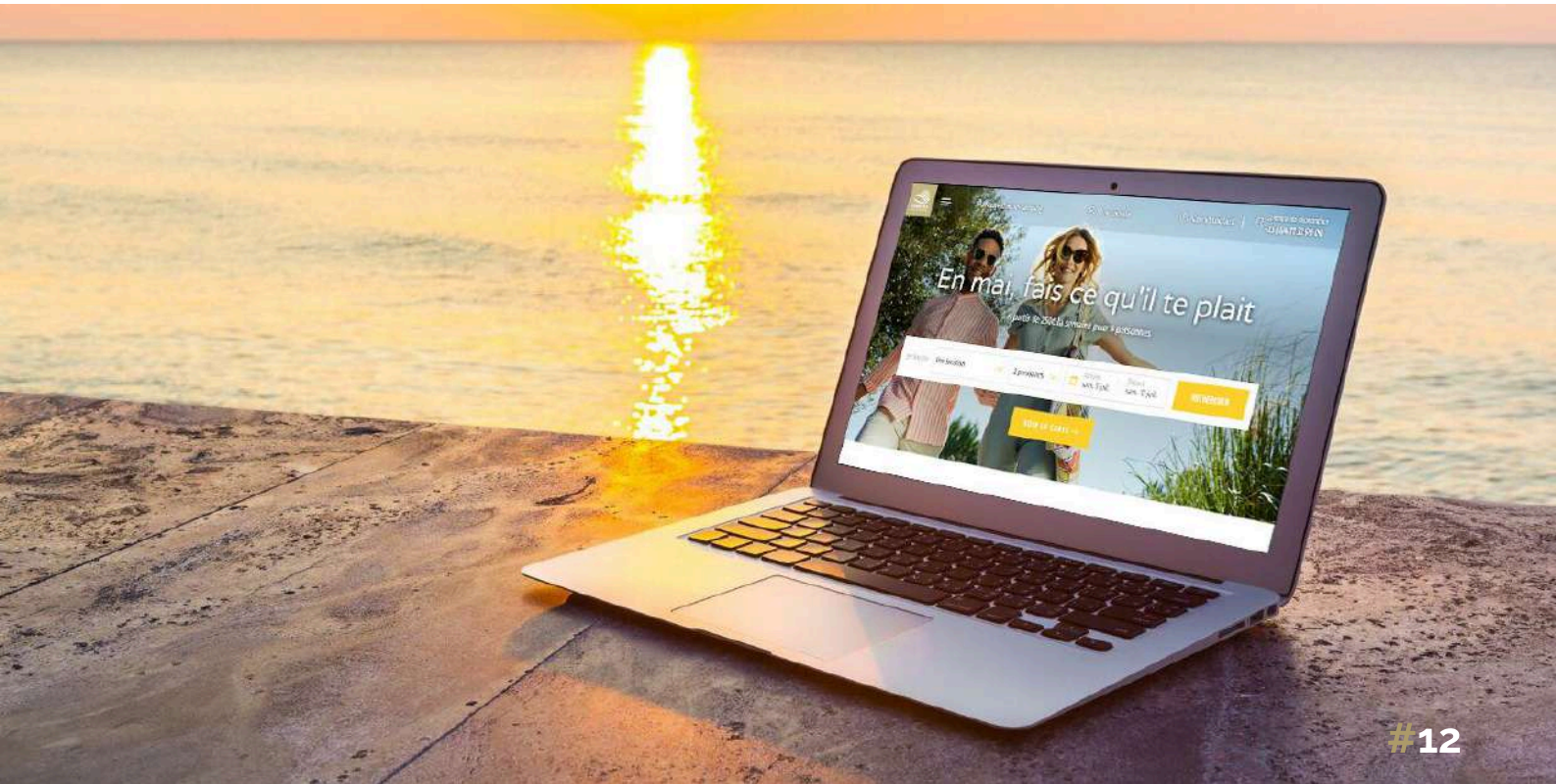
#3 Solutions

PIPELINE DE DONNÉES

Le pipeline de données mise en place pour Sandaya repose sur une architecture moderne et performante, conçue pour **collecter, transformer, stocker et exploiter les données de manière fiable et évolutive**.

Il constitue un socle stratégique pour garantir la qualité des flux d'information et accompagner la migration réussie du PMS.





Collecte

- **Sources multiples :** API, fichiers CSV, données web.
- Agrégation des données brutes provenant de différentes origines (paiements, PMS, autres systèmes).

Chargement

- **Données brutes (raw) :** Chargées sans transformation préalable grâce à Apache Airflow, garantissant une collecte fidèle des données.

Stockage

- **Data Lake S3 :** Les données brutes sont centralisées dans un stockage scalable et sécurisé.
- **Schéma raw :** Les données sont également persistées dans un schéma dédié de la base de données pour une accessibilité rapide.

Transformation

- **Nettoyage et consolidation :** dbt assure le traitement des données pour les rendre exploitables.
- **Appariement et enrichissement :** Création de données temporaires et proxy dans le schéma "staging" pour une préparation optimisée avant l'analyse.

Stockage des données consolidées

- **Data Warehouse :** Les données finales, dites "gold", sont structurées et prêtes pour l'analyse, stockées dans un environnement adapté aux besoins analytiques.

Exploitation

- **Reverse ETL et exportation :** Données utilisées pour alimenter des systèmes externes (comptabilité, BI, CRM) via API, fichiers CSV, ou TXT.
- **Mises à jour dynamiques :** Les flux critiques sont exécutés toutes les 10 minutes, tandis que les autres flux sont mis à jour quotidiennement.

Flexibilité et scalabilité

- **Orchestration ajustable :** Airflow permet de moduler la fréquence d'exécution en fonction des besoins opérationnels.
- **Clusterisation d'Airflow :** La solution est conçue pour gérer les montées en charge avec une exécution distribuée et scalable.

Grâce à cette architecture, Sandaya dispose d'un pipeline de données flexible, scalable et résilient, capable de répondre aux exigences actuelles tout en s'adaptant aux besoins futurs. Cette solution garantit une gestion optimisée des flux de données, essentielle à la réussite du projet et à l'amélioration des performances globales du système d'information.

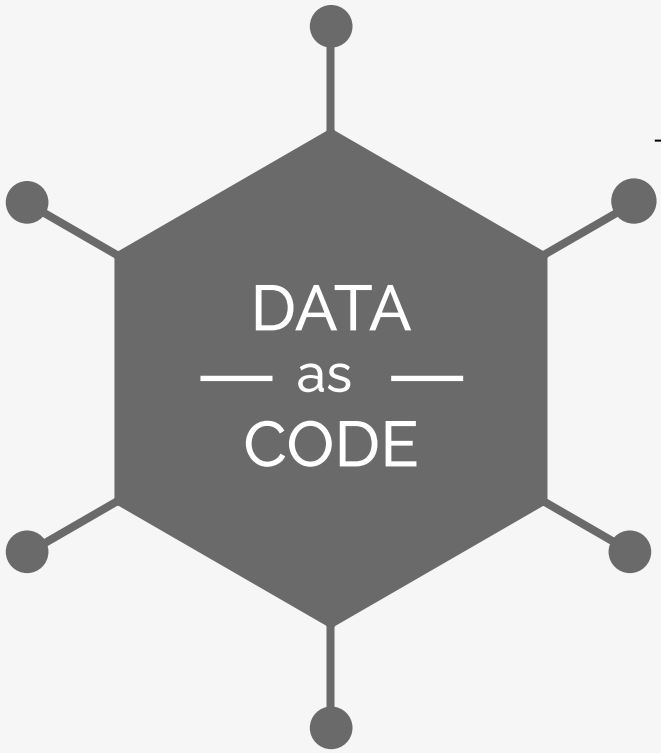
#4

Data as Code

AVANTAGES DE L'APPROCHE

L'approche "**Data as Code**" transforme significativement la manière dont les pipelines de données sont conçus, développés et maintenus, en s'appuyant sur les meilleures pratiques du développement logiciel moderne.

Cette approche transforme la manière dont les pipelines de données sont conçus, développés et maintenus, en s'appuyant sur les meilleures pratiques du développement logiciel moderne. Elle repose sur six piliers fondamentaux qui garantissent robustesse, évolutivité et collaboration.



Versioning et Traçabilité

Suivi précis et création de versions stabilisées

Propriété et Transparence

Plein contrôle grâce aux briques open source

Collaboration facilitée

Travail en équipe simplifié grâce à Git

Extensibilité et Personnalisation

Adaptabilité aux besoins spécifiques

Qualité du code

Code reviews et tests automatiques

Industrialisation des Processus

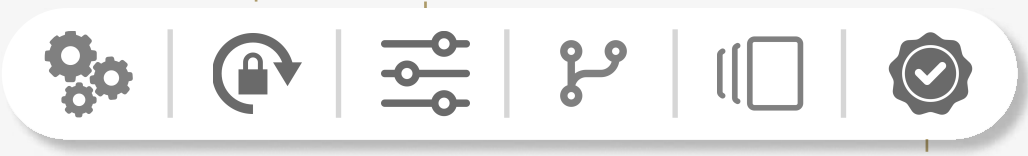
Déploiement automatisé et contrôlé

Propriété et transparence du code

- En s'appuyant sur des briques open source, le client devient pleinement propriétaire du code développé.
- Cette indépendance permet au client de confier le projet à différents prestataires ou de le gérer en interne, garantissant une flexibilité maximale et une pérennité à long terme.

Extensibilité et personnalisation

- L'architecture est conçue pour être fortement extensible : de nouvelles fonctionnalités ou solutions externes (par exemple, API ou outils d'analyse) peuvent être intégrées sans remettre en cause l'infrastructure existante.
- La personnalisation est au cœur de l'approche, permettant d'adapter les pipelines aux besoins spécifiques du client.



Amélioration de la qualité du code

- Les revues de code (Code Reviews) systématiques renforcent la qualité et la maintenabilité des livrables.
- Les pipelines de CI/CD (Intégration et Déploiement Continus) détectent les erreurs à un stade précoce, assurant un déploiement sécurisé.
- Des tests automatisés garantissent la robustesse des transformations et minimisent les risques d'erreurs en production.

Industrialisation des Processus

- L'approche standardise le cycle de vie des pipelines de données, de leur développement à leur mise en production.
- Des outils tels que Git, CI/CD, Capistrano (déploiement automatisé), et SonarQube (analyse de qualité du code) apportent un niveau élevé d'automatisation et de contrôle qualité.
- Le déploiement devient répétable, prévisible et industrialisé, assurant un gain significatif de temps et de fiabilité.

Versioning et Traçabilité

- Chaque action est historisée grâce au versioning Git, permettant un suivi précis des évolutions et modifications.
- Les versions stabilisées du projet peuvent être balisées via des releases et des tags, garantissant des bases fiables pour les déploiements ou retours en arrière en cas de besoin

Collaboration Facilitée

- L'intégration d'outils collaboratifs tels que Git facilite la montée en charge des équipes, en permettant un travail simultané et structuré sur les pipelines de données.
- Chaque membre de l'équipe peut contribuer sans crainte de conflits grâce à la gestion des branches, ce qui accélère le développement.

LES TECHNOLOGIES

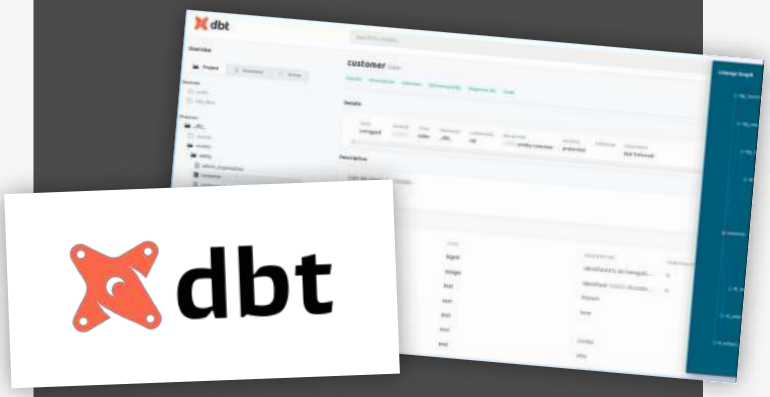
L'association d'Apache Airflow et dbt (Data Build Tool) constitue une solution robuste et performante pour orchestrer, monitorer et transformer des pipelines de données de manière fluide et efficace.

Apache Airflow

Apache Airflow est une plateforme open-source reconnue pour **l'orchestration de flux de travail**. Elle permet de **planifier**, **exécuter** et **surveiller** des **tâches complexes** au sein de pipelines de données, tout en offrant une **visibilité en temps réel** grâce à une interface utilisateur intuitive. Sa flexibilité et son extensibilité en font un outil idéal pour **automatiser les flux de données** dans des environnements complexes et hétérogènes.


© Crédit photo : Sandaya

dbt



De son côté, **dbt est un outil spécialisé dans la transformation des données**, qui adopte une approche centrée sur les développeurs ("data as code"). Il permet de modéliser, tester et documenter les données de manière efficace, en tirant parti des bonnes pratiques du développement logiciel telles que le **versioning**, les **tests unitaires** et les **révisions de code**.
En plus de sa capacité à structurer les données en couches (raw, staging, gold), **dbt génère automatiquement une documentation claire et interactive**, essentielle pour garantir une gouvernance des données de haute qualité.

En combinant Airflow pour l'orchestration des tâches et dbt pour les transformations, cette solution apporte :

- **Une orchestration puissante** : Airflow garantit le bon enchaînement des étapes, la gestion des dépendances, et un monitoring détaillé de chaque tâche.
- **Une transformation avancée et documentée** : dbt simplifie la création de pipelines de transformation optimisés, tout en générant une documentation détaillée et accessible pour assurer la transparence et la maintenabilité.
- **Une architecture modulaire et scalable** : Cette combinaison permet de répondre aux besoins des organisations, qu'il s'agisse de traiter des volumes massifs de données ou de gérer des pipelines de données complexes et interconnectés.

Pour conclure

Avec Airflow et dbt, les entreprises disposent d'une solution complète pour **automatiser**, **transformer**, et **superviser leurs flux de données**, tout en bénéficiant d'une traçabilité accrue et d'une **maintenance simplifiée**. Cette synergie est idéale pour répondre aux exigences croissantes des environnements data-driven modernes.

Romain Bonnal
Responsable Data et BI
Codéin, Agence web open source

A PROPOS DE CODÉIN

Codéin est une agence web open-source spécialisée en **développement, data, hosting et conseil en SI.** Depuis plus de 10 ans, nous accompagnons nos clients dans leur transformation numérique avec une expertise technique éprouvée. Nos chefs de projet, tous experts techniques, assurent une gestion rigoureuse grâce à une méthodologie affinée, un pilotage régulier et des KPI suivis en temps réel.

Nous collaborons avec des clients comme le **CIRAD**, **l'INRAE**, la **Fédération Française de Golf**, la **Banque Française Mutualiste**, les **Aéroports de Nice Côte d'Azur**, **l'OPPBTP** et le **Groupe Sandaya**. Chez Codéin, expertise, méthodologie et qualité de service sont au cœur de chaque projet.

Montpellier

09.72.42.26.03

6 rue de Maguelone
34000 Montpellier

Strasbourg

09.72.58.09.96

3 place de Haguenau
67000 Strasbourg

contact@codein.fr ▌www.codein.fr

" L'équipe Codéin est principalement en charge du site e-commerce de Sandaya depuis de nombreuses années. Très satisfait du résultat, nous avons décidé de leur confier la mise en œuvre d'une stack data / ETL au cœur de notre écosystème web. Notre site e-commerce répond sans faillir aux centaines de milliers de requêtes hebdo de nos internautes. Le rythme de mise à jour mensuel permet une grande réactivité, le tout dans une ambiance relationnelle sympathique et détendue. Mention spéciale pour notre nouvelle équipe data qui commence très fort ;) "

Christophe Musielak
Directeur des systèmes d'information

codein.

Digital **addict**

contact@codein.fr | www.codein.fr