BRIEF 1C

Ma première application

Présentation

- **Contexte** :

  Je suis développeur chez **Analysis Features Preprocessing And Research**, une ESN spécialisée dans la réalisation d'applicatifs de type BI et intelligence artificielle.

  On m'a confié la tâche de réaliser un proof of concept (PoC) dans le cadre d'un projet de dashboard d'aide à la décision pour un client exigeant. J'ai accès à un fichier de données brutes, matérialisant un export depuis leurs bases de données opérationnelles.

  Ce fichier CSV alimentera la base analytique et tient lieu de situation initiale. Les CSV des mois suivants me seront régulièrement transmis.

  IMPORTANT. Le dashboard - en accès retreint - devra permettre à l'utilisateur de déclencher son import :

  - avec suffisamment de feedback pour comprendre les décisions ETL automatisées ou semi automatisées proposées
  - de façon cumulative : d'autres CSV vont arriver

  Le dashboard comprendra obligatoirement ces éléments suivants :

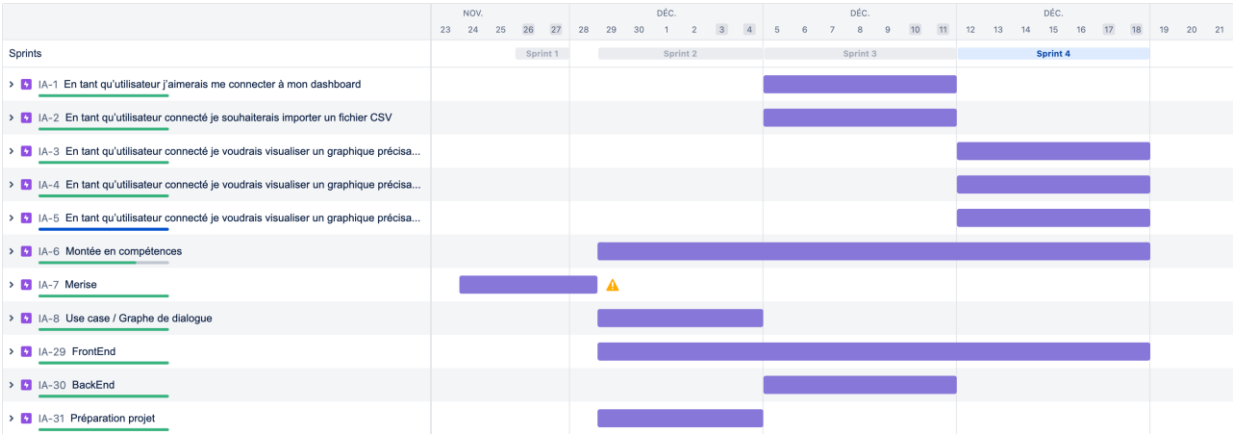
  - un graphique précisant la répartition des ventes par produit
  - un graphique précisant la répartition des ventes par région
  - un dernier graphique précisant la répartition des ventes par région et par produit

- **Priorité** :
  - nettoyer les données brutes (suppression des doublons et "faux", standardisation, etc.)
  - concevoir et paramétrer la base de données analytique
  - récupérer les données en base nécessaires pour le dashboard
  - développer les interfaces et les graphiques du dashboard

Objectifs

1. Accès restreint au dashboard
2. Import de fichier csv
3. Déclencher le nettoyage des données
4. Récupérer les données en BDD et affichage des graphiques

Plan d’actions



R & D

Chart.js

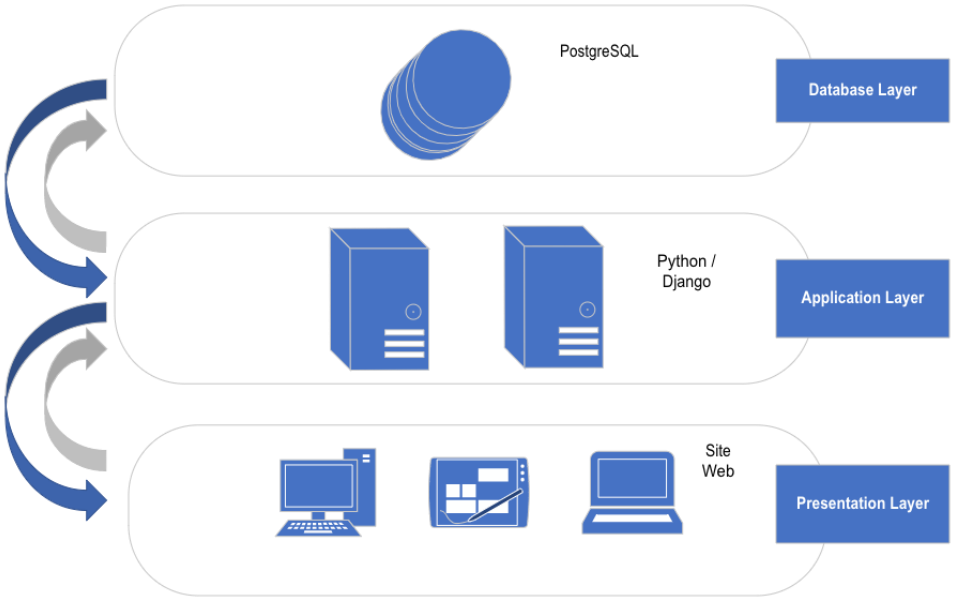
Pandas

Transfert csv vers PostgreSQL

Requête vers PostgreSQL

Spécificités techniques

Architecture 3-Tiers



Technologies utilisées

- Html
- CSS
- Javascript
- Bootstrap
- Python
- Django
- Chart.js => plugin DataLabels
- PostegreSQL
- Visual Studio Code
- Jupyter Notebook

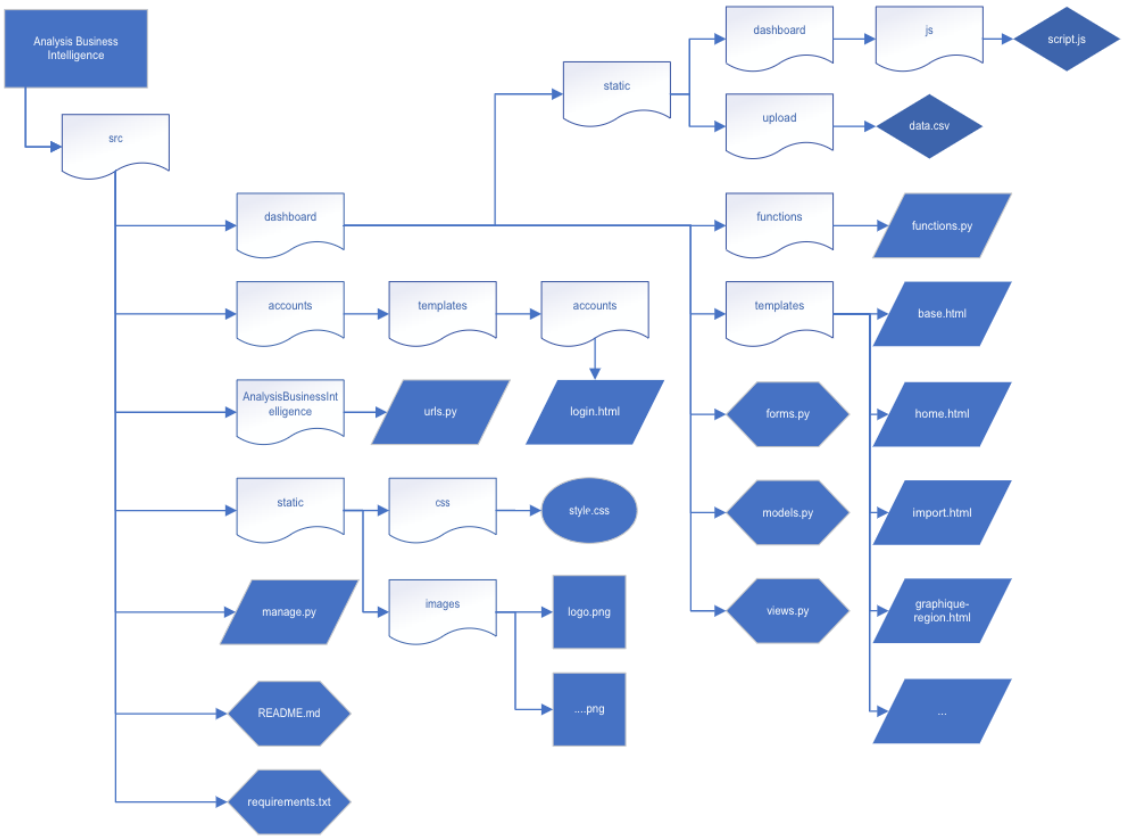
Outils Django utilisés

- L'interface d'administration (authentification)
- Les fichiers statiques (css, images, …)
- Les formulaires (Upload file, …)
- Les gabarits ({{ … }})
- Les modèles, relations entre modèles
- Les requêtes
- Les path (urls)

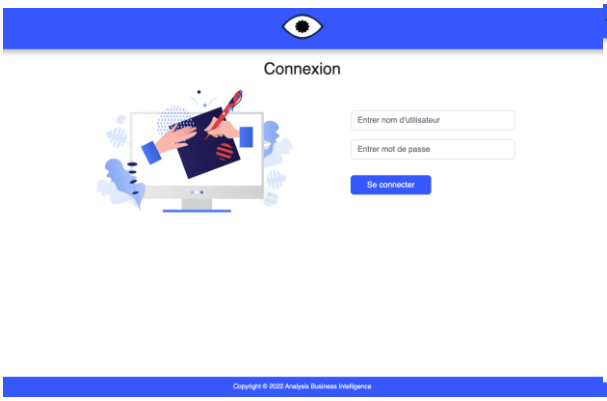
Outils Python utilisés

- Pandas
- Numpy
- Sqlalchemy (interroger la BDD)

Arborescence du site



Maquettes et / ou Screens du site



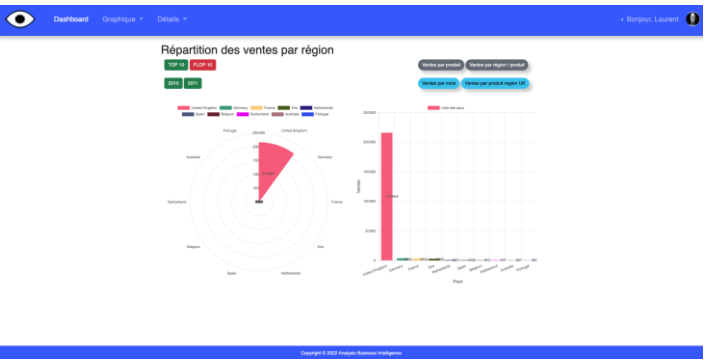

Dashboard
Bonjour, Laurent
Choisir un fichier
Aucun fichier choisi
Importer csv
Fichier CSV, en-tête à respecter


Dashboard
Bonjour, Laurent
Choisir un fichier
data2010201111.csv
Importer csv
Fichier CSV, en-tête à respecter


Copyright © 2022 Analysis Business Intelligence
Dashboard
Bonjour, Laurent
CSV importé avec succès!
Analyse du fichier CSV importé
Feedback
Nombre de lignes de facturations: 245903
A nettoyer
Nombre de lignes à supprimer: 10263
Nombre de champs réparable: 315841
Après nettoyage
Nombre de lignes de facturations: 235640
Pourcentage suppression: 4.17 %
Vous pouvez nettoyer les données
Nettoyer données

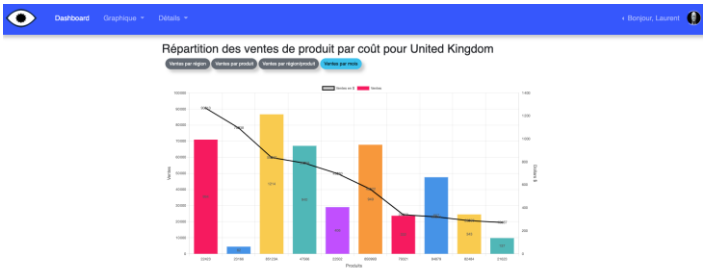

Copyright © 2022 Analysis Business Intelligence
Dashboard
Graphique
Détails
Bonjour, Laurent
CSV nettoyé et enregistré dans la base de donnée!
Choisir un fichier
Aucun fichier choisi
Importer csv
Fichier CSV, en-tête à respecter


Copyright © 2022 Analysis Business Intelligence
Dashboard
Graphique
Détails
Bonjour, Laurent
Dashboard
Bienvenue sur votre dashboard. Importer vos CSV, nettoyer les données et analyser vos graphiques
Importer un fichier CSV
Importer
CHIFFRES D'AFFAIRES
4614062 $
FACTURES
9592
PRODUITS
3495
LIVRAISONS
37
Graphiques
Ventes par région
Ventes par produit
Ventes par région et par produit
Ventes par mois
Ventes par produit region UK


Copyright © 2022 Analysis Business Intelligence
Dashboard
Graphique
Détails
Bonjour, Laurent
Répartition des ventes par région
TOP 10
FLOP 10
2010
2011
Ventes par produit
Ventes par région / produit
Ventes par mois
Ventes par produit region UK


Copyright © 2022 Analysis Business Intelligence
Dashboard
Graphique
Détails
Bonjour, Laurent
Répartition des ventes par produit
TOP 10
FLOP 10
2010
2011
Ventes par région
Ventes par région/produit
Ventes par mois
Ventes par produit region UK


Copyright © 2022 Analysis Business Intelligence
Dashboard
Graphique
Détails
Bonjour, Laurent
Répartition des ventes par région / produit
TOP 10
FLOP 10
2010
2011
Ventes par région
Ventes par produit
Ventes par mois
Ventes par produit region UK

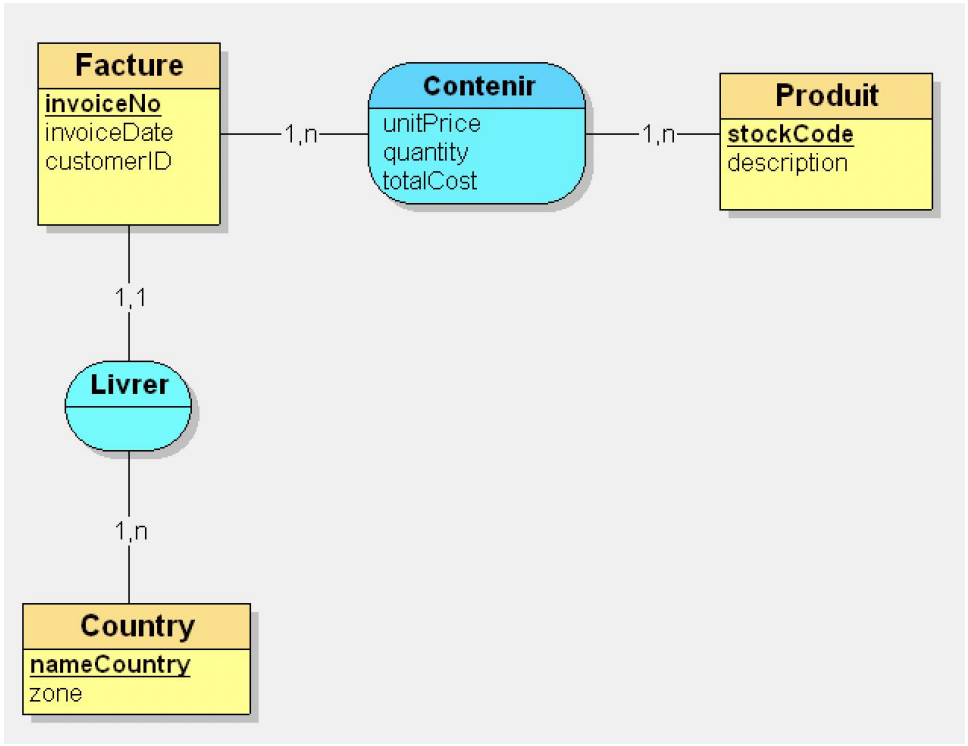

Dashboard
Graphique
Détails
Bonjour, Laurent
Répartition des ventes de produit par coût pour United Kingdom


Dashboard
Graphique
Détails
Bonjour, Laurent
Répartition des ventes par mois

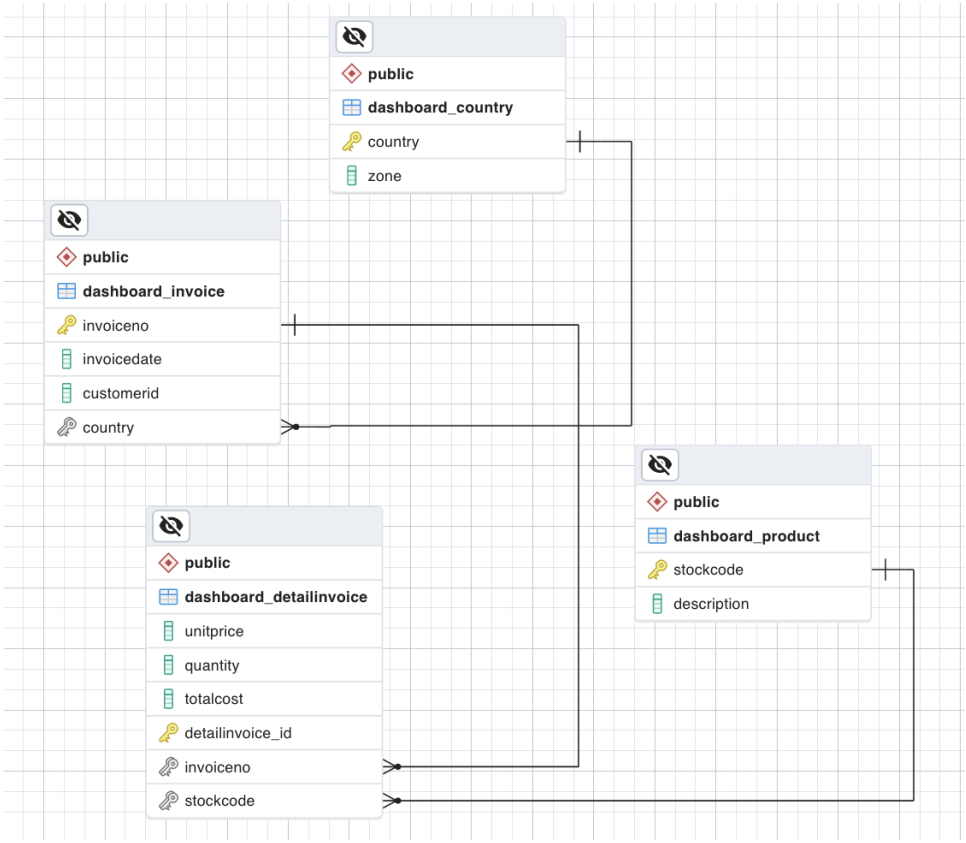

Dashboard
Graphique
Détails
Bonjour, Laurent
Liste des produits référencés et leurs prix moyens de facturations

Merise

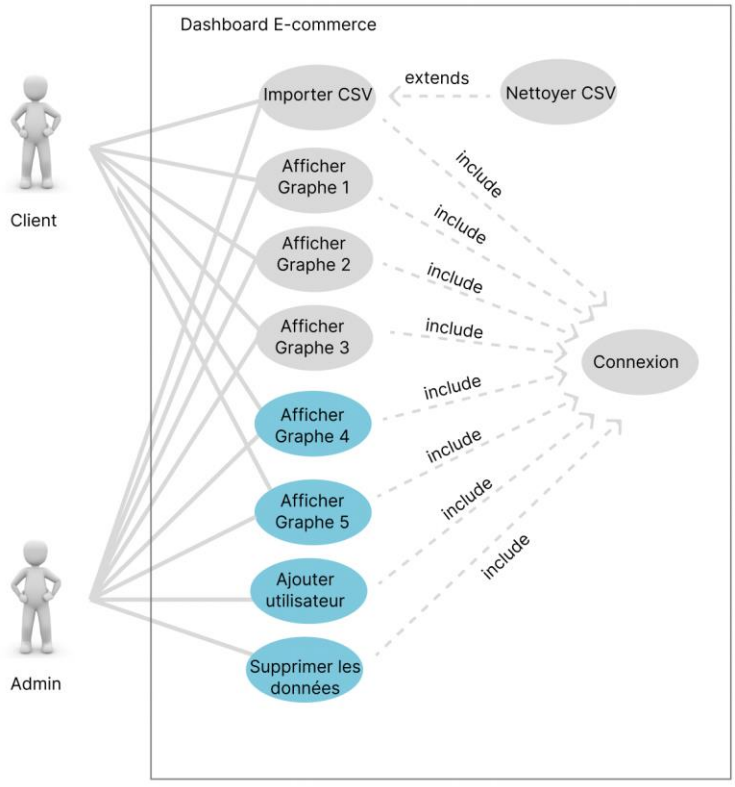
- MCD


Facture
invoiceNo
invoiceDate
customerID
1,n
Contenir
unitPrice
quantity
totalCost
1,n
Produit
stockCode
description
1,1
Livrer
1,n
Country
nameCountry
zone

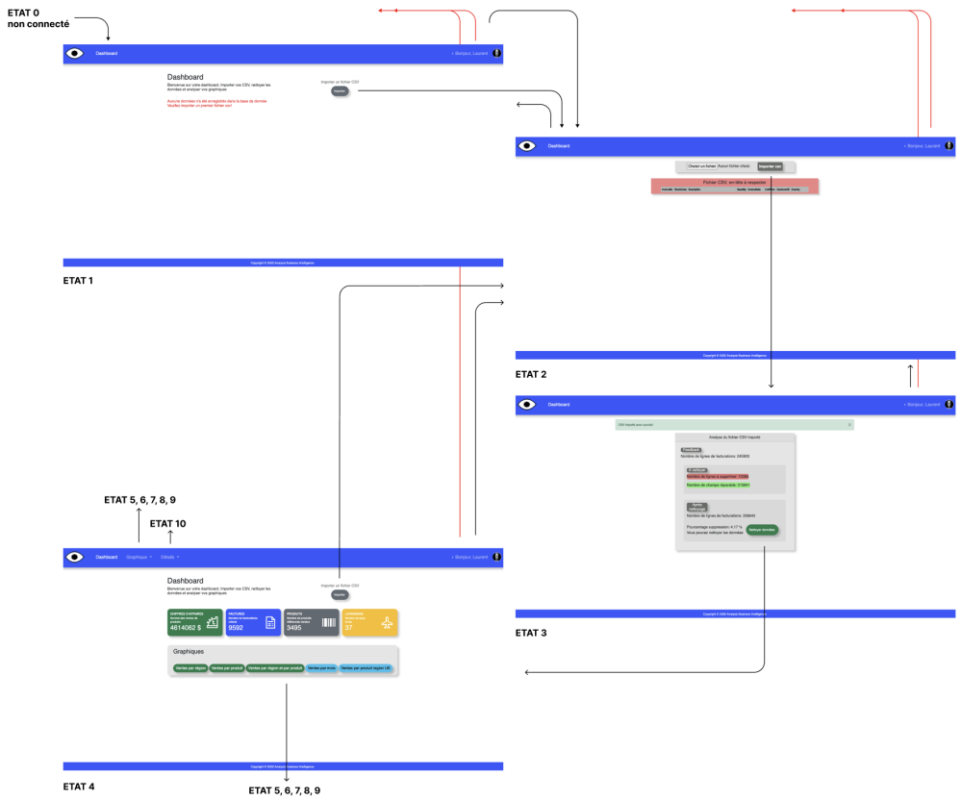
- MPD


public
dashboard_country
country
zone
public
dashboard_invoice
invoiceno
invoicedate
customerid
country
public
dashboard_product
stockcode
description
public
dashboard_detailinvoice
unitprice
quantity
totalcost
detailinvoice_id
invoiceno
stockcode

UseCase


Dashboard E-commerce
Client
Admin
Importer CSV
extends
Nettoyer CSV
Afficher Graphe 1
Afficher Graphe 2
Afficher Graphe 3
Afficher Graphe 4
Afficher Graphe 5
Ajouter utilisateur
Supprimer les données
Connexion
include
include
include
include
include
include
include
include

Graphe de dialogue



ETAT 0 :

Non connecté

ETAT 1 – 2 états possibles, une page Dashboard sans données importées présent dans la Base De Données (BDD) et une autre page Dashboard avec données importées dans la BDD (voir ETAT 4) :

1 bouton cliquable - “Importer” + 1 menu déroulant - 2 liens cliquables “Importer CSV” + “Déconnexion”

ETAT 2 :

2 boutons cliquables - “Choisir un fichier” et “Importer CSV” + 1 menu déroulant - 2 liens cliquables “Importer CSV” + “Déconnexion”

1 zone de message

ETAT 3 :

1 bouton cliquable - “Nettoyer données” + 1 menu déroulant - 2 liens cliquables “Importer CSV” + “Déconnexion”

1 zone feedback

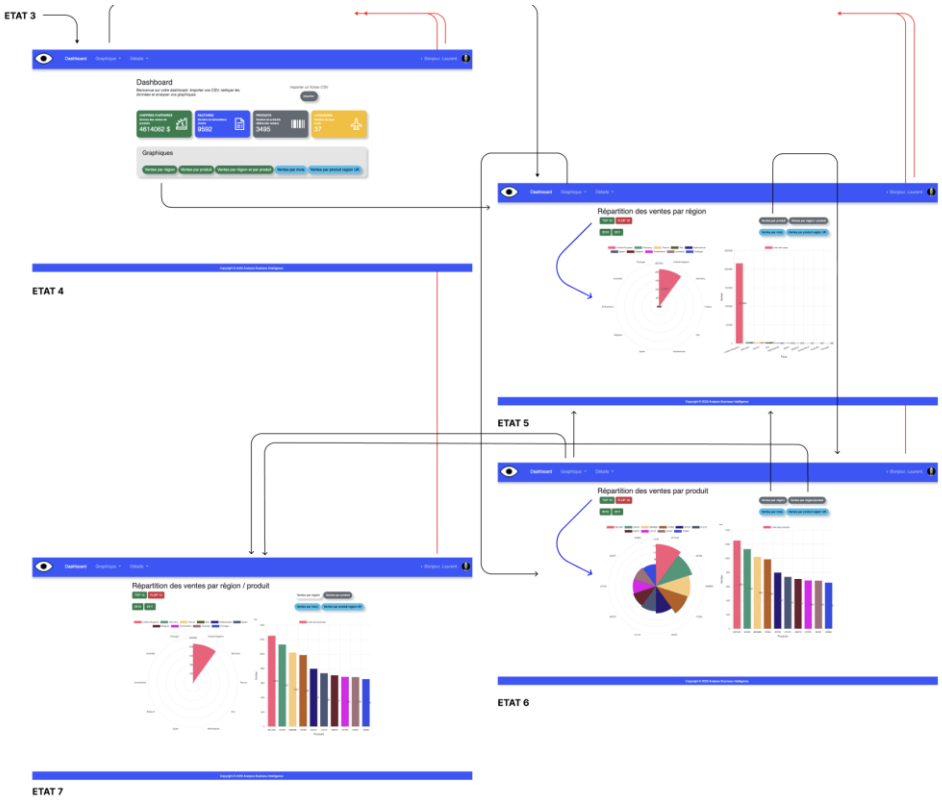
ETAT 4 :

1 + 5 boutons cliquables - “Importer”, “Ventes par région”, “Ventes par produit”, Ventes par région et par produit”, Ventes par mois et “Ventes par produit région UK”

1 menu déroulant - 2 liens cliquables “Importer CSV” + “Déconnexion”

1 menu déroulant “Graphique” - 5 liens cliquables “Ventes par région”, “Ventes par produit”, Ventes par région et par produit”, Ventes par mois et “Ventes par produit région UK”

1 menu déroulant “Détails - 1 lien cliquable “Liste des produits référencés”



ETAT 4 :

Voir ci-dessus

ETAT 5, 6 et 7 :

2 boutons cliquables “TOP 10” et “FLOP 10” + 2 boutons “2010” et “2011”

4 boutons cliquables des différents graphiques

1 menu déroulant - 2 liens cliquables “Importer CSV” + “Déconnexion”

1 menu déroulant “Graphique” - 5 liens cliquables “Ventes par région”, “Ventes par produit”, Ventes par région et par produit”, Ventes par mois et “Ventes par produit région UK”

1 menu déroulant “Détails - 1 lien cliquable “Liste des produits référencés”

2 zones graphiques

ETAT 8 et 9 :

4 boutons cliquables des différents graphiques

1 menu déroulant - 2 liens cliquables “Importer CSV” + “Déconnexion”

1 menu déroulant “Graphique” - 5 liens cliquables “Ventes par région”, “Ventes par produit”, Ventes par région et par produit”, Ventes par mois et “Ventes par produit région UK”

1 menu déroulant “Détails - 1 lien cliquable “Liste des produits référencés”

1 zone graphique

ETAT 10 :

1 menu déroulant - 2 liens cliquables “Importer CSV” + “Déconnexion”

1 menu déroulant “Graphique” - 5 liens cliquables “Ventes par région”, “Ventes par produit”, Ventes par région et par produit”, Ventes par mois et “Ventes par produit région UK”

1 menu déroulant “Détails - 1 lien cliquable “Liste des produits référencés”

1 zone table

ETL

Stratégie de nettoyage

- Standardisation des noms de colonnes : tous en minuscules (facilite la liaison avec postgreSQL)

```
# #Mettre en minuscule les noms de colonnes
df.columns = [x.lower() for x in df.columns]
```

- Suppression des doublons et combinaison InvoiceNo – StockCode

```
#Supprimer les lignes doublon
df = df.drop_duplicates()
```

```
#Supprimer les doublons invoice/stockcode
df = df.drop_duplicates(['invoiceno','stockcode'])
```

- Suppression des lignes avec quantités négatives / **à revoir si on veut comptabiliser le nombre de lignes de ventes et leurs coûts totaux car cela fausse les données sur les graphiques (lignes d’avoir, entrées manuelles, produits endommagés, …)**

```
#Supprimer les lignes quantity <=0
indexQuantity = df[df['quantity']<=0].index
df.drop(indexQuantity,inplace=True)
```

- Suppression des lignes avec stockcode qui commence par un caractère, cela représente des entrées manuelles sans numéro référencés ou des produits divers (des bons, …)

```
#Supprimer les lignes stockcode qui commence par une chaîne de caractère
indexStockCode = df[df["stockcode"].str.match("^[A-Za-z]")==True].index
df.drop(indexStockCode,inplace=True)
```

- Champs vides
  - Colonne Country
    - Ligne sans valeur ou avec country numérique => bascule en *Unspecified*

```
#Mettre les lignes avec country NULL ou Numérique en "unspecified"
df.replace('NaN',np.nan)
df['country'].fillna(value='Unspecified')
```

  - Colonne Description
    - Je remplis les champs vides d'une phrase "*Description du produit à intégrer*" afin de pouvoir les récupérer si création d'une fonctionnalité d'analyse des données en erreur (téléchargement ficher Excel ou visualisation directement dans l'application)

```
df['description'].fillna(value='Description du produit à intégrer')
```

- Normalisation
  - Colonne Country
    - Pays en majuscule => bascule en minuscule (ex : EIRE -> Eire)

```
# Mettre en Minuscule les lignes country EIRE -> Eire
df['country'] = df['country'].str.title()
```

Réalisation

Le système d'authentification => accès à l'application pour un administrateur et un utilisateur final (le client)

La page Dashboard qui à 2 états, un état initial sans donnée (aucune importation ayant été réalisée) et un état avec données importées

Retranscription de certaines données de la BDD sur la page Dashboard

Création des boutons accès aux différents graphiques sur la page Dashboard, apparaissent quand des données sont présentes en BDD

Création du système d'importation de fichier CSV, bouton sur la page Dashboard qui amène vers la page dédiée pour le système d'importation

La page d'importation de fichier CSV avec recommandations avant importation et ensuite page d'analyse des données (une fois fichier importé) et enfin apparition bouton "Nettoyage données" et recommandations si nettoyage doit se faire ou pas en fonction d'un pourcentage qui sera définit par le PO ou le client (par défaut ce pourcentage est à 5%)

Création des pages des différents graphiques avec comme éléments de filtrage, les TOP 10 et FLOP 10 ainsi que périodiques 2010 et 2011

Création de 2 graphiques supplémentaires en option aux vues des données importées et des premiers retours d'analyses (à valider ou non par le Product Owner)

Problèmes rencontrés

Montées en compétences pour les parties chart.js, pandas et requêtes SQL

Transfert des données du fichier CSV vers la BDD (utilisation de la fonction tosql() de Pandas et les requêtes SQL)

Réalisation du dernier graphique demandé "Répartition des ventes par région et par produit"

Identifier les graphiques les plus représentatifs pour analyser les données

Savoir évaluer la difficulté d'une tâche, avec ou non une nouvelle technologie

Etat des lieux

Ceux qui fonctionnent

- Le système d'authentification => admin et end-user

- l'importation du fichier CSV

- Le feedback (analyse du fichier CSV avant nettoyage)

- Le nettoyage et l'envoi vers la BDD

- Utilisation des données de la BDD pour affichage graphiques

- Sur les graphiques, TOP et FLOP année 2010 et 2011

Ceux qu'ils restent à faire

- Le graphique des ventes par région / produit n'est pas encore fonctionnel

- L'étape de nettoyage doit être encore plus robuste pour palier à toute éventualité dans le fichier CSV (ex : valider que toutes les colonnes avec le bon nom soit présent, validation du format des dates, …)

- En fonction de la validation ou non des graphiques proposés en option, revoir le processus de nettoyage afin d'éviter de supprimer des lignes essentielles à la bonne analyse de ces graphiques

- Sur tous les graphiques, faire mieux ressortir les données (labels) au niveau de leur couleur

- Créer une fonctionnalité pour récupérer les données supprimées et les mettre à disposition pour analyse soit sur l'application soit en téléchargement fichier Excel par exemple

Améliorations continues

- Mieux découper les applications au niveau de la programmation afin d'éviter de surcharger le views.py de l'application Dashboard

    . accounts

    . dashboard

    . graphic

    ...

- Mieux définir le code en petites fonctions au lieu d'avoir des dizaines de lignes dans une seule fonction
- Continuer à utiliser Jira comme support d'outil SCRUM

Ressources

Lien du repo :

https://github.com/LaurentDIA16/IA_AnalysisBusinessIntelligence.git

Django :

https://www.w3schools.com/django/index.php

https://docs.djangoproject.com/fr/4.1/intro/tutorial01/

Utilisation PostgreSQL :

https://www.postgresqltutorial.com/postgresql-getting-started/install-postgresql/

Maquette / UseCase / Graphe de dialogue :

https://www.figma.com/file/AkVtrEJ4b3YERifp4tGKsl/BRIEF-1C?node-id=114%3A153&t=5PuVwalKO2vkmxJP-1

Méthode scrum :

https://romannart.atlassian.net/jira/software/projects/IA/boards/4/roadmap?timeline=WEEKS&shared=&atlOrigin=eyJpIjoiYzBiOTY1YzFiODUyNDFjMjljODJlZGI0ZDZlYmYxZTciLCJwIjoiaiJ9

Import fichier CSV :

https://www.javatpoint.com/django-file-upload

https://www.geeksforgeeks.org/how-to-delete-a-csv-file-in-python/

https://stackoverflow.com/questions/56115382/rename-changing-csv-file-name-in-directory

Pandas :

https://pandas.pydata.org/pandas-docs/stable/index.html