

RAPPORT : Démarche pour l'analyse des Logs avec Pentaho

Imad El Maftouhi

2023/2024

—

Ingénierie des données

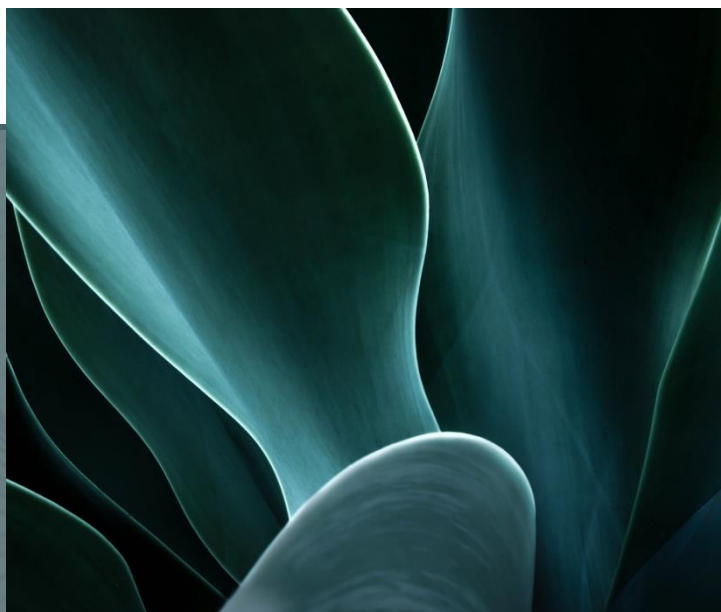
—

Prof. Hassan ZILI

Contexte

L'analyse des logs informatiques est cruciale pour la compréhension du fonctionnement d'un système et la détection de problèmes potentiels. Pentaho, une suite logicielle décisionnelle open-source, permet un accès, une visualisation et une exploration facilités des données impactant les résultats de l'entreprise.

Ce rapport présente une méthodologie complète pour réaliser une analyse approfondie des logs informatiques, de l'intégration des données Excel à la création de tableaux de bord dynamiques. Il démontre ainsi la puissance et la flexibilité de Pentaho dans le domaine de l'analyse des données.





Introduction

Contexte

Pentaho est une suite logicielle décisionnelle qui permet aux utilisateurs professionnels et aux informaticiens d'accéder, de visualiser et d'explorer facilement toutes les données ayant un impact sur les résultats de l'entreprise. Le but de ce travail est la réalisation d'un dossier qui synthétise la démarche à suivre pour effectuer un dossier d'analyse des logs informatiques partant de l'intégration des données Excel jusqu'à la visualisation des tableaux de bord.

Objectifs

Les objectifs principaux de ce projet sont :

- Intégrer des données Excel dans Pentaho
- Transformer et analyser les données avec Pentaho
- Créer des tableaux de bord pour visualiser les résultats

Méthodologie

La méthodologie suivie pour ce projet comprend les étapes suivantes :

- Collecte et intégration des données Excel dans Pentaho
- Transformation et préparation des données avec les outils ETL de Pentaho
- Analyse des données et extraction des insights
- Conception et création des tableaux de bord avec Pentaho
- Interprétation des résultats et conclusions

Table of Contents

- I. Présentation de Pentaho 6
 - Caractéristiques de Pentaho 6
 - Architecture de Pentaho..... 6
 - Utilisateurs de Pentaho 6
- II. Intégration de Données avec ETL.....7
 - Présentation de l'ETL Pentaho7
 - Intégration de Données avec l'ETL Pentaho7
- III. Dashboard :..... 11
 - 1. Installation et configuration de Pentaho Server 11
 - 2. Connexion au Serveur Pentaho.....14
 - 3. Création d'un Nouveau Tableau de Bord.....15
 - 4. Configuration du Tableau de Bord15
 - 5. Importation des données :.....16
 - 6. Ajout de Contenu17
- IV. Conclusion19

I. Présentation de Pentaho

Caractéristiques de Pentaho

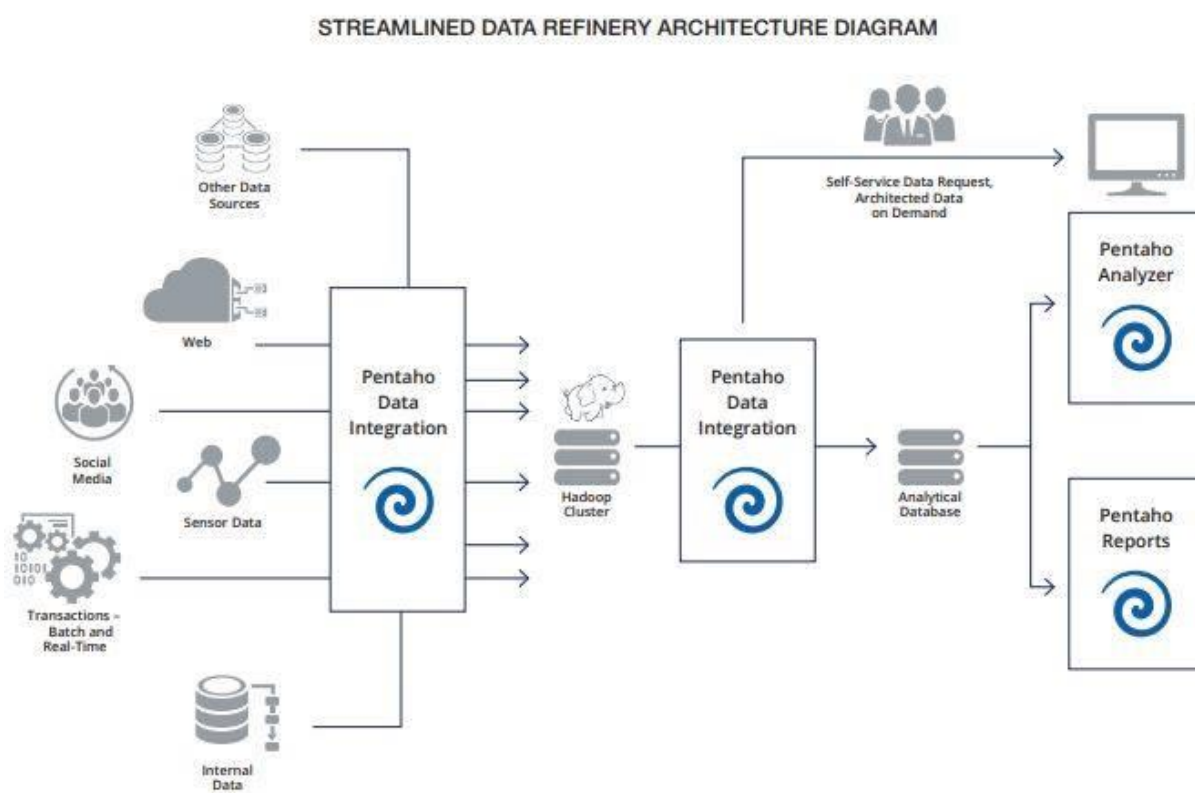
Pentaho est une plateforme décisionnelle open-source offrant des fonctionnalités telles que l'ETL, les rapports, les tableaux de bord, l'analyse ad

hoc et l'OLAP. Elle s'adresse aux utilisateurs de base et avancés, avec une architecture web 2.0, une console utilisateur et divers clients riches pour la conception et la publication de ressources.



Architecture de Pentaho

Pentaho a une architecture orientée processus, permettant la modélisation de workflows BI avancés sans nécessiter de connaissances en Java, seulement une maîtrise de SQL et des compétences de base en XML, HTML et JavaScript.



Utilisateurs de Pentaho

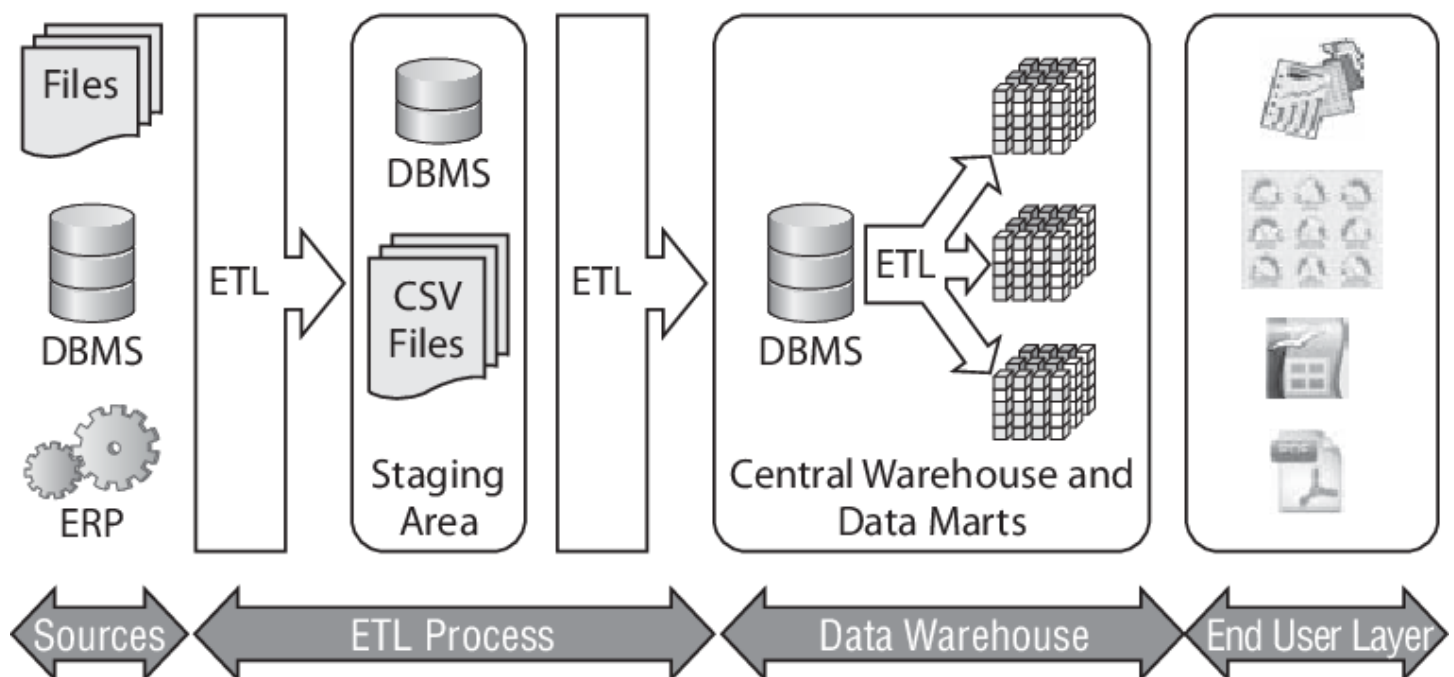
Pentaho a une communauté dynamique contribuant à des plugins et des projets communautaires, avec la version gratuite fournissant une plateforme BI complète. Ce rapport présente les étapes clés pour réaliser une analyse des logs informatiques à l'aide de Pentaho, en partant de l'intégration des données Excel jusqu'à la création de tableaux de bord. Les sections suivantes détailleront chaque phase du projet.

II. Intégration de Données avec ETL

Présentation de l'ETL Pentaho

Pentaho Data Integration (PDI), également connu sous le nom de Kettle, est un outil d'ETL (Extract, Transform, Load) open-source qui permet de concevoir et d'exécuter des processus complexes de manipulation et de transformation de données[1][3]. Il offre une bibliothèque d'objets de mapping, des fonctionnalités avancées pour le datawarehousing, un moteur d'exécution scalable et des connecteurs à diverses technologies [1].

PDI est entièrement intégré à la suite Pentaho et est disponible en version communautaire et payante [1]. Il permet de récupérer des données de sources diverses, de les traiter et de les transformer, puis de les exporter dans un format souhaité vers une destination cible [1].



Intégration de Données avec l'ETL Pentaho

L'intégration de données avec PDI se fait en deux étapes principales :

Création d'une transformation : qui décrit le flux de données pour l'ETL [2]

Exécution de la transformation : pour extraire, transformer et charger les données [1]

PDI offre une interface graphique intuitive pour concevoir les transformations en faisant glisser-déposer des étapes (steps) et en les reliant entre elles [1][3]. Chaque étape peut être configurée individuellement pour effectuer des opérations spécifiques sur les données [1].

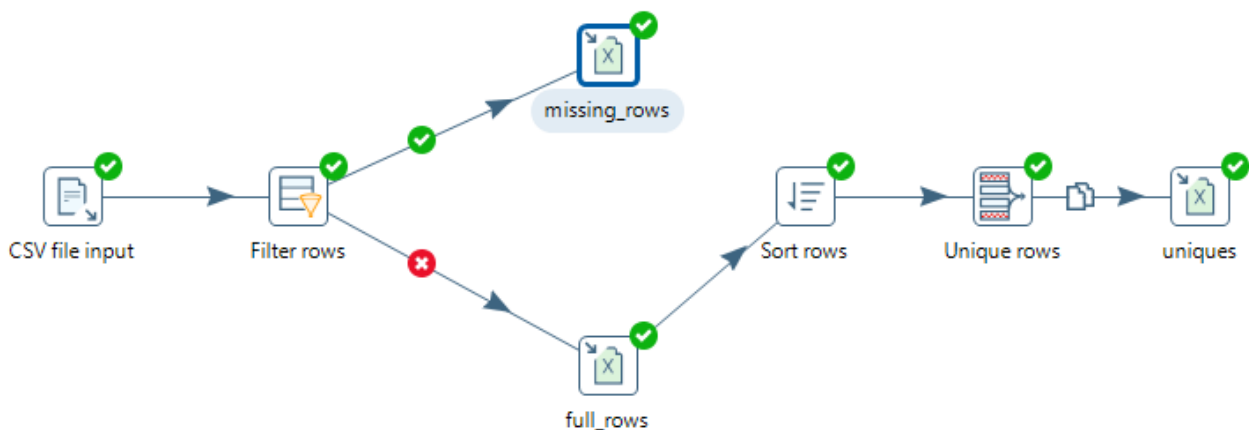
Citations :

[1] <https://www.youtube.com/watch?v=cX2oH6qW1-E>

[2] <https://docs.hitachivantara.com/r/en-us/pentaho-data-integration-and-analytics/9.4.x/mk-95pdia000/getting-started-with-pdi/pentaho-data-integration-pdi-tutorial>

[3] <https://pentahopublic.atlassian.net/wiki/spaces/EAI/pages/364316813/Pentaho%2BData%2BIntegration%2BSteps>

Voici notre transformation :



Etape de la transformation :

1. Chargement des données nécessaire :

Ca concerne le chargement d'un fichier Excel depuis un source local sur notre machine : [logs.xlsx](#)

2. Filtrage des lignes vides et/ou non complet :

Filter rows

Step name:

Send 'true' data to step:

Send 'false' data to step:

The condition:

```
Module = []  
  
AND  
(  
    Bug = []  
    AND  
    (  
        Date = []  
        AND  
        null = []  
    )  
)
```

Help OK Cancel

Dans cette étape, on cherche à détecter les lignes qui possèdent des champs vides, et les séparer dans une autre base de données sous forme d'un fichier Excel « missing_rows.xls »
Les lignes qui ne possèdent aucun champ vide sont collectées dans un autre fichier Excel nommée « full_rows »

3. Triage et préparation pour détection des entrées doublée.

Step name:

Sort directory:

TMP-file prefix:

Sort size (rows in memory):

Free memory threshold (in %):

Compress TMP Files? ☐

Only pass unique rows? (verifies keys only) ☐

Fields:

#	Fieldname	Ascending	Case sensitive compare?	Sort based on current locale?	Collator Strength	Presorted?
1	Num	Y	N	N	0	N
2	Module	Y	N	N	0	N
3	Bug	Y	N	N	0	N
4	Date	Y	N	N	0	N

Cette étape est une préparation au suivant, ceci cherche seulement a trier nos lignes dans notre tableau avec un ordre ascendant.

4. Elimination des entrée double et conservation d'une seule valeur unique.

Step name:

Settings

Add counter to output? ☐ Counter field

Redirect duplicate row ☒ Error description

Fields to compare on (no entries means: compare complete row)

#	Fieldname	Ignore case
1	Num	N
2	Module	N
3	Bug	N
4	Date	N

5. Enregistrement de la version finale transformée.

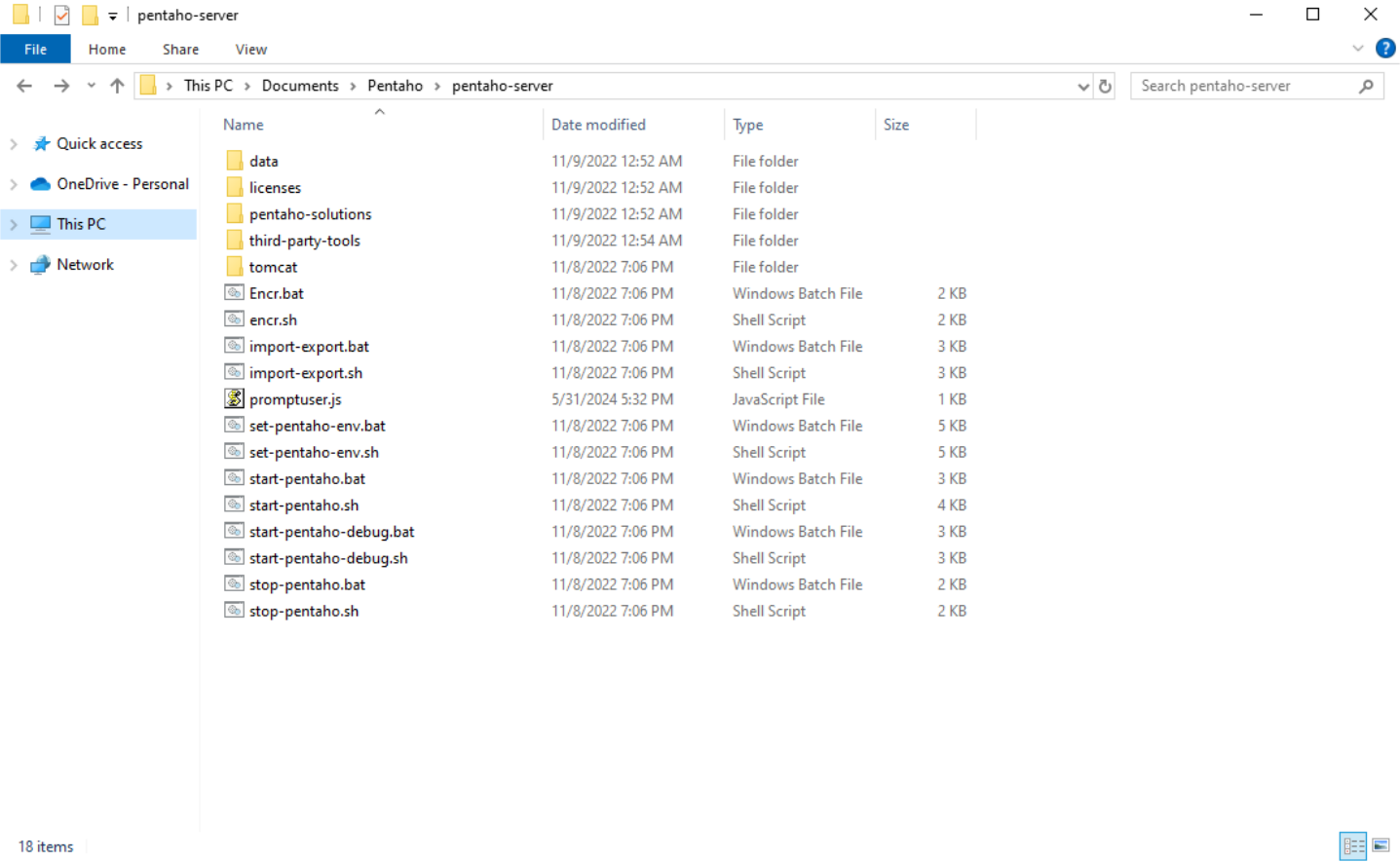
III. Dashboard :

Pour visualiser les données préparer avec Pentaho PDI, on utiliser Pentaho Server
Le serveur Pentaho offre des fonctionnalités puissantes pour créer et déployer des tableaux de bord interactifs et visuellement attrayants. Voici les étapes clés pour créer un tableau de bord avec le serveur Pentaho :

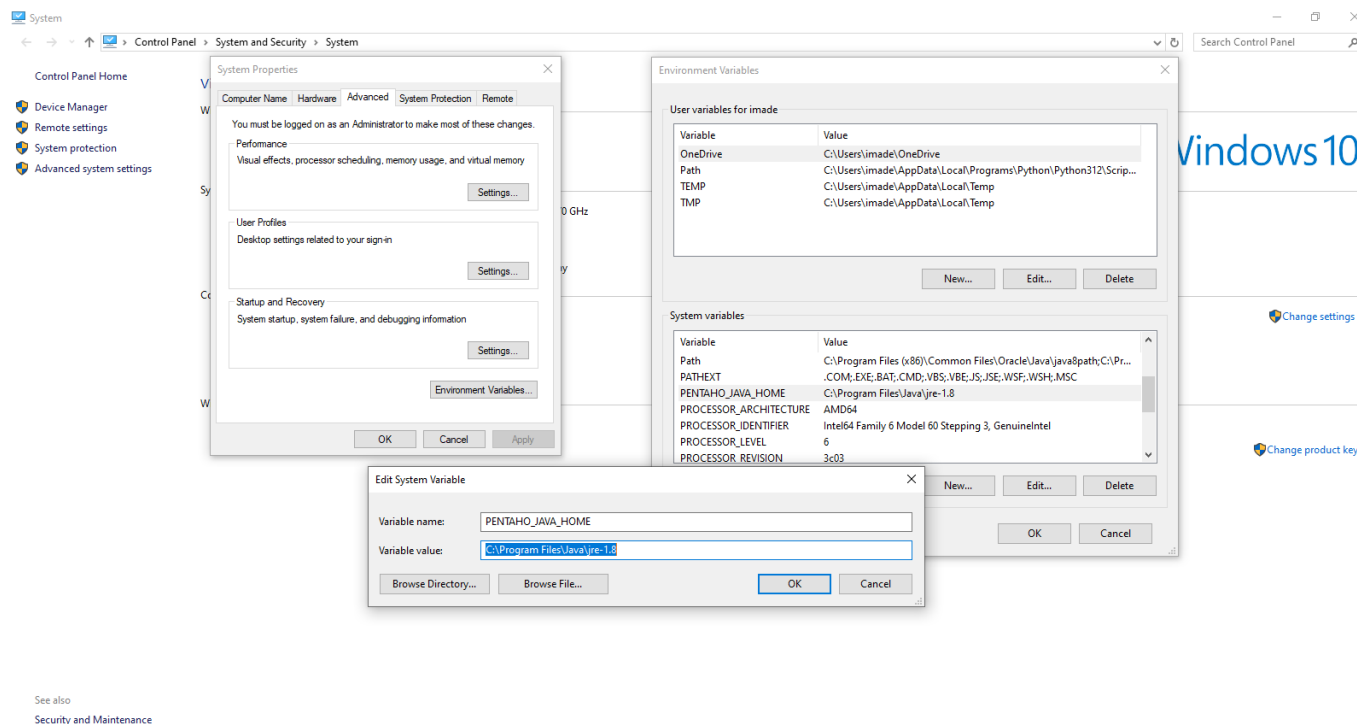
1. Installation et configuration de Pentaho Server

Lien de téléchargement : <https://privatefilesbucket-community-edition.s3.us-west-2.amazonaws.com/9.4.0.0-343/ce/server/pentaho-server-ce-9.4.0.0-343.zip>

On extrait les contenus de fichier ZIP dans en emplacement :



Maintenant il faut configure les variables de système en définissant un variable système Java :



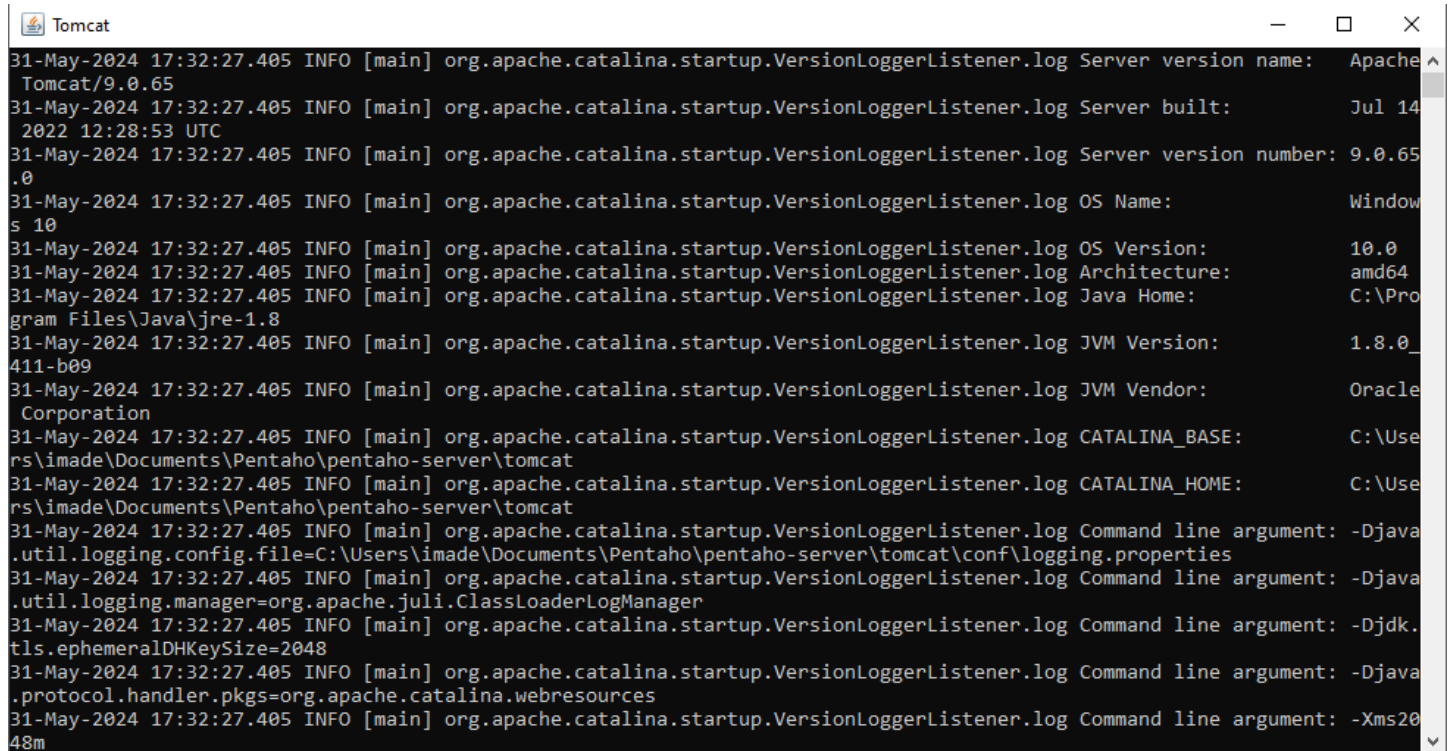
Après avoir effectué correctement ces étapes, on peut lancer le serveur Pentaho (Par défaut le serveur est lancé sur le port 8080 de votre machine)

```
C:\Windows\System32\cmd.exe
Microsoft Windows [Version 10.0.19045.4412]
(c) Microsoft Corporation. All rights reserved.

C:\Users\imade\Documents\Pentaho\pentaho-server>start-pentaho.bat
DEBUG: Using PENTAHO_JAVA_HOME
DEBUG: _PENTAHO_JAVA_HOME=C:\Program Files\Java\jre-1.8
DEBUG: _PENTAHO_JAVA=C:\Program Files\Java\jre-1.8\bin\java.exe
Using CATALINA_BASE: "C:\Users\imade\Documents\Pentaho\pentaho-server\tomcat"
Using CATALINA_HOME: "C:\Users\imade\Documents\Pentaho\pentaho-server\tomcat"
Using CATALINA_TMPDIR: "C:\Users\imade\Documents\Pentaho\pentaho-server\tomcat\temp"
Using JRE_HOME: "C:\Program Files\Java\jre-1.8"
Using CLASSPATH: "C:\Users\imade\Documents\Pentaho\pentaho-server\tomcat\bin\bootstrap.jar;C:\Users\imade\Documents\Pentaho\pentaho-server\tomcat\bin\tomcat-juli.jar"
Using CATALINA_OPTS: "-Xms2048m -Xmx6144m -Dsun.rmi.dgc.client.gcInterval=3600000 -Dsun.rmi.dgc.server.gcInterval=3600000 -Dfile.encoding=utf8 -Djava.locale.providers=COMPAT,SPI -DDI_HOME=C:\Users\imade\Documents\Pentaho\pentaho-server\pentaho-solutions\system\kettle"

C:\Users\imade\Documents\Pentaho\pentaho-server>
```

Une nouvelle fenêtre s'ouvrira :

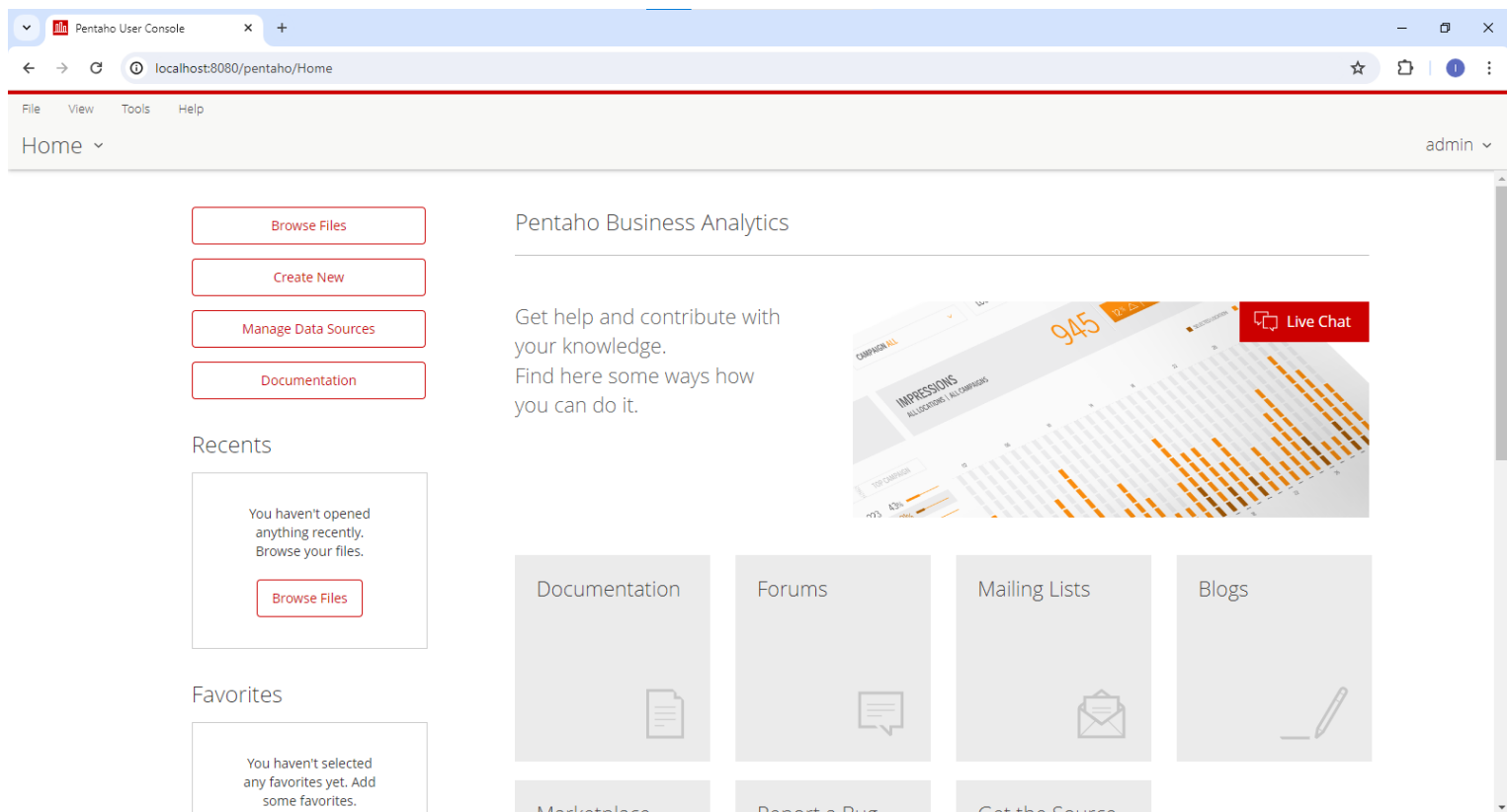


```
31-May-2024 17:32:27.405 INFO [main] org.apache.catalina.startup.VersionLoggerListener.log Server version name: Apache Tomcat/9.0.65
31-May-2024 17:32:27.405 INFO [main] org.apache.catalina.startup.VersionLoggerListener.log Server built: Jul 14 2022 12:28:53 UTC
31-May-2024 17:32:27.405 INFO [main] org.apache.catalina.startup.VersionLoggerListener.log Server version number: 9.0.65
31-May-2024 17:32:27.405 INFO [main] org.apache.catalina.startup.VersionLoggerListener.log OS Name: Windows 10
31-May-2024 17:32:27.405 INFO [main] org.apache.catalina.startup.VersionLoggerListener.log OS Version: 10.0
31-May-2024 17:32:27.405 INFO [main] org.apache.catalina.startup.VersionLoggerListener.log Architecture: amd64
31-May-2024 17:32:27.405 INFO [main] org.apache.catalina.startup.VersionLoggerListener.log Java Home: C:\Program Files\Java\jre-1.8
31-May-2024 17:32:27.405 INFO [main] org.apache.catalina.startup.VersionLoggerListener.log JVM Version: 1.8.0_411-b09
31-May-2024 17:32:27.405 INFO [main] org.apache.catalina.startup.VersionLoggerListener.log JVM Vendor: Oracle Corporation
31-May-2024 17:32:27.405 INFO [main] org.apache.catalina.startup.VersionLoggerListener.log CATALINA_BASE: C:\Users\rs\imade\Documents\Pentaho\pentaho-server\tomcat
31-May-2024 17:32:27.405 INFO [main] org.apache.catalina.startup.VersionLoggerListener.log CATALINA_HOME: C:\Users\rs\imade\Documents\Pentaho\pentaho-server\tomcat
31-May-2024 17:32:27.405 INFO [main] org.apache.catalina.startup.VersionLoggerListener.log Command line argument: -Djava.util.logging.config.file=C:\Users\imade\Documents\Pentaho\pentaho-server\tomcat\conf\logging.properties
31-May-2024 17:32:27.405 INFO [main] org.apache.catalina.startup.VersionLoggerListener.log Command line argument: -Djava.util.logging.manager=org.apache.juli.ClassLoaderLogManager
31-May-2024 17:32:27.405 INFO [main] org.apache.catalina.startup.VersionLoggerListener.log Command line argument: -Djdk.tls.ephemeralDHKeySize=2048
31-May-2024 17:32:27.405 INFO [main] org.apache.catalina.startup.VersionLoggerListener.log Command line argument: -Djava.protocol.handler.pkgs=org.apache.catalina.webresources
31-May-2024 17:32:27.405 INFO [main] org.apache.catalina.startup.VersionLoggerListener.log Command line argument: -Xms2048m
```

Ne ferme pas celle-ci, elle est responsable de fonctionnement de serveur, il peut prendre un peu du temps pour l'installation et configuration mais après une certaine période nous pourrons accéder à notre serveur sur : **localhost :8080**

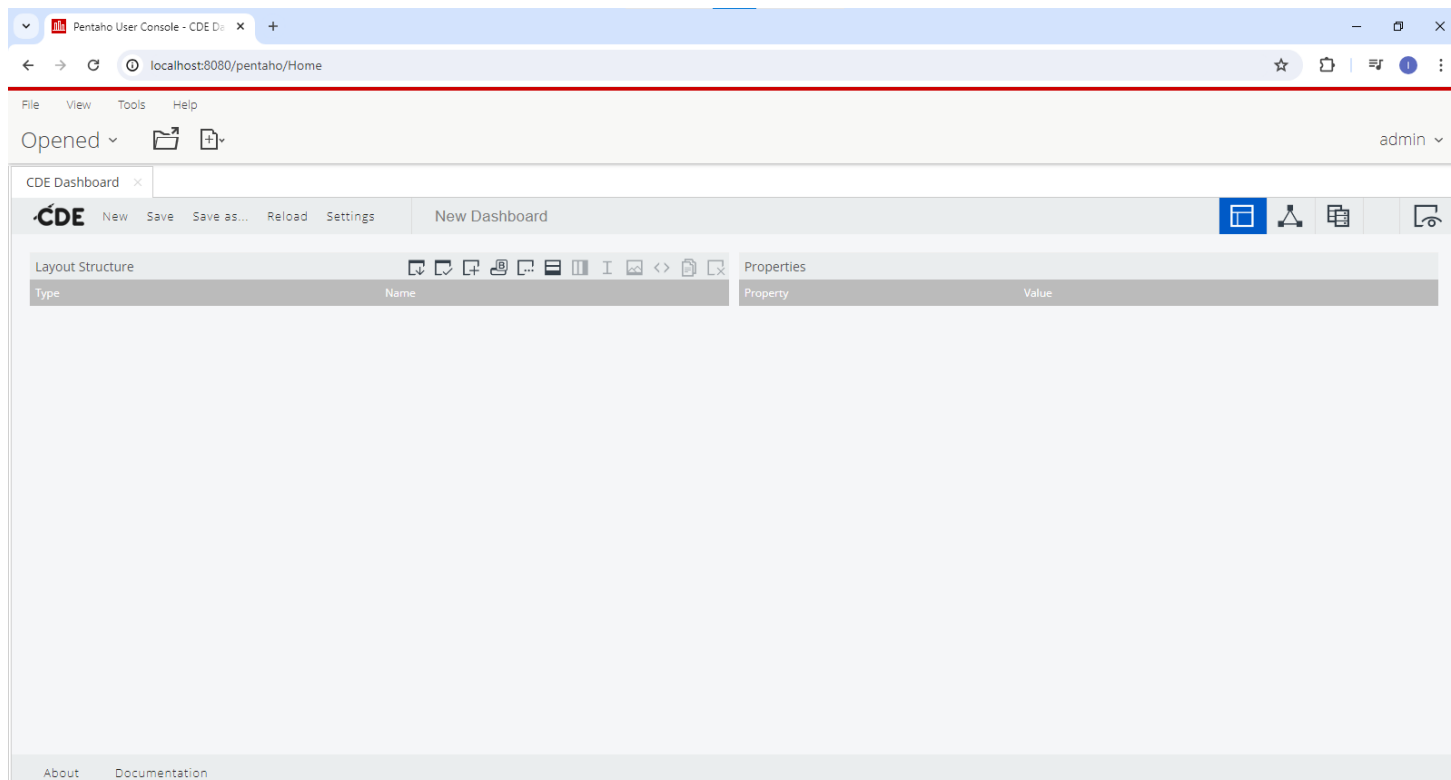
2. Connexion au Serveur Pentaho

- Ouvrez un navigateur web et accédez à l'URL du serveur Pentaho : localhost :8080
- Connectez-vous avec vos identifiants d'utilisateur



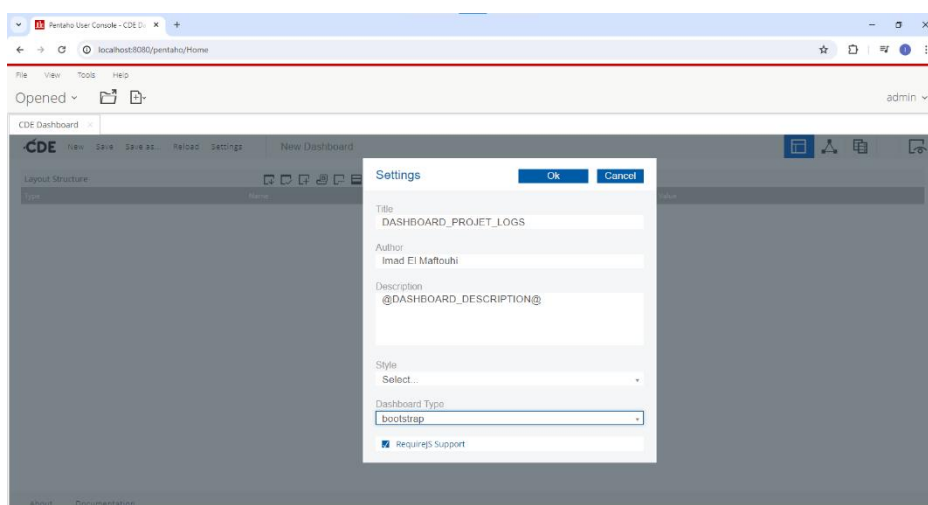
3. Création d'un Nouveau Tableau de Bord

- Dans la barre de lancement rapide de la console Pentaho, cliquez sur "Nouveau tableau de bord"
- Alternativement, vous pouvez cliquer sur Fichier > Nouveau > Tableau de bord ou sur l'icône "Nouveau tableau de bord" dans la barre d'outils
- La page "Nouveau tableau de bord" apparaît

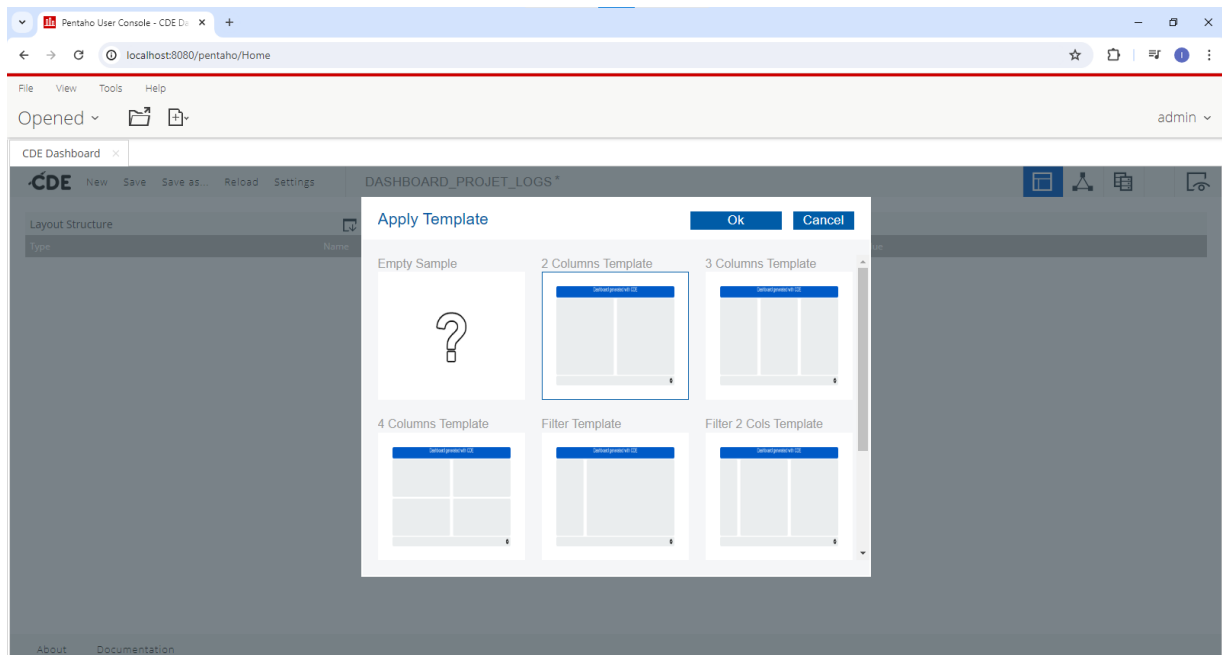


4. Configuration du Tableau de Bord

- Dans le volet d'édition (partie inférieure de la page), cliquez sur "Propriétés" et entrez un titre pour votre tableau de bord dans la zone de texte "Titre de la page"



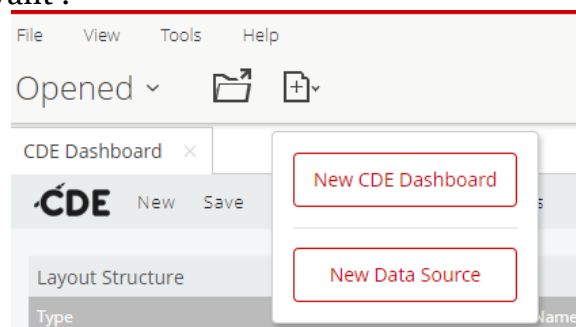
- Cliquez sur "Modèles" pour sélectionner une mise en page pour votre tableau de bord



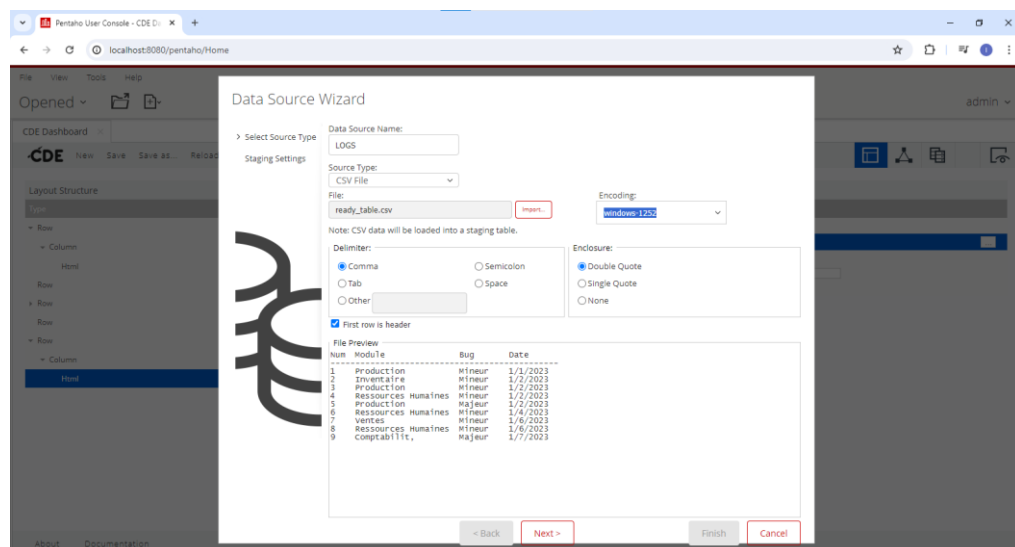
- Un tableau de bord vierge avec la mise en page sélectionnée apparaît
- Cliquez sur "Thème" pour sélectionner un thème pour votre tableau de bord
- Le thème sélectionné est appliqué à votre tableau de bord

5. Importation des données :

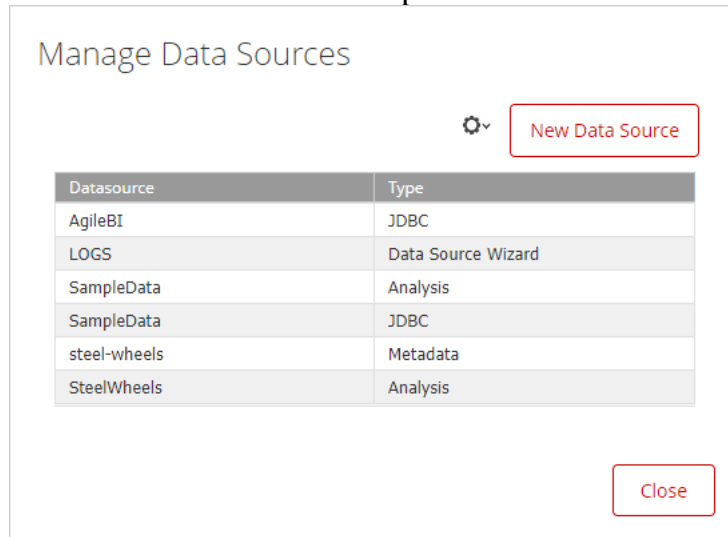
- Naviguez vers l'option suivant :



- Importer vos fichiers à l'aide de cette forme :

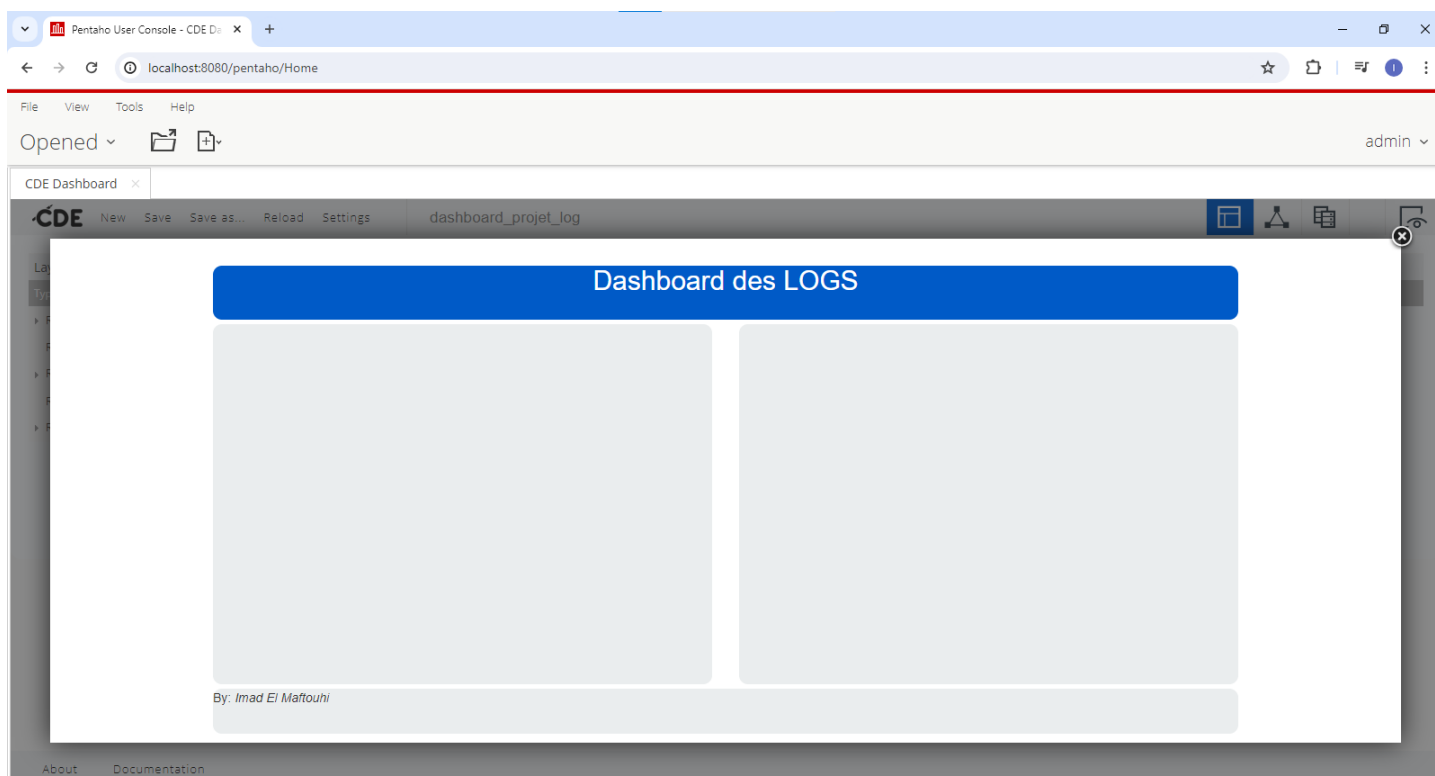


Finalement nous auront nos bases de données importées sur nos serveur :



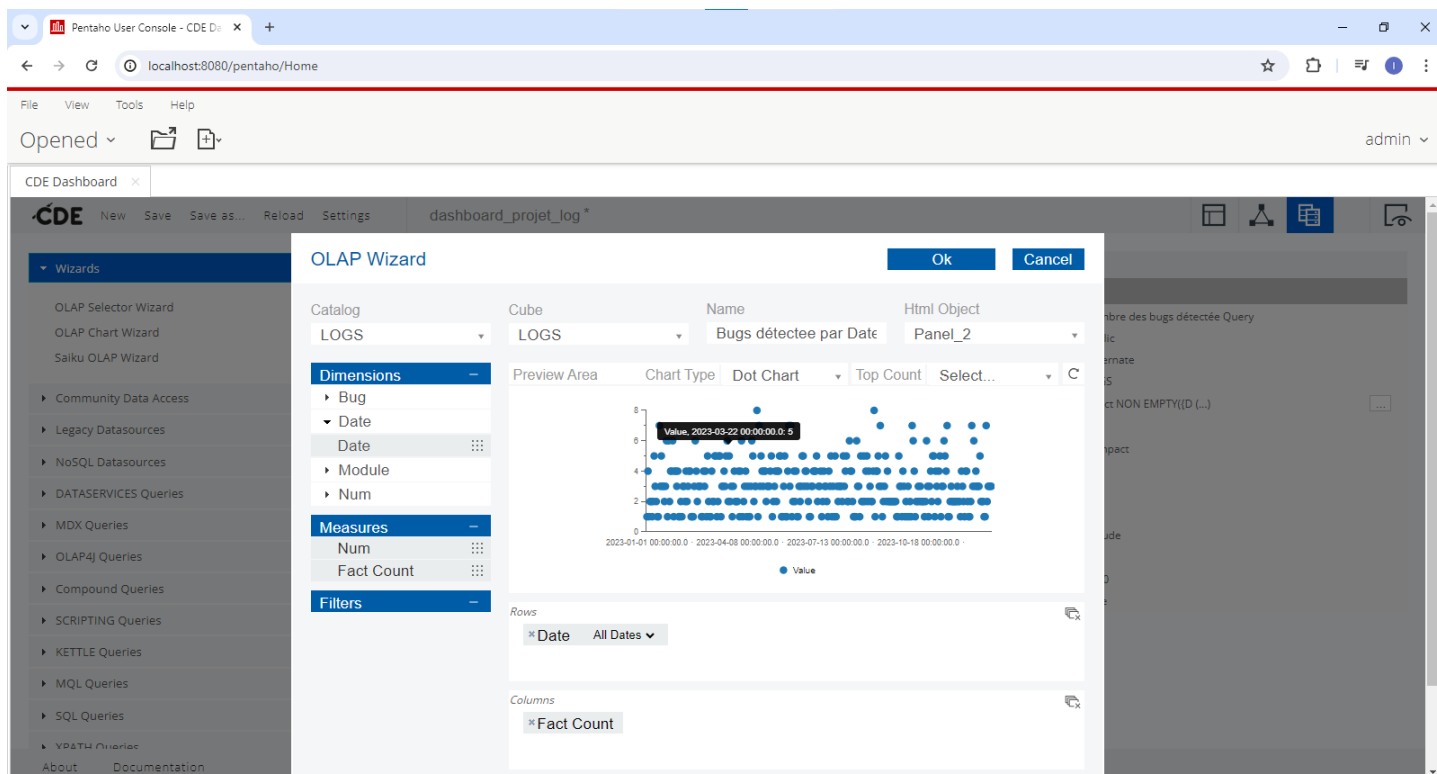
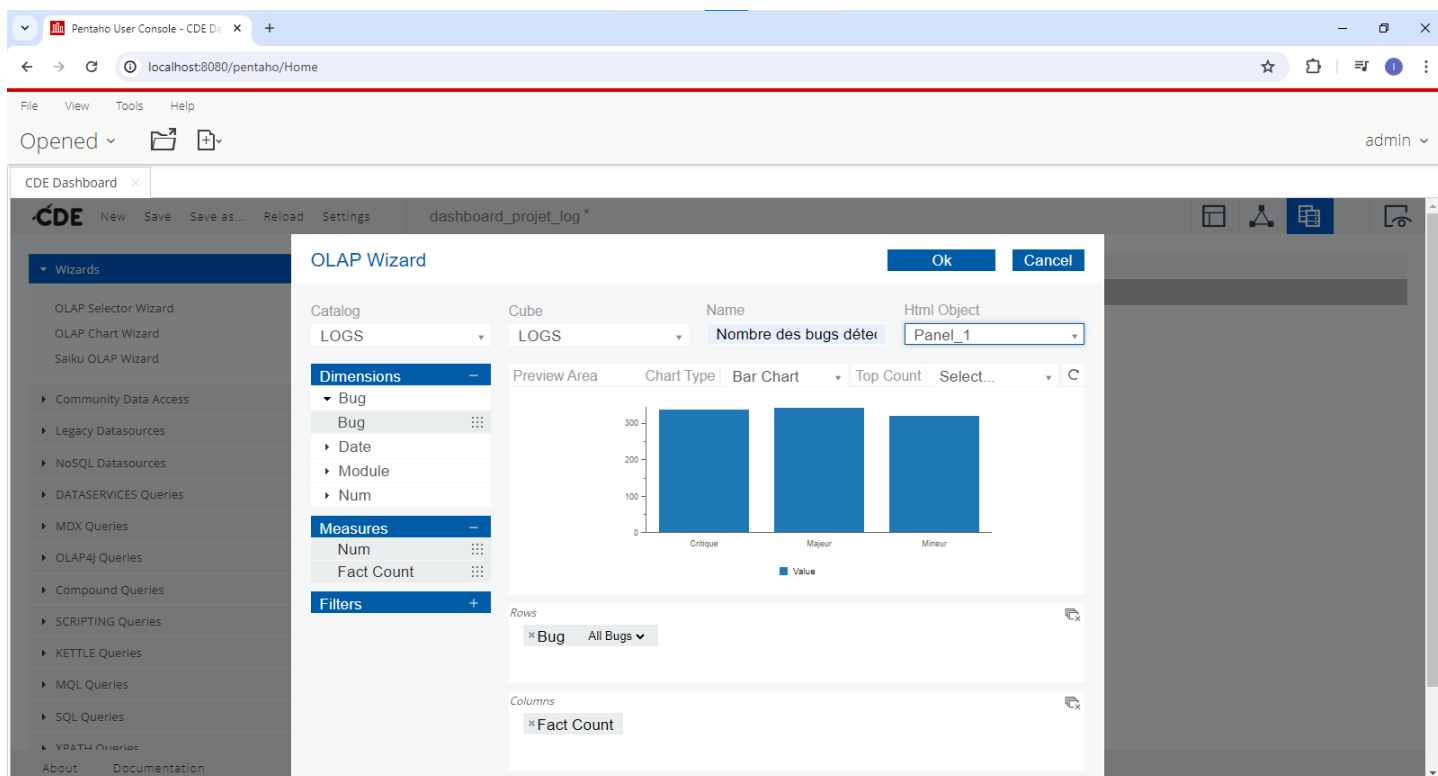
6. Ajout de Contenu

- Vous pouvez ajouter différents types de contenu à votre tableau de bord, comme des graphiques, des tableaux de données, des URLs, etc.



- Utilisez l'outil Chart Designer pour créer des graphiques personnalisés

- Configurez les filtres dynamiques pour permettre aux utilisateurs de contrôler le contenu du tableau de bord



IV. Conclusion

En suivant ces étapes, vous pouvez créer des tableaux de bord interactifs et visuellement attrayants avec le serveur Pentaho, en tirant parti de ses fonctionnalités de création de rapports et de visualisation des données.