Projet étudiant
MASTER II

DATA SCIENCE & ANALYSE DE DONNEES

CONSTRUCTION D'UN LAC DE DONNÉES AVEC APACHE SPARK

Proposé par :

M. MOUNGOLE ARISTIDE

Réalisé par :

HYONTA KENGAP BLERIOT

Année : 2024/2025

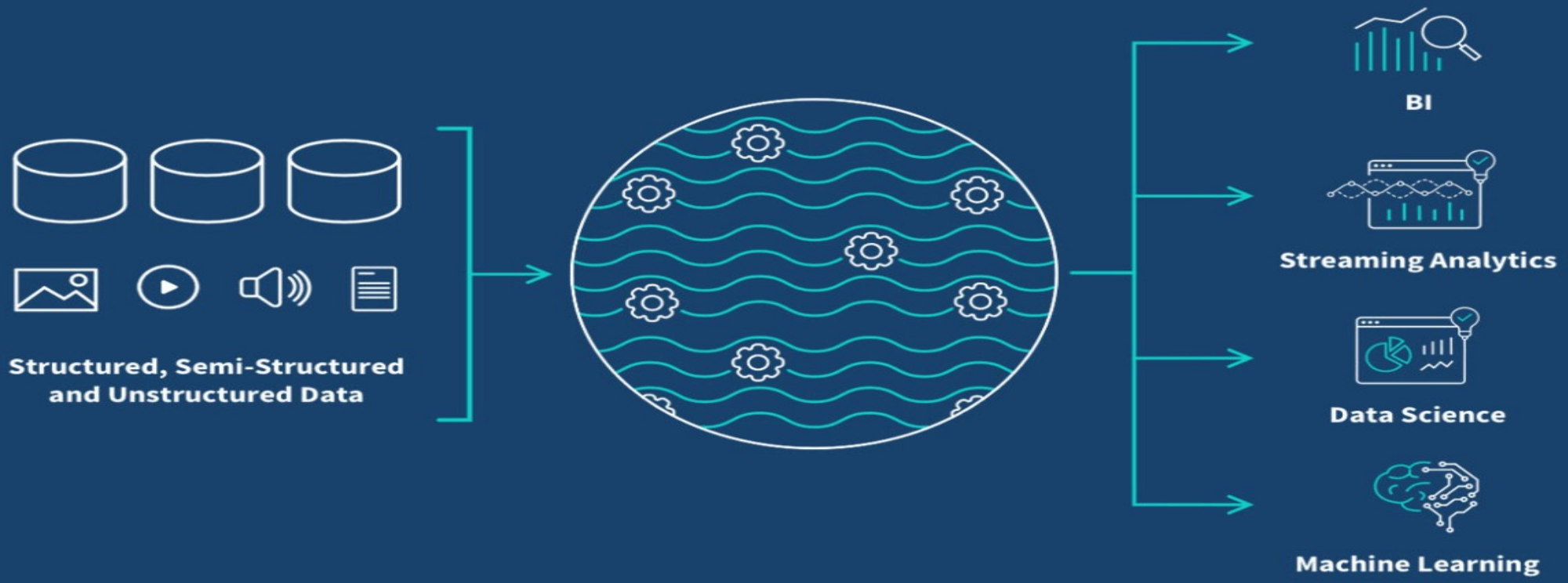

Structured, Semi-Structured
and Unstructured Data
BI
Streaming Analytics
Data Science
Machine Learning

Table des matières

Introduction au Projet

Présentation Générale du Projet

Ce projet vise à construire un lac de données en utilisant Apache Spark pour centraliser et analyser des données provenant de diverses sources.

Il répond aux défis posés par la croissance exponentielle des données dans des secteurs clés comme la finance et le commerce.

Contexte et Problématique

Croissance des Données

Augmentation exponentielle des données générées dans divers secteurs.

Nécessité d'une gestion efficace des données pour soutenir la prise de décision.

Défis pour les Entreprises

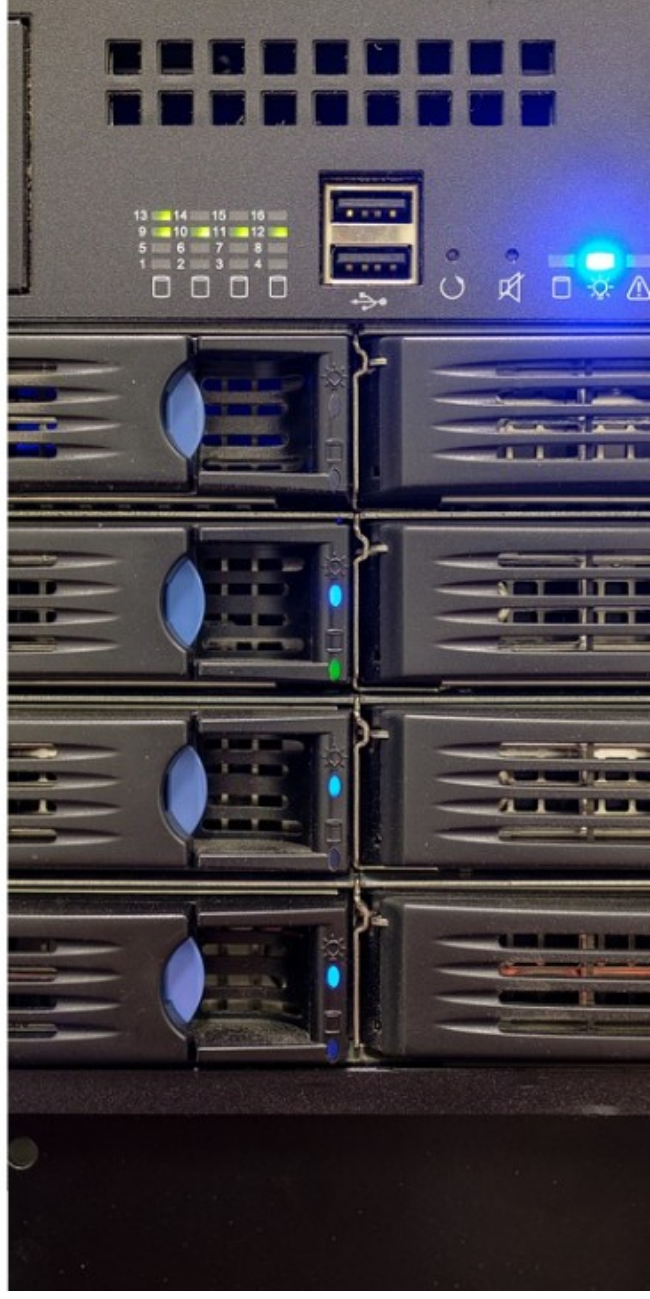
Sécurité des données face à la croissance rapide des volumes.

Conformité réglementaire et intégration des systèmes existants.

Centralisation des Informations

Avantages d'une base de données centralisée pour une meilleure accessibilité.

Facilitation de l'analyse des données pour des décisions éclairées.



Objectifs du Projet

Conception du lac de données

Concevoir un lac de données capable de stocker des informations brutes hétérogènes issues de bases SQL, de fichiers logs et d'avis clients.

Pipeline d'ingestion et transformation

Mettre en place un pipeline d'ingestion et de transformation avec Apache Spark.

Analyse et tableaux de bord interactifs

Analyser ces données en vue de produire des tableaux de bord interactifs pour faciliter l'aide à la décision.

Introduction au Lac de Données

Stockage Distribué

Un lac de données est un système de stockage massivement distribué, permettant de gérer de grandes quantités de données. Cette architecture facilite l'accès et la gestion des données provenant de sources variées.

Flexibilité Analytique

Les lacs de données offrent une flexibilité analytique, permettant d'analyser les données sans nécessiter un schéma prédéfini. Cela permet une exploration plus libre et dynamique des données.

Données Hétérogènes

Ils prennent en charge des données hétérogènes, incluant des données structurées, semi-structurées et non structurées. Cette diversité enrichit les possibilités d'analyse et d'interprétation.

Volume Élevé

Les lacs de données sont conçus pour gérer des volumes élevés de données, ce qui est essentiel dans le contexte actuel de croissance exponentielle des données. Cela permet aux entreprises de centraliser et d'exploiter efficacement leurs informations.

Absence de Schéma Prédéfini

L'absence de schéma prédéfini dans un lac de données permet une plus grande agilité dans le traitement des données. Cela favorise l'innovation et l'adaptation rapide aux besoins changeants des utilisateurs.

Technologies Utilisées

Apache Spark

Apache Spark est une plateforme de traitement distribué utilisée pour le traitement des données massives.

Elle permet d'ingérer, de nettoyer et d'analyser efficacement des données provenant de diverses sources.

MySQL & PostgreSQL

MySQL et PostgreSQL sont des systèmes de gestion de bases de données relationnelles.

Ils sont utilisés pour le stockage des données transactionnelles et l'extraction d'informations clients.

Formats de Données

Les formats de données tels que Parquet et JSON sont utilisés pour le stockage efficace dans le lac de données.

Ces formats permettent de gérer des données structurées et non structurées.

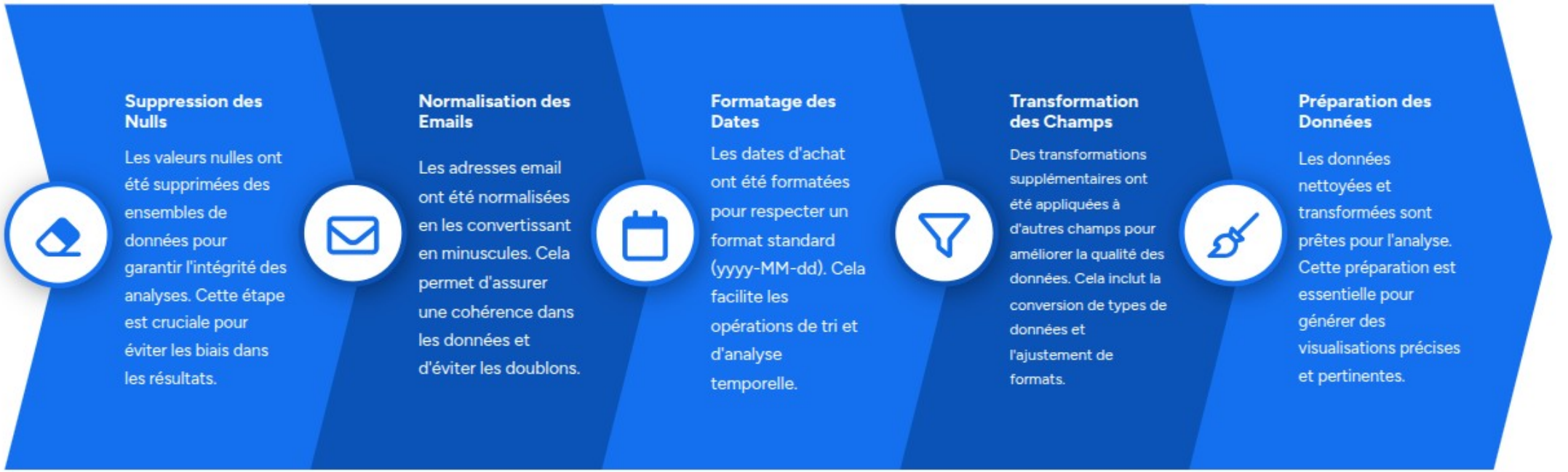
Outils de Visualisation des Données

Matplotlib et Pandas sont des bibliothèques Python utilisées pour l'analyse et la visualisation des données.

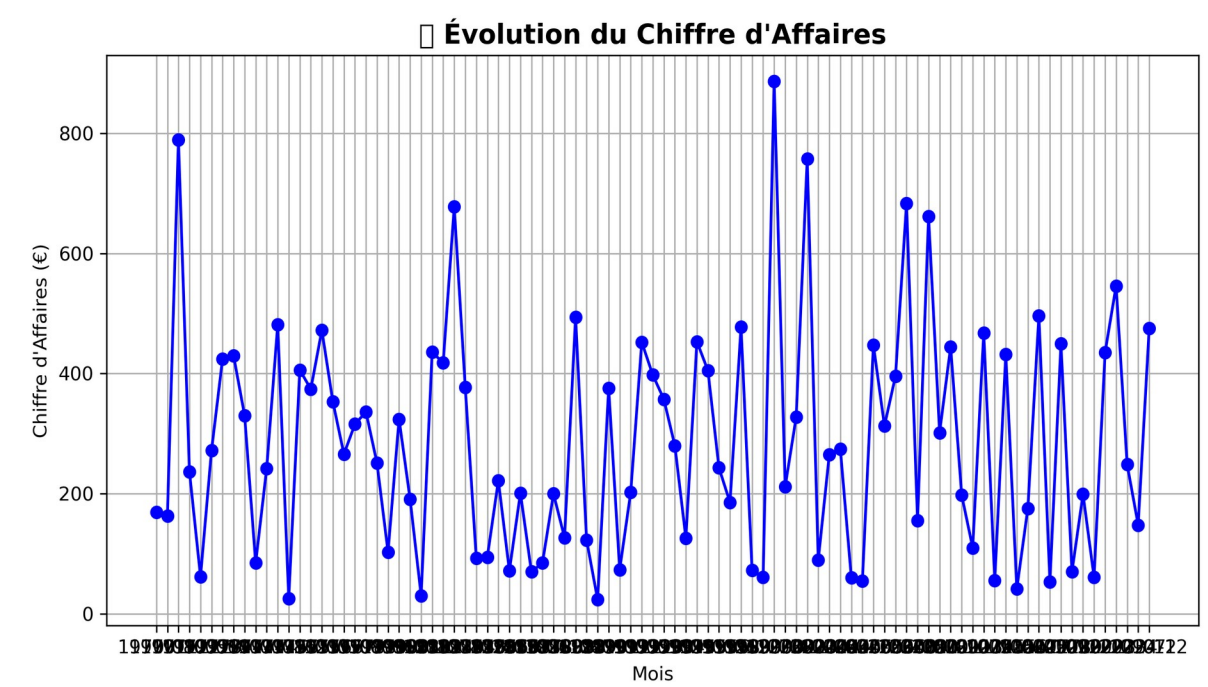
Jupyter Notebook offre un environnement interactif pour la présentation des résultats d'analyse.

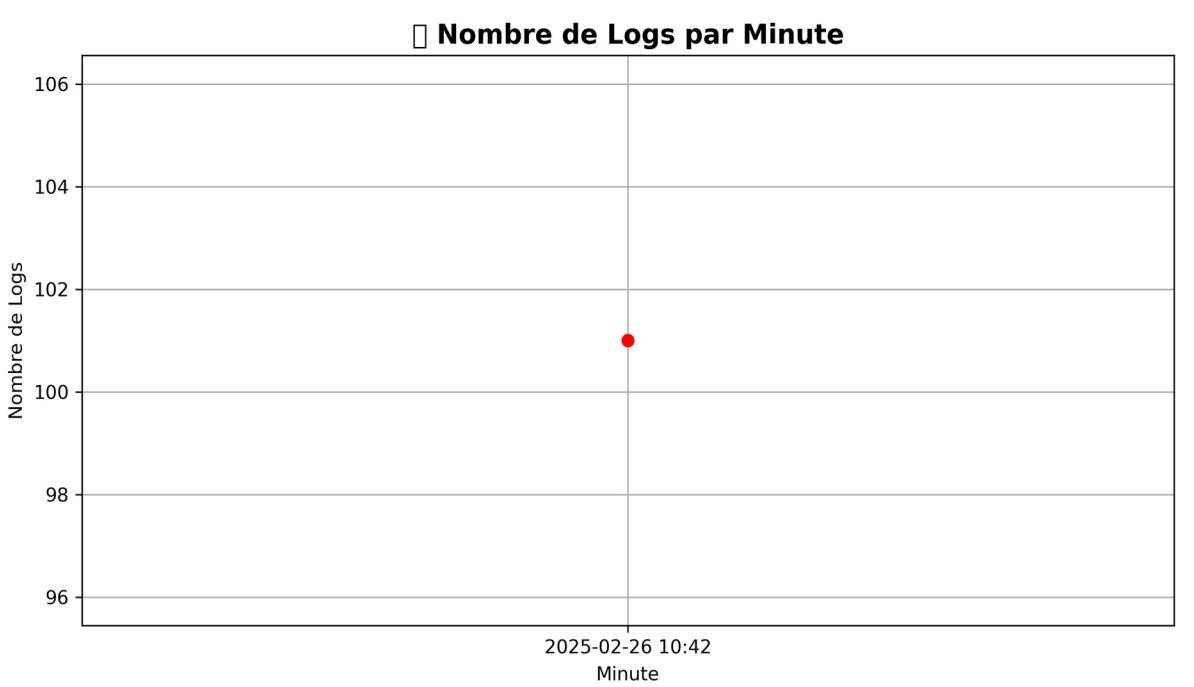
Ingestion des Données

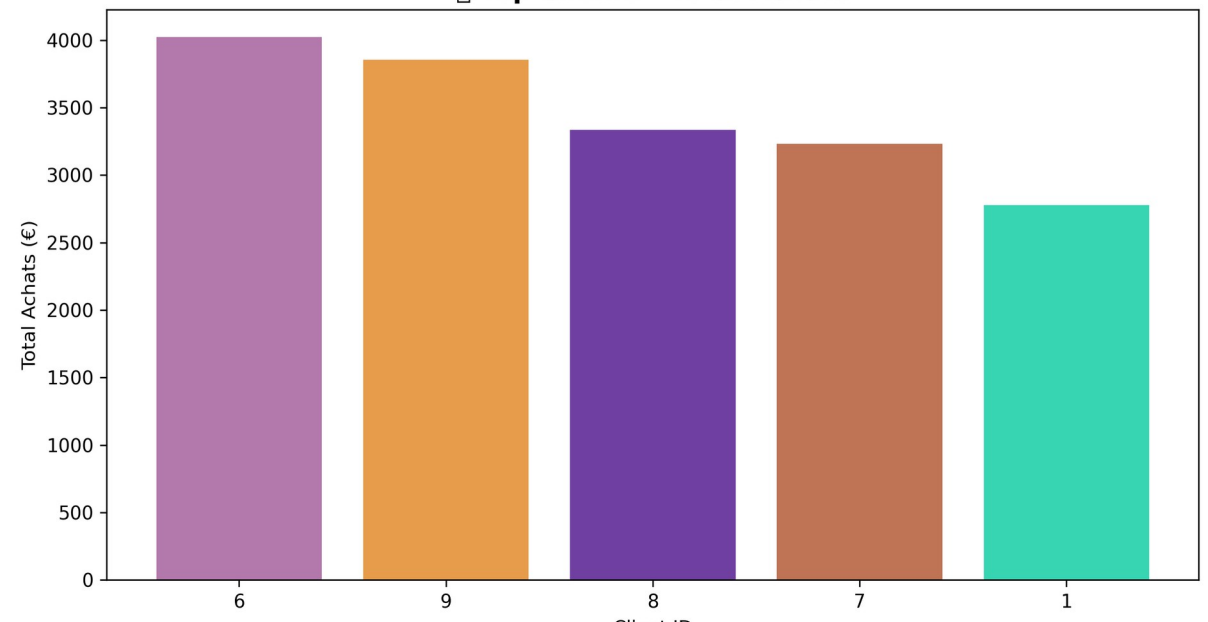
Nettoyage et Transformation des Données

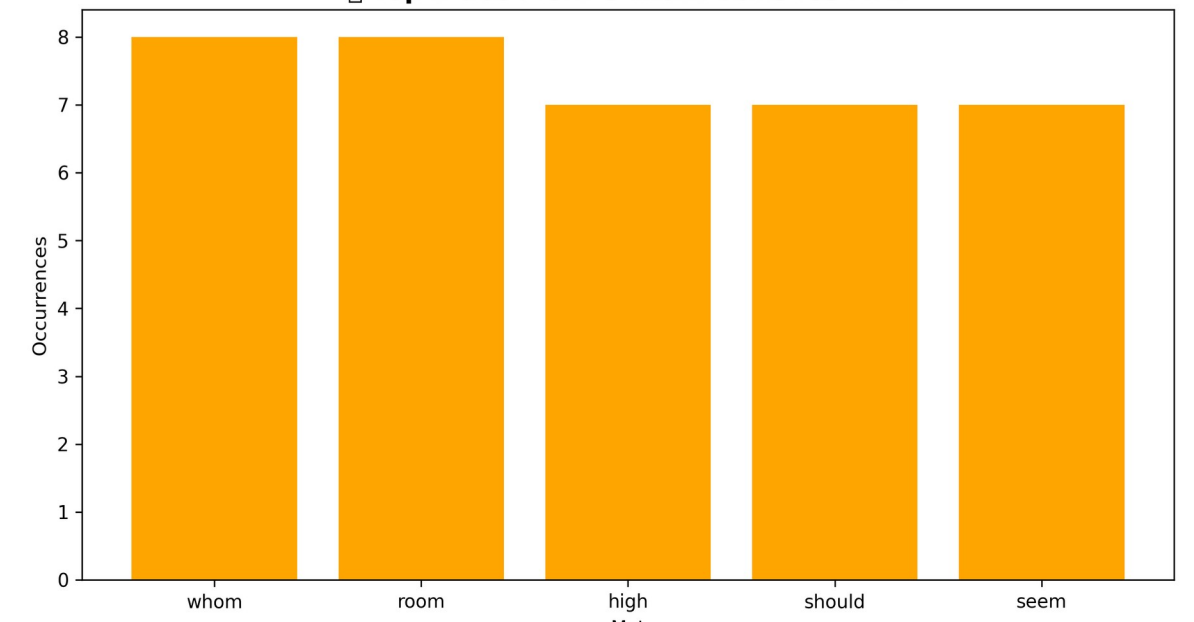


Analyse et Visualisation des Données


Évolution du Chiffre d'Affaires
Chiffre d'Affaires (€)
0
200
400
600
800
Mois


Nombre de Logs par Minute
Nombre de Logs
96
98
100
102
104
106
2025-02-26 10:42
Minute


Top Clients - Total Achats
Total Achats (€)
0
500
1000
1500
2000
2500
3000
3500
4000
6
9
8
7
1


Top 5 des Mots Utilisés dans les Avis
Occurrences
0
1
2
3
4
5
6
7
8
whom
room
high
should
seem

Conclusion et Perspectives

Résumé des Résultats du Projet

Le projet a démontré l'efficacité d'un lac de données pour le traitement de grandes quantités d'informations variées.

Efficacité du Lac de Données

Le lac de données a prouvé sa capacité à gérer des données hétérogènes de manière efficace.

Intégration de Hadoop

Hadoop améliorera la gestion des flux de données en temps réel.

Utilisation de Kafka

Kafka facilitera la gestion des données en continu, augmentant ainsi la réactivité du système.

Perspectives d'Amélioration

Nous envisageons d'automatiser le pipeline de traitement et de développer une interface web intuitive.

Mercie pour votre attention