Table des matières

[I. Introduction 2](#_Toc126160997)

[A. Contexte 2](#_Toc126160998)

[B. Définition du problème 3](#_Toc126160999)

[II. Etat de l’art 4](#_Toc126161000)

[A. **Définition de la maintenance prédictive** 4](#_Toc126161001)

[B. **Revue de littérature** 4](#_Toc126161002)

[C. **Importance de la maintenance prédictive dans le data management** 5](#_Toc126161003)

[D. **Méthodes de machine learning utilisées dans la maintenance prédictive** 6](#_Toc126161004)

[E. **Limites actuelles de l'utilisation du machine learning dans la maintenance prédictive** 8](#_Toc126161005)

[III. Méthodologie 9](#_Toc126161006)

[A. Collecte de la donnée 9](#_Toc126161007)

[1. Source de données : 9](#_Toc126161008)

[2. Architecture de la solution 10](#_Toc126161009)

[3. Extraction de la donnée 11](#_Toc126161010)

[4. Préparation des données d'entrée du modèle 12](#_Toc126161011)

[5. **Modification de l’index** 13](#_Toc126161012)

[6. **Définition de la fréquence** 13](#_Toc126161013)

[7. **Filtrage des données** 14](#_Toc126161014)

[B. **Analyse exploratoire des données** 14](#_Toc126161015)

[1. **Visualisation de notre série et premières interprétations** 14](#_Toc126161016)

[2. **White Noise (Bruit Blanc)** 15](#_Toc126161017)

[3. **La stationnarité** 15](#_Toc126161018)

[4. **Décomposition de la Série temporelle** 17](#_Toc126161019)

[5. **Saisonnalité** 17](#_Toc126161020)

[6. **Tendance** 18](#_Toc126161021)

[7. **L’autocorrélation** 18](#_Toc126161022)

[8. Préparation des données 19](#_Toc126161023)

 3- Méthodologie

* Collecte de la donnée
* Description des données utilisées
* Description de la méthode de machine Learning utilisée
* Validation de la méthode de machine Learning

4- Résultats

* Construction du modèle
* Déploiement
* Evaluation

5- Resultats

* Présentation des résultats obtenus
* Comparaison avec les méthodes de maintenance prédictive existantes
* Discussions sur les résultats obtenus et améliorations 1

# Introduction

## Contexte

L'équipe data joue un rôle crucial dans toute entreprise en garantissant la distribution et la valorisation des données. Chez Boursorama, cette équipe se compose de 40 personnes réparties en trois sous-équipes distinctes. La première équipe, appelée Business Intelligence, est responsable de l'analytique et du reporting. La seconde équipe, appelée Data Hub, est chargée de traiter les données Big Data. Enfin, la troisième équipe, appelée Data Lake, s'occupe des traitements en matière de science des données.

Pour garantir une distribution efficace des données, l'équipe data doit intégrer des sources de données diverses, telles que des partenaires externes, des APIs ou des données internes. Cette intégration est gérée et automatisée grâce à des processus ETL. Chez Boursorama, c'est l'équipe Business Intelligence qui gère l'ensemble de la partie ETL. Un processus ETL est un processus qui permet de traiter les données pour les rendre prêtes à être analysées

 Depuis 2015, Microsoft SSIS est l'outil ETL utilisé par l'équipe data pour traiter les données. Le point central de stockage des données traitées par SSIS est le Data Warehouse, qui est également géré par l'équipe Business Intelligence. Les reportings, les tableaux de bord et les rapports sont produits à partir des données présentes dans le Data Warehouse.Ces rapports produits par l'équipe Business Intelligence sont importants pour les activités quotidiennes de plusieurs métiers, notamment la conformité, la lutte contre la fraude, le marketing, la finance, etc. Les rapports sont conçus de manière à être envoyés automatiquement par courriel à intervalles hebdomadaires, quotidiens, mensuels, etc. à l'aide de l'outil Cognos Analytics. Les tableaux de bord, quant à eux, sont créés à l'aide de l'outil Tableau

 A mesure que le temps passe, les entreprises, y compris Boursorama, deviennent de plus en plus axées sur les données, ce qui signifie que les métiers dépendent de plus en plus des équipes de données, en particulier de l'équipe de BI. Cette forte dépendance se manifeste par une augmentation des demandes de rapports et de tableaux de bord. Pour répondre à ces nouvelles demandes, il est nécessaire de créer de nouvelles tables dans le datawarehouse et de récupérer des données dans de nouvelles sources.   
Les besoins en intégration de nouvelles données se traduisent par la création de nouveaux packages SSIS, qui sont des fichiers contenant une série de tâches automatisées pour transférer et transformer les données    
Le processus d'intégration des nouvelles données comprend la création de nouveaux packages SSIS et leur mise en production via un logiciel appelé dollarUniverse. Ce dernier est responsable du lancement et de l'ordonnancement des paquets en fonction de leur dépendance les uns des autres. Par exemple, si un paquet dépend de l'exécution réussie d'un ou plusieurs autres, il doit être programmé correctement dans dollarUniverse en conséquence   
 Lors de la création de paquets SSIS, ils sont configurés pour générer des journaux. Les journaux permettent de surveiller l'état du système à chaque instant, avec plusieurs types de journaux (erreur, alerte, information) enregistrés dans une base de données relationnelle SQL Server en temps réel pour chaque événement. Cependant, comme tout système n'est pas parfait, les paquets SSIS peuvent souvent rencontrer des erreurs, telles que des erreurs de connectivité, de surcharge de performance, et des erreurs liées à l'absence de données attendues d'un autre paquet ou source externe.   
Les paquets SSIS lors de leur création sont programmés pour sortir des logs. Les logs permettent de connaitre l’état du système à chaque seconde. Logs produits sont de plusieurs types (Erreur, Alertes, Informations) pour chaque évènement, l’état du système est enregistré dans une base de données relationnelle SQL Server en temps réel.     
Comme aucun système n’est parfait, les packages SSIS rencontrent très souvent des erreurs. Les erreurs sont variées, les principales sont celles liées à la connectivité, d’autres liées à la surcharge des performances, d'autres erreurs liées à l’absence de données qui sont censés être fournies par un autre package ou une source externe    
Ces erreurs dans les packages SSIS peuvent entraîner des retards dans les processus métier, car elles peuvent empêcher la production de rapports et la mise à jour des tableaux de bord. Les erreurs sont signalées aux équipes de Business Intelligence par l'intermédiaire de DollarUniverse et sont généralement corrigées manuellement par ces équipes. La correction des erreurs peut prendre plusieurs minutes car il est nécessaire de détecter l'origine de l'erreur et de la corriger ensuite.

## Définition du problème

L'échec des packages SSIS peut entraîner des retards considérables dans les activités de l'équipe de business intelligence et les processus métiers. La correction de ces erreurs peut nécessiter plusieurs minutes, voire des heures, en raison de divers facteurs, tels que l'utilisation de ressources par une équipe externe ou l'absence de données fournies par des sources externes. Afin de prévenir les interruptions système, il est important de prédire les échecs potentiels des packages SSIS.

En utilisant des algorithmes de machine learning pour analyser les données enregistrées dans les logs, il est possible de détecter des tendances ou des anomalies qui peuvent indiquer un risque d'échec futur. Cela permettrait aux équipes de prendre des mesures proactives pour éviter les erreurs et les retards, ce qui améliorerait la fiabilité et la qualité du système.

Comment Est-ce que L'utilisation de l'intelligence artificielle peut être bénéfique pour anticiper les erreurs de packages SSIS ?

# Etat de l’art

## **Définition de la maintenance prédictive**

La maintenance prédictive est une stratégie de maintenance qui vise à anticiper et à prévenir les défaillances des équipements avant qu'elles ne se produisent. Cette approche s'appuie sur l'analyse de données et de modèles mathématiques pour prédire l'état futur des équipements et détecter les problèmes potentiels avant qu'ils ne surviennent. Elle se différencie de la maintenance corrective, qui intervient après la survenue d'une défaillance, et de la maintenance préventive, qui suit un calendrier de maintenance prédéterminé indépendamment de l'état de l'équipement.

La maintenance prédictive utilise des techniques de prévision et de diagnostic avancées pour optimiser le temps de fonctionnement des équipements en prévenant les pannes avant qu'elles ne se produisent. Elle peut également aider à réduire les coûts de maintenance à long terme en permettant de planifier les interventions de maintenance de manière plus efficace et en réduisant le temps passé à réparer les équipements en panne. De nombreux outils et technologies sont utilisés dans la maintenance prédictive, tels que les capteurs de données, les systèmes de surveillance en temps réel, les logiciels de diagnostic et les algorithmes de machine learning.

## **Revue de littérature**

 La maintenance prédictive est un domaine qui utilise des techniques d'analyse de données pour identifier les défaillances potentielles avant qu'elles ne se produisent, afin d'éviter les temps d'arrêt non planifiés et de prolonger la durée de vie des équipements. Elle est devenue de plus en plus importante pour les entreprises, car elle permet de réduire les coûts de maintenance, d'augmenter la disponibilité des équipements et de prolonger leur durée de vie. Cependant, la maintenance prédictive est un domaine complexe qui nécessite une grande quantité de données de qualité et des compétences en analyse de données pour être efficace.

Il y a plusieurs études qui ont montré l'efficacité de l'utilisation de techniques de machine learning pour la maintenance prédictive en Data management. Une publication scientifique de l’entreprise [Procedia manufacturing (2018)](https://www.sciencedirect.com/science/article/pii/S2351978918305262) démontre l'utilisation du machine learning, plus précisément des méthodes de clustering, pour classer des machines d'impression 3D en clusters basés sur leur comportement et ainsi prévoir les futures pannes.   
D'autres études, comme celle de [Awsan Mohammed dans son livre Engineering Failure analysis(2023)](https://www.sciencedirect.com/science/article/abs/pii/S1350630722009864), propose Un modèle basé sur la régression  pour prédire les défaillances des pompes à eau de mer dans l'industrie pétrolière et gazière.

Il y a aussi des études qui se concentrent sur la maintenance prédictive pour des données spécifiques comme la maintenance des moteurs , les données de capteur ou la maintenance dans les industries de processus . Ces études ont montré qu'il est possible d'utiliser des techniques de machine learning pour résoudre des problèmes spécifiques de maintenance prédictive dans différents domaines.

Il y a différents types de données utilisées pour la maintenance prédictive, comme les données de fonctionnement, les données de capteurs, les données de maintenance, les données climatiques et les données de production. Ces données peuvent être collectées à l'aide de capteurs connectés, d'enregistreurs de données et d'autres dispositifs de surveillance, et doivent être stockées et gérées de manière efficace pour pouvoir être utilisées pour la maintenance prédictive.

Plusieurs challenges sont liés à la mise en place de la maintenance prédictive basée sur l'apprentissage automatique, comme la qualité et la quantité des données, la complexité des modèles, les défis liés à l'intégration des systèmes et la gestion des défaillances réelles. Il est donc important d'avoir une équipe multidisciplinaire et une infrastructure adéquate pour mettre en œuvre une telle approche.

 La maintenance prédictive peut nécessiter une analyse des données en temps réel, avoir une infrastructure adéquate pour collecter et traiter les données en temps réel est la case prioritaire à cocher. Ainsi, les entreprises peuvent utiliser des technologies comme l'IoT (Internet des objets) pour connecter les équipements et collecter les données en temps réel.

Il est important également de mentionner les avantages de la maintenance prédictive par rapport à d'autres types de maintenance. La maintenance corrective réactive se concentre sur les problèmes après qu'ils se sont produits, tandis que la maintenance préventive planifiée est basée sur un calendrier fixe et ne tient pas compte des variations de la condition physique des équipements. La maintenance prédictive, d'autre part, utilise des données en temps réel pour identifier les défaillances potentielles avant qu'elles ne se produisent, et donc permet une maintenance plus efficace et une meilleure utilisation des ressources.

En somme, la maintenance prédictive des processus de gestion de données est un domaine en évolution qui utilise des techniques de machine learning pour identifier les défaillances potentielles avant qu'elles ne se produisent. Pour obtenir des résultats optimaux, il est important de disposer d'une grande quantité de données de fonctionnement de qualité et de disposer des compétences nécessaires pour entraîner et utiliser des modèles de machine learning, mettre en place une infrastructure adéquate pour collecter et traiter

## **Importance de la maintenance prédictive dans le data management**

La maintenance prédictive est devenue particulièrement importante dans les processus de data management et de data Analytics, car elle peut aider à maximiser la disponibilité et la fiabilité des systèmes de stockage et de traitement de données. En utilisant des techniques de prévision et de diagnostic avancées, il est possible de détecter les problèmes potentiels avant qu'ils ne surviennent et de planifier les interventions de maintenance de manière plus efficace. Cela peut non seulement réduire les temps d'arrêt et les coûts de maintenance, mais également améliorer la qualité des données collectées et analysées.

De nos jours, les entreprises dépendent de plus en plus de leurs systèmes de stockage et de traitement de données pour collecter, analyser et utiliser les données de leurs activités. Ces données sont souvent critiques pour la prise de décision stratégique et pour l'optimisation des processus d'affaires. Si les systèmes de stockage et de traitement de données tombent en panne ou sont inefficaces, cela peut entraîner des retards coûteux, des erreurs de décision et une perte de confiance de la part des clients et des partenaires d'affaires.

En utilisant la maintenance prédictive, il est possible de minimiser les risques de défaillance et de garantir la disponibilité et la fiabilité des systèmes de stockage et de traitement de données. Cela peut être accompli en utilisant des capteurs de données et des systèmes de surveillance en temps réel pour collecter des données sur l'état des équipements, en utilisant des algorithmes de machine learning pour détecter les problèmes potentiels et en planifiant les interventions de maintenance de manière proactive plutôt que de devoir réagir en cas de panne.

En outre, la maintenance prédictive peut être utilisée pour optimiser l'utilisation des ressources informatiques et améliorer l'efficacité des processus de data management. Par exemple, en utilisant des algorithmes de machine learning pour détecter les problèmes de performance des bases de données, il est possible de planifier des optimisations avant qu'elles ne soient nécessaires. De même, en surveillant en temps réel l'utilisation des ressources informatiques, il est possible de répartir de manière plus efficace la charge de traitement et de réduire les coûts informatiques à long terme.

Enfin, la maintenance prédictive peut être utilisée pour améliorer l'efficacité des processus de data analytics et de business intelligence. En surveillant en temps réel l'état des systèmes de stockage et de traitement de données, il est possible de garantir que les données collectées sont de haute qualité et de les utiliser de manière plus efficace dans les analyses. De cette manière, la maintenance prédictive peut contribuer à améliorer la prise de décision stratégique et à optimiser l'efficacité des processus d'affaires.

## **Méthodes de machine learning utilisées dans la maintenance prédictive**

Il existe plusieurs méthodes de machine learning qui peuvent être utilisées dans la maintenance prédictive pour détecter les problèmes potentiels et planifier les interventions de maintenance de manière plus efficace.

Les algorithmes de classification sont couramment utilisés pour prédire si un équipement présentera ou non un problème dans un avenir proche. Ils fonctionnent en apprenant à partir de données historiques sur les performances des équipements et en identifiant des modèles de comportement anormaux ou des indicateurs de problèmes à venir. Il existe plusieurs types d'algorithmes de classification, tels que les réseaux de neurones, les arbres de décision et les k plus proches voisins. Chacun de ces algorithmes utilise une approche différente pour apprendre à partir de données et prédire les problèmes futurs.   
Les réseaux de neurones sont des algorithmes de classification qui sont inspirés du fonctionnement du cerveau humain. Ils sont constitués de couches de neurones qui sont connectées entre elles et qui travaillent ensemble pour analyser les données et prédire les problèmes futurs. Les réseaux de neurones peuvent être utilisés pour traiter des données structurées et non structurées, et ils sont particulièrement efficaces pour détecter les modèles de comportement complexes dans les données.   
Les arbres de décision sont des algorithmes de classification qui fonctionnent en construisant un arbre de décision à partir de données d'entraînement. Chaque nœud de l'arbre représente une décision à prendre, et les branches de l'arbre représentent les différentes options qui peuvent être prises en fonction de la décision. Les arbres de décision sont particulièrement utiles pour traiter des données structurées et pour prédire les problèmes futurs en prenant en compte plusieurs facteurs différents.   
Les k plus proches voisins sont particulièrement utiles pour traiter des données non structurées et pour prédire les problèmes futurs en prenant en compte plusieurs facteurs différents. Cependant, ils ont tendance à être moins précis que d'autres algorithmes de classification lorsqu'il y a peu de données disponibles ou lorsque les données sont très bruyantes ou incomplètes. En outre, les k plus proches voisins ont besoin de plus de temps de calcul pour prédire les problèmes futurs que d'autres algorithmes de classification, ce qui peut être un problème dans les environnements où la vitesse de traitement est critique.

Les algorithmes de régression sont également utilisés pour prédire les problèmes futurs dans la maintenance prédictive. Ils fonctionnent en apprenant à partir de données historiques sur les performances des équipements et en identifiant des modèles de comportement qui permettent de prédire comment un équipement donné se comportera dans un avenir proche. Une fois qu'un modèle de régression a été entraîné, il peut être utilisé pour prédire les performances futures d'un équipement donné et planifier les interventions de maintenance de manière proactive.

 Nous avons aussi les algorithmes de détection d'anomalies dans la maintenance prédictive. Ils fonctionnent en surveillant en temps réel l'état des équipements et en détectant les comportements anormaux ou les indicateurs de problèmes potentiels. Une fois qu'un problème potentiel a été détecté, il peut être corrigé de manière proactive avant qu'il ne devienne un problème plus grave.    
Il existe plusieurs types d'algorithmes de détection d'anomalies, tels que les algorithmes de seuil, les algorithmes de distribution et les algorithmes de densité de point. Chacun de ces algorithmes utilise une approche différente pour détecter les anomalies dans les données en temps réel.   
Les algorithmes de seuil fonctionnent en définissant un seuil sur les données en temps réel et en détectant les valeurs qui dépassent ce seuil. Si une valeur dépasse le seuil, elle est considérée comme une anomalie et peut être utilisée pour détecter un problème potentiel. Les algorithmes de seuil sont simples à utiliser et rapides à mettre en œuvre, mais ils ont tendance à être moins précis que d'autres algorithmes de détection d'anomalies lorsque les données sont bruyantes ou incomplètes   
Les algorithmes de distribution fonctionnent en apprenant à partir de données historiques sur les performances des équipements et en construisant un modèle de distribution des données. Ensuite, ils peuvent utiliser ce modèle pour détecter les valeurs qui sont anormales par rapport à la distribution attendue. Les algorithmes de distribution sont particulièrement utiles pour détecter les anomalies dans les données non structurées et pour prendre en compte plusieurs facteurs différents. Cependant, ils nécessitent plus de temps de calcul que les algorithmes de seuil et peuvent être difficiles à mettre en œuvre dans des environnements où la vitesse de traitement est critique.   
Les algorithmes de densité de point sont particulièrement utiles pour détecter les anomalies dans les données en temps réel et pour prendre en compte plusieurs facteurs différents. Cependant, ils ont besoin de beaucoup de données historiques pour être efficaces et peuvent être difficiles à mettre en œuvre dans des environnements où les données sont bruyantes ou incomplètes. En outre, ils nécessitent plus de temps de calcul que les algorithmes de seuil et peuvent être difficiles à mettre en œuvre dans des environnements où la vitesse de traitement est critique.

## **Limites actuelles de l'utilisation du machine learning dans la maintenance prédictive**

Bien que l'utilisation du machine learning dans la maintenance prédictive des processus de gestion de données présente de nombreux avantages, il existe également certaines limites actuelles à son utilisation.

Une des limites techniques les plus importantes de l'utilisation du machine learning dans la maintenance prédictive des processus de gestion de données est liée à la complexité et à la variabilité des données. Dans de nombreux cas, les données utilisées pour entraîner les algorithmes de machine learning sont complexes et hétérogènes, avec de nombreuses variables différentes qui peuvent influencer les performances des équipements et des processus de gestion de données. Cette complexité et cette variabilité peuvent rendre difficile l'entraînement et le déploiement efficace des algorithmes de machine learning, car ils peuvent être sensibles aux changements dans les données et aux erreurs de calcul.

De plus, il existe également des limites liées à la qualité et à la quantité des données disponibles. Pour que les algorithmes de machine learning soient efficaces dans la maintenance prédictive des processus de gestion de données, ils ont besoin d'un grand nombre de données de haute qualité sur les performances des équipements et sur les processus de gestion de données. Cependant, dans de nombreux cas, les données sont bruyantes, incomplètes ou peu fiables, ce qui peut compromettre la précision des prévisions.

La scalabilité est un problème important dans l'utilisation du machine learning dans la maintenance prédictive des processus de gestion de données. Les algorithmes de machine learning doivent être capables de traiter de grandes quantités de données et de fonctionner efficacement dans des environnements de gestion de données volumineux et complexes. Cela peut nécessiter l'utilisation de techniques de parallélisation et de distribution de calcul pour optimiser les performances des algorithmes et gérer les exigences de temps de calcul et de puissance de traitement.

Il existe plusieurs techniques couramment utilisées pour améliorer la scalabilité des algorithmes de machine learning dans la maintenance prédictive des processus de gestion de données. L'une de ces techniques est l'utilisation de modèles de machine learning distribués, qui permettent de répartir les calculs sur plusieurs machines ou sur plusieurs nœuds d'un cluster. Cela permet de réduire les temps de calcul et de réaliser des prévisions plus rapidement, ce qui est essentiel dans les processus de gestion de données en temps réel.

Il est également possible d'utiliser des techniques de réduction de dimensionnalité pour améliorer la scalabilité des algorithmes de machine learning dans la maintenance prédictive des processus de gestion de données. La réduction de dimensionnalité consiste à réduire le nombre de variables prises en compte par l'algorithme, ce qui peut permettre de réduire les temps de calcul et d'améliorer les performances de l'algorithme. Cependant, il est important de noter que la réduction de dimensionnalité peut également compromettre la précision des prévisions, et il est donc important de trouver un bon compromis entre précision et scalabilité.

Nous notons qu’il est également possible d'utiliser des techniques de sampling pour améliorer la scalabilité des algorithmes de machine learning dans la maintenance prédictive des processus de gestion de données. Le sampling consiste à échantillonner une partie des données disponibles et à utiliser cet échantillon pour entraîner et tester l'algorithme, plutôt que de traiter l'ensemble des données. Cette technique peut être particulièrement utile lorsque les données sont volumineuses et que les temps de calcul sont longs. Cependant, il est important de veiller à ne pas compromettre la qualité des données en utilisant un échantillon trop petit ou non représentatif.

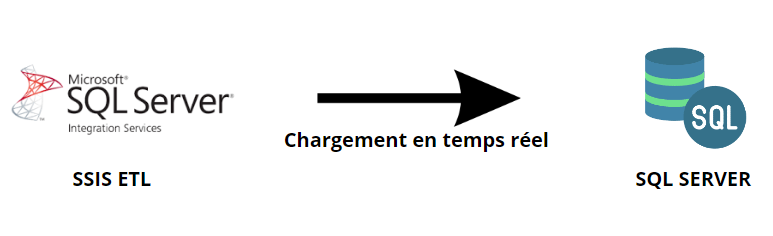
# Méthodologie

## Collecte de la donnée

### Source de données :

Les données utilisées proviennent de SSIS. SQL Server Integration Services (SSIS) est un outil de Microsoft qui permet de créer des packages de données pour l'intégration, la migration et la transformation de données. Il est utilisé pour automatiser les tâches de flux de données complexes, telles que la récupération de données à partir de différentes sources, la transformation de données, le chargement de données dans des bases de données et la synchronisation de données.   
SSIS prend en charge un large éventail de sources de données, y compris les bases de données relationnelles, les fichiers, les services Web, les flux de données en temps réel, les cloud et les sources de données big data. Il prend également en charge un grand nombre de types de transformations de données, comme les calculs, les regroupements, les triages et les recherches de correspondances.

SSIS produit des logs de plusieurs types (Erreur, Alertes, Informations) pour chaque évènement, l’état du système est enregistré dans une base de données relationnelle SQL Server en temps réel.

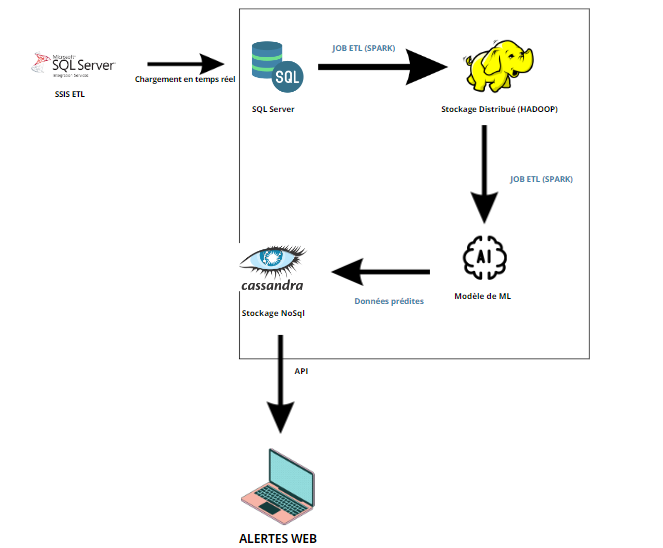


Les logs sont enregistrés depuis septembre 2015 (année d’installation de SSIS dans le département Data de Boursorama). Deux principales tables utilisées sont :  une table d’historique et une table d’enregistrement.

### Architecture de la solution

Une fois les données initiales stockées, dans SQL Server par SSIS il est important de penser à une architecture convenable allant du traitement des données à la visualisation finale. De façon générale, une architecture big Data et machine learning est composée de 6 couches.  

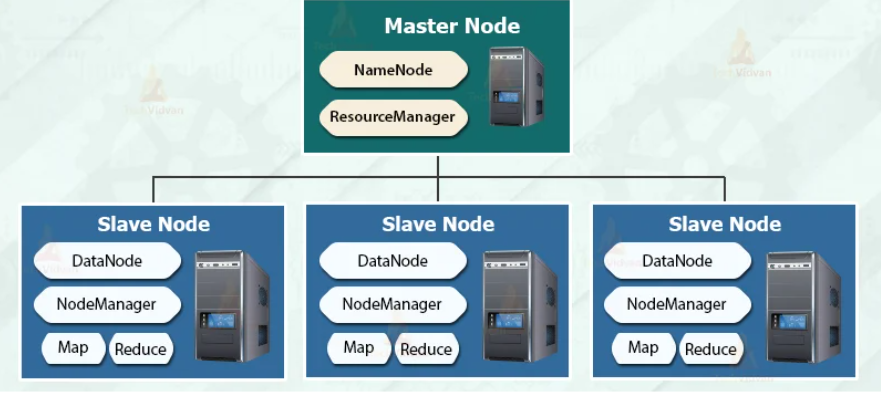
* Couche de stockage : Cette couche est responsable de stocker les données à grande échelle. Les technologies courantes utilisées pour stocker les données incluent HDFS (Hadoop Distributed File System), Apache Cassandra, Apache HBase et Amazon S3.
* Couche de prétraitement : Cette couche est responsable de préparer les données pour l'analyse en utilisant des outils tels que Apache Spark ou Apache Storm. Les tâches courantes incluent la nettoyage des données, l'extraction des caractéristiques et le traitement des données manquantes.
* Couche d'analyse : Cette couche est responsable de l'analyse des données, en utilisant des outils tels que Apache Mahout, Spark MLlib ou TensorFlow. Les tâches courantes incluent l'apprentissage automatique, l'analyse statistique et la modélisation des données.
* Couche de visualisation : Cette couche est responsable de présenter les résultats de l'analyse de manière compréhensible pour les utilisateurs, en utilisant des outils tels que Tableau, QlikView ou Apache Zeppelin. Les utilisateurs peuvent interagir avec les données pour explorer les résultats et les visualiser sous différentes formes.
* Couche d'intégration : Cette couche est responsable de l'intégration de tous les composants de l'architecture, en utilisant des technologies telles que Apache Kafka, Apache Nifi ou RabbitMQ. Cela permet de gérer les flux de données entrant et sortant de l'architecture et de garantir une communication efficace entre les différentes couches.
* Couche de gestion : Cette couche est responsable de gérer les différents composants de l'architecture, en utilisant des outils tels que Apache Ambari, Cloudera Manager ou Databricks . Cela permet de surveiller les performances, de gérer les ressources et de maintenir la disponibilité de l'architecture.



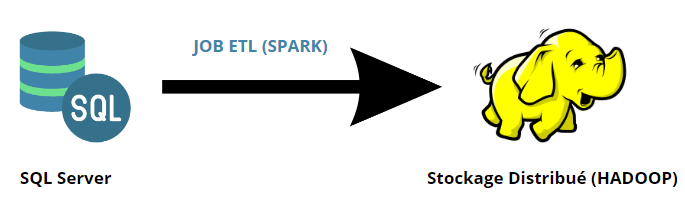
### Extraction de la donnée

Les données stockées dans SQL Server sont utilisées pour des analyses ponctuelles et simples, mais pour des besoins d'analyse et de prévision plus avancés, il est nécessaire de migrer ces données vers une architecture de stockage distribuée plus adaptée.    
   
Cette architecture permet de traiter des volumes de données plus importants, de gérer des débits de données plus élevés et de supporter des demandes de calcul intensif, telles que celles liées aux modèles de prévision de séries temporelles. Les technologies de stockage distribué, comme Apache Hadoop ou Apache Spark, permettent de stocker et de traiter des données sur plusieurs nœuds, augmentant ainsi la capacité de traitement des données et permettant d'analyser les données en temps réel.

Chez Boursorama, on utilise HDFS (Hadoop Distributed File System) qui est un système de fichiers distribué développé dans le cadre du projet Apache Hadoop. Il est conçu pour gérer de grandes quantités de données à travers un réseau de machines, en répartissant les données sur plusieurs nœuds, ce qui permet de traiter et d'analyser les données de manière efficace. HDFS est également conçu pour être tolérant aux pannes, en permettant à des nœuds de tomber en panne sans compromettre la disponibilité des données. En utilisant HDFS, Boursorama peut stocker et traiter des données volumineuses, telles que les données de marché en temps réel, les données de transactions et les données de clients, de manière efficace et fiable.



Une fois que les données ont été collectées et stockées dans SQL Server, elles sont migrées vers HDFS à l'aide d'un Job Spark. Ce processus automatisé est effectué de manière hebdomadaire. L'utilisation de Spark permet de nettoyer les données en éliminant les enregistrements de logs inutiles et en ne conservant que les données pertinentes pour l'analyse. Spark, nous permet de traiter de grandes quantités de données de manière efficace et offre des fonctionnalités avancées pour le nettoyage de données.



### Préparation des données d'entrée du modèle

Les données de logs à l’état brutes sont classées par ligne pour chaque enregistrement, et ceux-ci ont chacun 12 colonnes caractéristiques.    
Les données viennent de 2 tables  principales **AUDIT\_BI..LOG\_LOAD\_DETAIL\_HIST** et **AUDIT\_BI..sysssislog\_hist**

Les données historiques disponible exploitables sont celles du 1er Janvier 2022 au 31 Novemebre 2022



* ID : Identifiant unique de chaque enregistrement dans les logs.
* Event : Catégorie de l'enregistrement pour classer les différentes actions enregistrées.
* Computer : Identifiant du serveur d'où proviennent les logs.
* Operator : Environnement (Prod, Dev, ...) d'où proviennent les logs.
* Source : Action effectuée dans SSIS qui a généré le log.
* Sourceid : Identifiant unique de la source.
* Strarttime : Heure de début de l'enregistrement du log.
* Endtime : Heure de fin de l'enregistrement du log.
* Datacode : Représente la taille de l'enregistrement.
* DataBytes : Nombre de bytes de l'enregistrement.
* Message : Détail du log
* Package : C’est le nom du package

Afin de maximiser l'efficacité de notre analyse, nous avons choisi de ne conserver que les colonnes les plus pertinentes dans nos logs. Les colonnes Operator, DataBytes et Computer, qui présentent des valeurs uniques pour tous les enregistrements, ne seront pas prises en compte dans notre analyse. De plus, les colonnes Starttime et Endtime, qui présentent les mêmes valeurs pour chaque enregistrement, nous n'en conserverons qu'une seule et nous la renommerons en "Date".

Notre objectif étant de prévoir la survenue d'évènements d'erreur en fonction du temps pour chaque package, nous ne conserverons donc que les colonnes suivantes : Package, Starttime(Date) et Event. Pour cela, nous allons créer une colonne "Eventcount" pour notre variable cible, qui prendra la valeur 1 si un évènement est considéré comme une erreur, et 0 sinon. Ensuite, nous regrouperons tous les Eventcount par Date puis supprimerons la colonne "Sourceid" pour obtenir un Dataframe avec seulement 2 colonnes.

### **Modification de l’index**

Définir la date comme l'index d'une série temporelle permet de conserver l'ordre chronologique des données et d'utiliser les fonctionnalités de manipulation de dates de pandas pour effectuer des opérations telles que la resampling, le lissage, l'agrégation, etc.

En définissant la date comme l'index, pandas peut facilement identifier les observations qui appartiennent à des périodes de temps spécifiques, comme les jours, les semaines, les mois ou les années, et les utiliser pour effectuer des opérations de groupe. Il peut également être utilisé pour identifier et supprimer les observations qui ne sont pas valides ou qui sont manquantes.En utilisant la date comme index, il est également plus facile de regrouper les données par des périodes de temps spécifiques pour des analyses telles que les tendances à long terme ou les saisons, il est également plus facile de synchroniser les données avec d'autres séries chronologiques pour des analyses comparatives.   
La fonction pandas qui permet de le faire est **set\_index**

### **Définition de la fréquence**

La fréquence d'une série temporelle en Python définit la période de temps à laquelle les observations de la série sont prises. Cela peut être défini comme hebdomadaire, mensuel, trimestriel, annuel, etc.   
Pour pouvoir analyser les séries chronologiques de manière significative, toutes les périodes de temps doivent être égales et clairement définies. La frequence se doit d’etre constante, ce qui permettra de voir comment les deux caractéristiques sont liées. La fréquence est une mesure du temps et peut aller de quelques millisecondes à plusieurs décennies. Toutefois, celles que nous rencontrons le plus souvent sont quotidiennes, mensuelles, trimestrielles et annuelles.   
La fonction pandas pour définir la frequence est **as\_freq**

### **Filtrage des données**

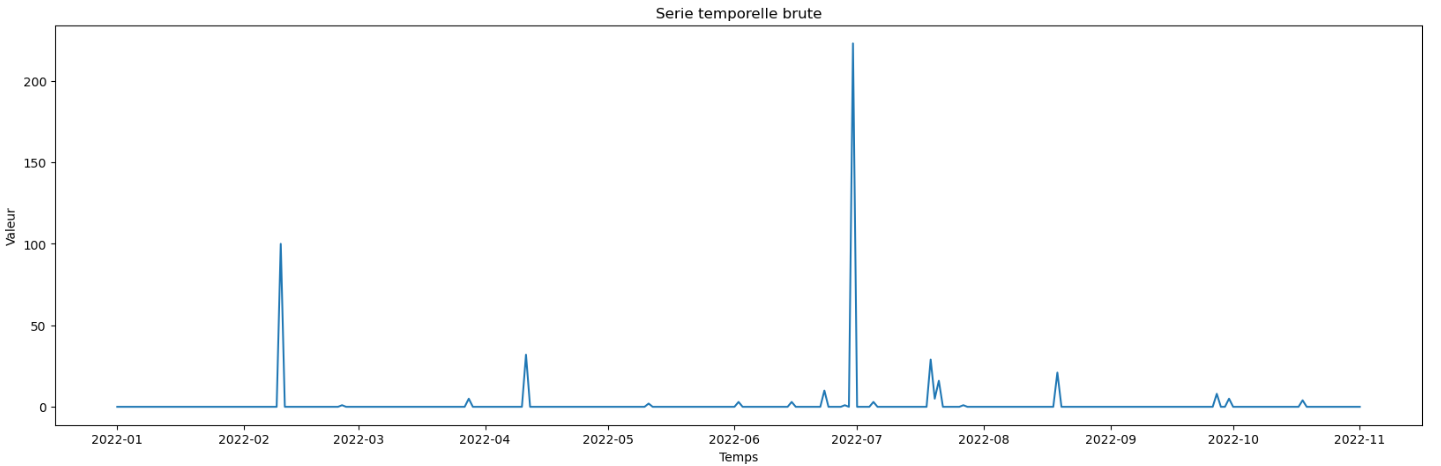
Afin de poursuivre notre analyse et d'obtenir des prévisions précises, nous allons trier les packages en fonction de leur fréquence d'erreurs. Dans notre jeu de données pour l'année 2022, nous avons un total de 256 packages. Certains d'entre eux n'ont jamais connu d'erreurs ou ont rencontré un faible nombre d'erreurs. Pour cette raison, nous allons exclure ces packages et nous concentrer sur ceux qui ont une fréquence d'erreurs plus élevée, en particulier les 10 packages qui ont connu le plus d'erreurs. Cette approche nous permettra de cibler notre analyse sur les packages les plus à risque et de fournir des prévisions plus précises.

D’autre part notre jeu de données est assez large et le traitement de grandes quantités de données peut prendre beaucoup de temps, il est donc pratique de travailler avec un échantillon pour des raisons de performance Enfin, afin de mieux comprendre les tendances, les saisons et les anomalies dans les données.

Pour la suite de l’analyse nous utiliserons le package **DIM\_PERSONNE\_SATTELITE**.

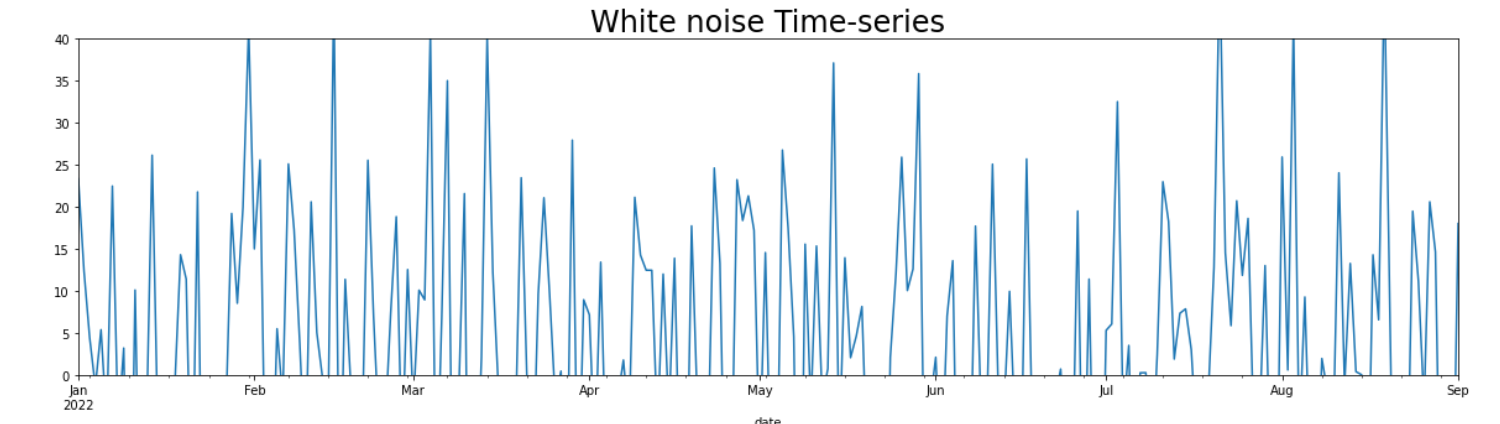
## **Analyse exploratoire des données**

### **Visualisation de notre série et premières interprétations**

A première vue, il est possible de remarquer des pics dans notre série chronologique, ce qui peut nous amener à penser qu'elle n'est pas stationnaire. Il est également possible de constater la présence de plusieurs valeurs nulles, qui pourraient représenter des événements normaux, qui ne lèvent pas d‘alerte. En examinant de plus près les données, il est possible de constater des augmentations importantes d'erreurs pendant les mois de juin à septembre. Il est important de noter que ces observations sont des hypothèses et qu'il sera nécessaire d'analyser les données de manière plus détaillée pour en déduire des conclusions plus précises.

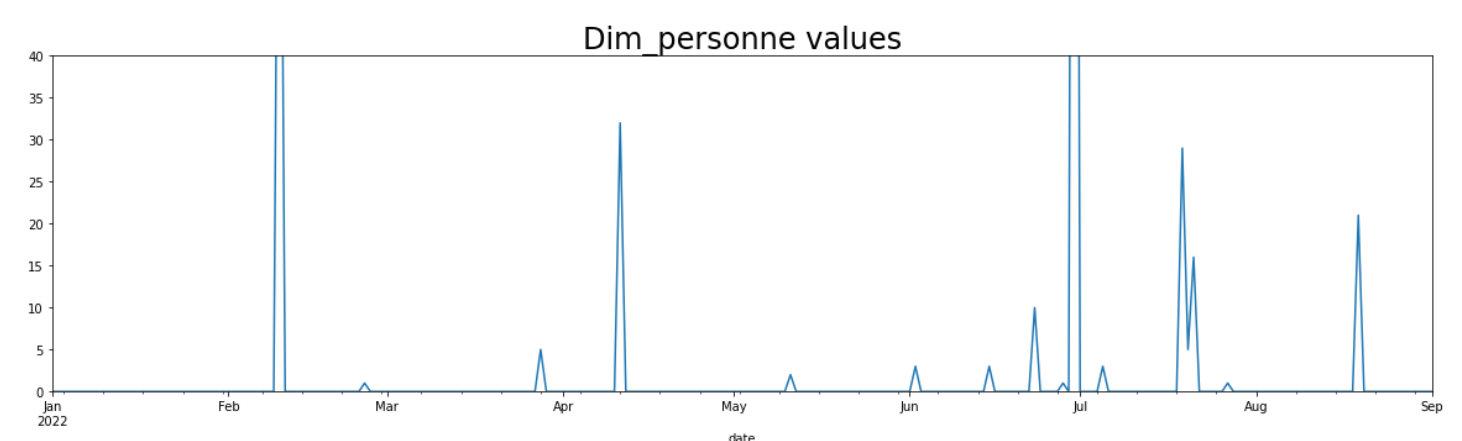
### **White Noise (Bruit Blanc)**

En séries chronologiques, le bruit blanc (white noise) est un processus aléatoire qui suit une distribution normale. Il est caractérisé par une constante moyenne et une constante variance, et les échantillons de bruit blanc sont indépendants les uns des autres. Autrement dit, le bruit blanc est un processus aléatoire qui ne présente pas de corrélation entre les échantillons.     
Si votre série temporelle est un bruit blanc, alors, par définition, elle est aléatoire Et nous ne pouvons pas raisonnablement la modéliser et faire des prédictions.

Pour vérifier, nous allons créer une série aléatoire qui suit une distribution normale.    
Ensuite nous attribuons à notre bruit blanc créé la moyenne et l’écart type de notre série temporelle (DIM\_PERSONNE\_SATELLITE)   
Nous affichons le white noise créé.    


Il est également possible de voir que les données ont des valeurs négatives et positives et une grande dispersion, ce qui est typique d'un White noise .

Affichons notre série sur la même période



Nous remarquons que les sauts entre les périodes sont beaucoup plus faibles pour les valeurs de DIM\_PERSONNE\_SATELLITE puisque les valeurs ne sont pas aléatoires, donc notre série n’est pas un White noise, nous pouvons continuer l’analyse de façon classique.

### **La stationnarité**

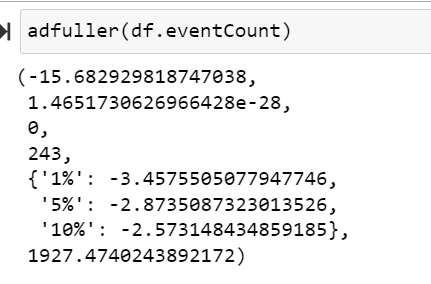
La stationnarité d'une série temporelle signifie que les propriétés statistiques de la série temporelle, telles que la moyenne et la variance, ne changent pas au fil du temps. Cela signifie que les fluctuations dans les données au fil du temps sont dues au bruit aléatoire et non à des variations à long terme.

Une série temporelle est dite stationnaire si elle satisfait les deux conditions suivantes :

* La moyenne de la série temporelle est constante au fil du temps.
* La variance de la série temporelle est constante au fil du temps.

La stationnarité est importante en analyse de séries temporelles car de nombreux modèles et techniques statistiques utilisés en analyse de séries temporelles supposent que les données sont stationnaires. Si une série temporelle n'est pas stationnaire, cela peut affecter la validité des résultats de l'analyse.

Pour déterminer si la serie est stationnaire nous allons utiliser le Le test augmenté de Dickey-Fuller ou test ADF.



Ce test renvoie plusieurs valeurs, notamment :

* **La valeur de test** : -15.682929818747038
* **La p-value :** 1.4651730626966428e-28
* **Le nombre de lags utilisés** : 0
* **Le nombre d'observations utilisées :** 243
* **Les seuils de signification pour différents niveaux de confiance** : {'1%': -3.4575505077947746, '5%': -2.8735087323013526, '10%': -2.573148434859185}
* **Le critère informatif :** 1927.4740243892172

La valeur de test est un indicateur de la force de la non-stationnarité dans la série chronologique. Plus elle est élevée, plus la série est non-stationnaire. La p-value est utilisée pour déterminer la signification statistique de la valeur de test. Elle mesure la probabilité qu'une valeur de test aussi élevée ou plus élevée se produise par hasard si la série chronologique est vraiment stationnaire. Dans ce cas, la p-value est très faible (1.4651730626966428e-28), ce qui signifie qu'il y a une très faible probabilité que la série chronologique soit stationnaire.

Les seuils de signification pour différents niveaux de confiance sont utilisés pour déterminer si la valeur de test est statistiquement significative. Si la valeur de test est plus élevée que le seuil de signification pour le niveau de confiance souhaité, la série chronologique est considérée comme non-stationnaire. Dans ce cas, la valeur de test est bien en dessous des seuils de signification pour tous les niveaux de confiance, ce qui signifie que la série chronologique est considérée comme stationnaire.

Le critère informatif est une mesure de la qualité de la prédiction de la valeur de test pour un modèle spécifique utilisé pour le test ADF. Il n'est pas nécessaire de s'attarder sur cette valeur pour l'interprétation générale des résultats.

En résumé, en se basant sur les résultats obtenus par le test ADF, on peut conclure que la série chronologique est statistiquement stationnaire. Cependant, il est important de noter que cela ne signifie pas que la série chronologique ne présente pas de tendances ni de saisonnalités, il est nécessaire de faire des analyses plus poussées pour en être sûr.

### **Décomposition de la Série temporelle**

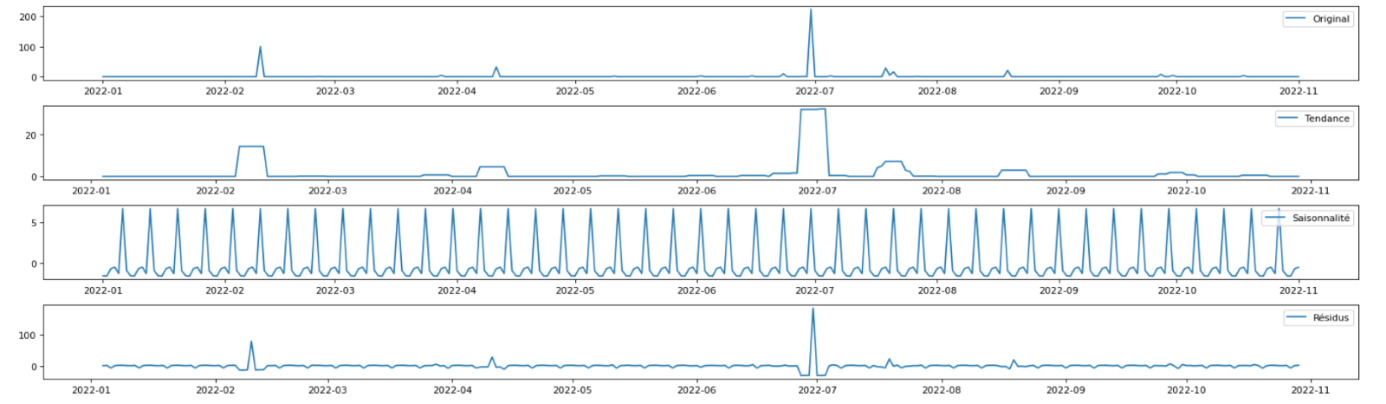
La décomposition de série temporelle permet de décomposer une série temporelle en ses composants tendance, saisonnalité et résidus. Cela peut être utile pour mieux comprendre et analyser les données, ainsi que pour prévoir des valeurs futures.

**La tendance** est la partie de la série temporelle qui représente la direction générale des données au fil du temps, c'est-à-dire une croissance ou une décroissance à long terme.

**La saisonnalité** est la partie de la série temporelle qui représente les variations périodiques des données, telles que les variations mensuelles ou annuelles.

**Les résidus** sont la partie de la série temporelle qui reste une fois que la tendance et la saisonnalité ont été enlevées. Ils peuvent être utilisés pour détecter des anomalies ou des événements imprévus dans les données.

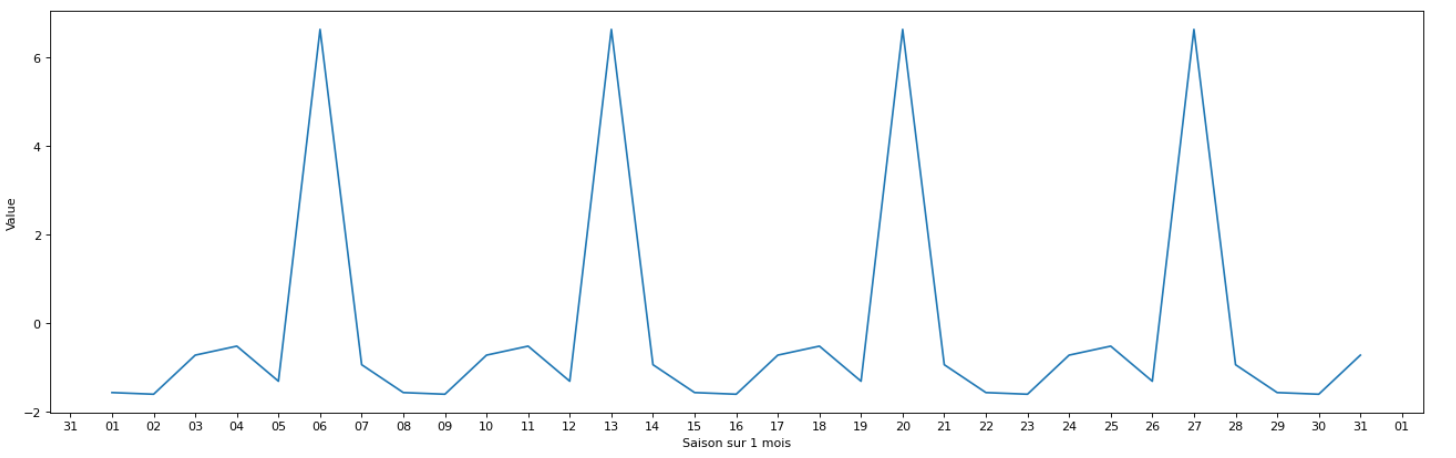
En décomposant une série temporelle, on peut mieux comprendre comment elle est construite et comment ses différents composants interagissent. Cela peut être utile pour identifier des tendances ou des saisons cachées dans les données. Grâce à la décomposition nous comprendrons mieux la serie et pourront choisir la bonne approche de prévision.



### **Saisonnalité**

La saisonnalité est un phénomène fréquent dans les séries chronologiques, où des variations récurrentes se produisent à des intervalles réguliers dans le temps, comme les jours de la semaine, les mois de l'année, etc.    
La saisonnalité peut être causée par des facteurs externes tels que les changements météorologiques, les vacances, les changements dans les politiques économiques, etc. En comprenant la saisonnalité, nous pouvons identifier les causes sous-jacentes et prendre des mesures pour les gérer

L'analyse de la saisonnalité de notre série montre 4 pics d'erreurs mensuellement, cela peut indiquer que la performance du système est affectée par des facteurs saisonniers spécifiques.

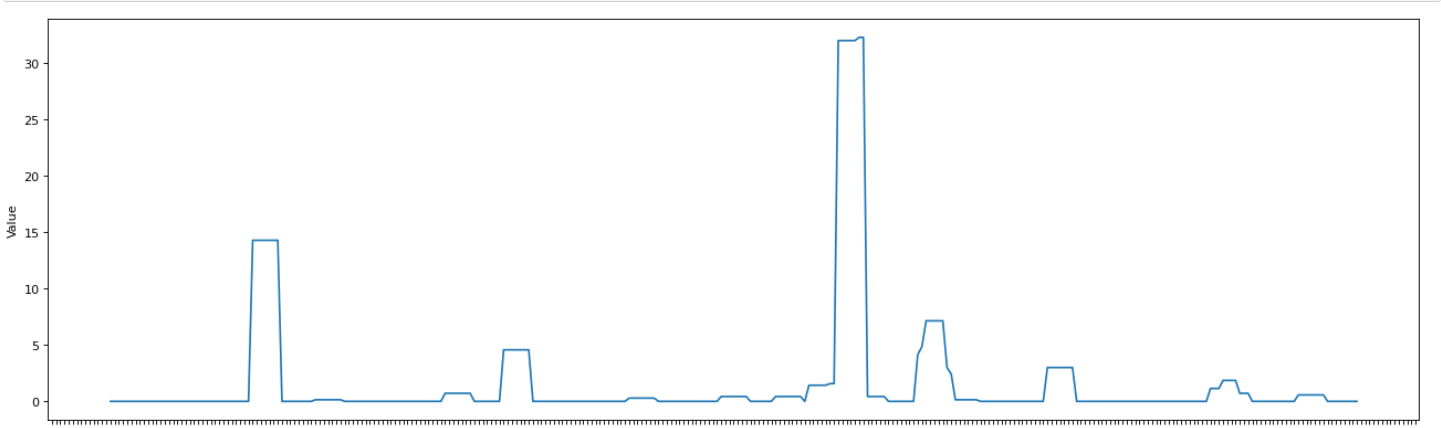


En regardant de plus près un mois, celui de janvier notamment, nous remarquons que ce sont les **6, 13, 20 et 27** du mois qui sont des dates susceptibles d’avoir des pics, donc des erreurs.

### **Tendance**

La tendance d'une série temporelle fait référence à la direction générale dans laquelle les données évoluent au fil du temps. Il peut y avoir différents types de tendances selon le comportement des données :

* Tendance croissante : la tendance est à la hausse, c'est-à-dire que les valeurs augmentent au fil du temps.
* Tendance décroissante : la tendance est à la baisse, c'est-à-dire que les valeurs diminuent au fil du temps.
* Tendance plate : la tendance est stable, c'est-à-dire qu'il n'y a pas de changement significatif des valeurs au fil du temps.
* Tendance non linéaire : la tendance est complexe, c'est-à-dire qu'il y a des changements brusques dans les valeurs au fil du temps



Notre tendance ici est non linéaire, car elle présente des changements brusques au fil du temps. Ce qui nous soulève l’hypothèse qu’une méthode de prévision classique pourrait ne pas être utilisée pour nos données

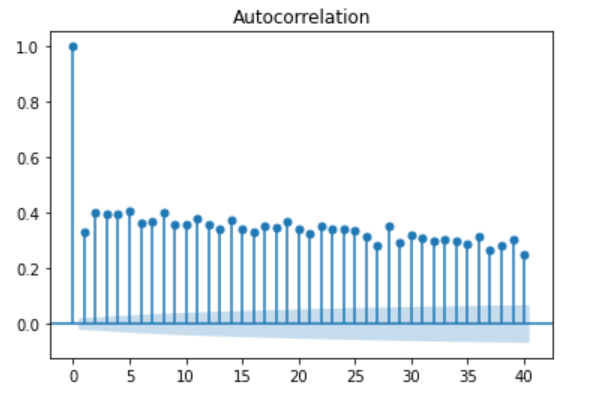
### **L’autocorrélation**

L'autocorrélation est un concept statistique qui mesure la corrélation entre les valeurs d'une série temporelle à des moments différents. Il est utilisé pour déterminer la force de la relation entre une observation à un instant donné et une observation à un instant ultérieur.

C’est dire si à des moments différents sont corrélées entre elles. Car cela permet de déterminer si les données passées ont un impact sur les données futures.

Il existe deux types d'autocorrélation : positive et négative. Une autocorrélation positive signifie que les observations similaires se suivent dans le temps, tandis qu'une autocorrélation négative signifie que les observations opposées se suivent dans le temps.

Traçons maintenant le graphe d’autocorrélation



Ce graphe nous montre que les corrélations sont proches de 0, en dessous même de 0,2.    
De plus en calculant le coefficient d'autocorrélation grâce à la librairie pandas , nous obtenons un score de 0.32    
Cela signifie qu'il y a peu ou pas d'autocorrélation entre les données et l’absence d’autocorrélation est une métrique importante à prendre en compte dans la construction des modèles prédictifs.

### Préparation des données

L'étape de prétraitement des données joue un rôle crucial dans la réussite de tout modèle d'apprentissage automatique. Elle vise à transformer les données brutes en une forme qui soit efficace pour notre modèle.   
Dans le cas de notre problème initial, nous souhaitons anticiper les erreurs dans la chaîne de traitement en déterminant si une erreur se produira ou non. Pour ce faire, nous utiliserons un modèle de classification pour prédire si une donnée sera considérée comme une erreur ou non en se basant sur ses caractéristiques.



* Description des données utilisées
* Description de la méthode de machine Learning utilisée
* Validation de la méthode de machine Learning