PROSIT 1 : ARCHIE

FODIL Nel | MARCELLI Enzo | GOUADFEL Rayan
05 avril 2024

Table des matières

Table des figures

I. Introduction

Le projet entrepris par le Cloud Healthcare Unit (CHU) vise à développer un entrepôt de données pour intégrer et analyser les informations médicales. Notre objectif est de créer un prototype en trois semaines, qui démontrera la capacité de l'entrepôt à répondre aux demandes analytiques des professionnels de santé.

II. Analyse du contexte

Le travail préliminaire du projet a été réalisé par un stagiaire qui a conçu une architecture Big Data et a établi un environnement virtualisé. Avec le projet actuellement en suspens faute de financement supplémentaire, notre équipe a été chargée de le reprendre et de produire un résultat tangible à court terme. Le succès de ce projet dépendra de notre capacité à comprendre et à utiliser les structures et les outils déjà en place, tout en respectant les délais et les réglementations de confidentialité des données.

III. Objectifs

Validation de l'Architecture Big Data :

- Évaluer et approuver le schéma d'architecture Big Data proposé par le stagiaire pour assurer qu'il répond aux besoins de performance, de scalabilité, et de sécurité requise par le CHU.

Modélisation des Données :

- Créer un modèle logique de données qui catégorise et organise les informations en provenance des différentes sources (CSV, MongoDB, PostgreSQL) de façon à faciliter leur traitement et leur analyse.

Préparation des Données pour l'ETL :

- Identifier et préparer les données nécessaires pour le processus ETL afin qu'elles puissent être efficacement extraites, transformées et chargées dans le système HDFS.

Planification de l'Intégration et de l'Implémentation des Données :

- Définir un plan d'action pour l'intégration des données dans HDFS et l'implémentation physique dans l'entrepôt de données.

Établissement de la Stratégie de Visualisation :

- Conception préliminaire des dashboards et rapports sur Power BI en fonction des requêtes et des analyses de données requises pour les utilisateurs finaux du CHU.

Respect de la Confidentialité et de la Sécurité des Données :

- Veiller à ce que toutes les étapes respectent les normes de confidentialité, notamment le RGPD, en traitant les données sensibles des patients avec les précautions nécessaires.

IV. Problématique

Quelle architecture Big Data peut-être mise en place en trois semaines pour analyser les données médicales du CHU tout en garantissant sécurité et conformité ?

V. Plan d'Action

Validation et Architecture

- Examiner le schéma d'architecture Big Data existant pour sa conformité aux exigences techniques et de sécurité.
- Confirmer l'adéquation des composants d'architecture avec les objectifs de traitement et d'analyse des données.

Modélisation et Préparation des Données

- Identifier les sources de données nécessaires et développer un modèle logique pour structurer ces données.
- Élaborer des processus ETL pour l'extraction, la transformation et le chargement des données.

Stockage et Gestion des Données

- Configurer l'environnement HDFS pour le stockage des données selon le modèle établi.
- Programmer et automatiser les jobs d'ETL pour maintenir l'entrepôt de données à jour.

Analyse et Visualisation

- Effectuer des tests pour valider l'intégrité des données et optimiser les performances des requêtes.
- Utiliser Power BI pour créer des visualisations répondant aux besoins d'analyse des utilisateurs finaux.

Sécurité et Conformité

- Intégrer des mesures pour assurer la sécurité des données et la conformité avec les normes, notamment le RGPD.

Révision et Optimisation

- Réviser l'ensemble du processus pour identifier et mettre en œuvre des améliorations potentielles dans l'architecture, le stockage, l'analyse et la visualisation des données.

VI. Définitions

1. Big Data

Le Big Data désigne des volumes considérables de données qui, en raison de leur taille, variété, et vitesse de génération, surpassent les capacités des systèmes traditionnels de gestion de base de données. Les "6V" fournissent un cadre pour comprendre les défis et opportunités associés au Big Data :

- Volume : La quantité de données générées par les interactions numériques, les capteurs, les transactions, etc., est énorme, se mesurant souvent en pétaoctets ou zettaoctets.
- Vitesse : Les données sont produites continuellement, souvent en temps réel ou près du temps réel, nécessitant des technologies capables de traiter rapidement de grands flux d'informations.
- Variété : Les données viennent sous de nombreuses formes - des données structurées comme dans les bases de données traditionnelles à des données non structurées comme le texte, les images, les vidéos, et les données semi-structurées comme les XML ou les JSON.
- Véracité : il se rapporte à la qualité et à l'authenticité des données. Dans un ensemble de données massif, il y a souvent des incohérences, des inexactitudes, ou des informations manquantes qu'il faut gérer.
- Valeur : C'est le potentiel des données à être transformées en informations utiles. Extraire des insights significatifs et actionnables à partir de vastes ensembles de données brutes représente le véritable défi du Big Data.
- Variabilité : Contrairement à la variété, la variabilité fait référence aux changements dans le format des données, dans leur signification (comme les changements de langue dans les données textuelles), ou dans les taux de flux de données, ce qui peut compliquer les processus d'agrégation et d'organisation des données.

Le traitement du Big Data implique des technologies spécialisées comme Hadoop, Spark, et des systèmes de gestion de bases de données NoSQL, qui permettent le stockage, l'analyse, et la gestion efficace de ces grandes quantités de données. Les entreprises et organisations utilisent le Big Data pour améliorer la prise de décision, découvrir de nouvelles opportunités de marché, optimiser les opérations, et personnaliser les expériences des utilisateurs.

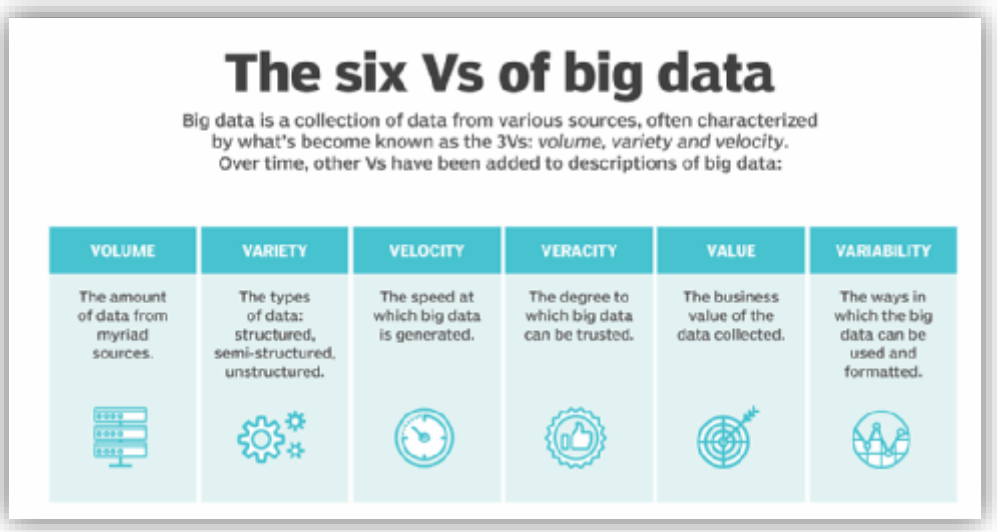

Figure 1 : Big Data
https://www.lemagit.fr/definition/Big-Data

2. Data Warehouse

Un Data Warehouse est une plateforme utilisée pour collecter et analyser des données en provenance de multiples sources hétérogènes. Le Data Warehouse est généralement séparé de la base de données opérationnelle d'une entreprise.

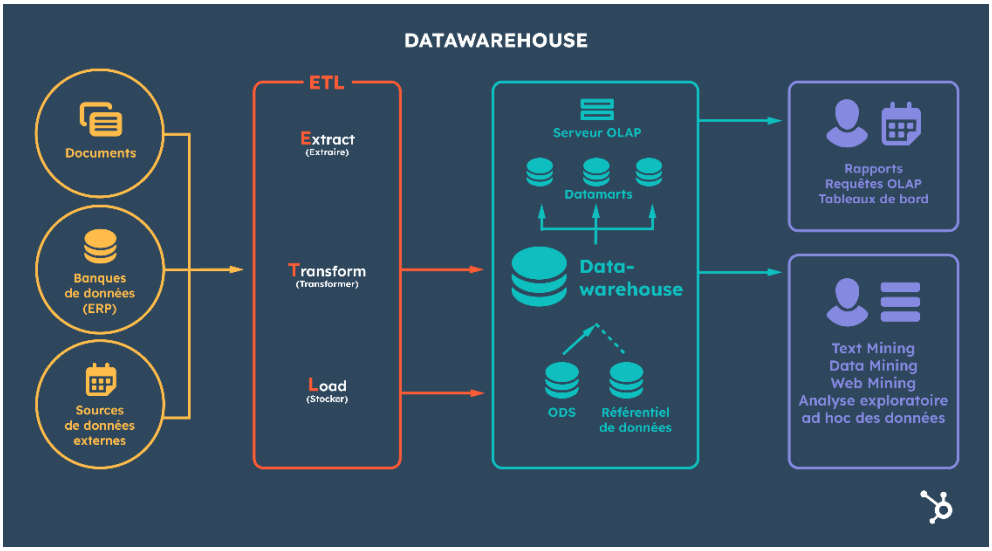

Figure 2 : DataWarehouse
https://blog.hubspot.fr/marketing/datawarehouse

3. Data Lake

Un Data Lake est une plateforme de stockage pour des données brutes en provenance de diverses sources, sans nécessité de les structurer préalablement. Contrairement à un Data Warehouse, il ne contraint pas les données à un format spécifique. Cela permet une flexibilité et une exploration aisée des données pour les analyses.

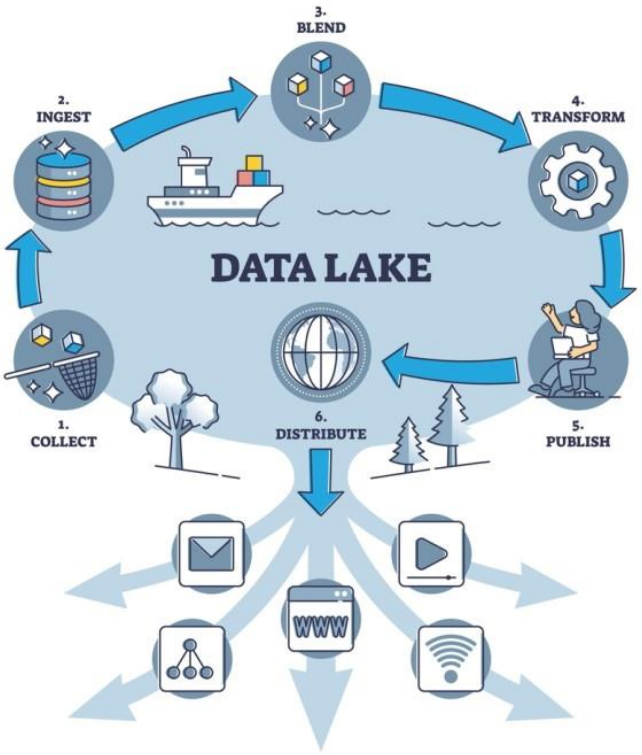

Figure 3 : Data Lake
https://www.corotsystems.com/data-lake/

4. Data Lakehouse

Un Data Lakehouse est une plateforme qui fusionne les fonctionnalités d'un Data Lake et d'un Data Warehouse. Il permet le stockage flexible des données brutes, tout en offrant des capacités de structuration et d'analyse similaires à celles d'un entrepôt de données.

Fonctionnalité	Data Lake	Data Warehouse	Data LakeHouse
Type de données	Brut, structuré, semi-structuré, non structuré	Structuré	Brut, structuré, semi-structuré, non structuré
Structure	Flexible, évolutive	Rigide, définie	Flexible, évolutive avec possibilités de définir des structures
Objectifs	Exploration de données, apprentissage automatique, analyse de l'IoT	Prise de décision opérationnelle, analyse de tendances, reporting	Combine les objectifs du data lake et du data warehouse
Cas d'utilisation	Découvrir de nouvelles informations, prendre des décisions basées sur les données	Répondre à des questions commerciales précises	Combine les cas d'utilisation du data lake et du data warehouse, tout en permettant une exploration des données plus facile sur des jeux de données (dataset) brutes

Figure 4 : Comparatif Data Lakehouse

5. Datamart

Le datamart, traduit par « magasin de données » ou « comptoir de données », est une base de données dont le contenu est en rapport avec une activité de l'entreprise. Il se situe généralement en fin d'une architecture big data, à la suite du Data Warehouse. Ils permettent de définir un accès aux données stockées dans un Data Warehouse. Alors que le Data Warehouse est prévu pour contenir l'intégralité des données d'une entreprise, alors qu'un Data Mart répondra seulement aux besoins d'un département donné ou d'une fonction commerciale spécifique. C'est depuis les Data Mart qu'on met en place des solutions de visualisation de données, tel que ClickView ou Power BI.

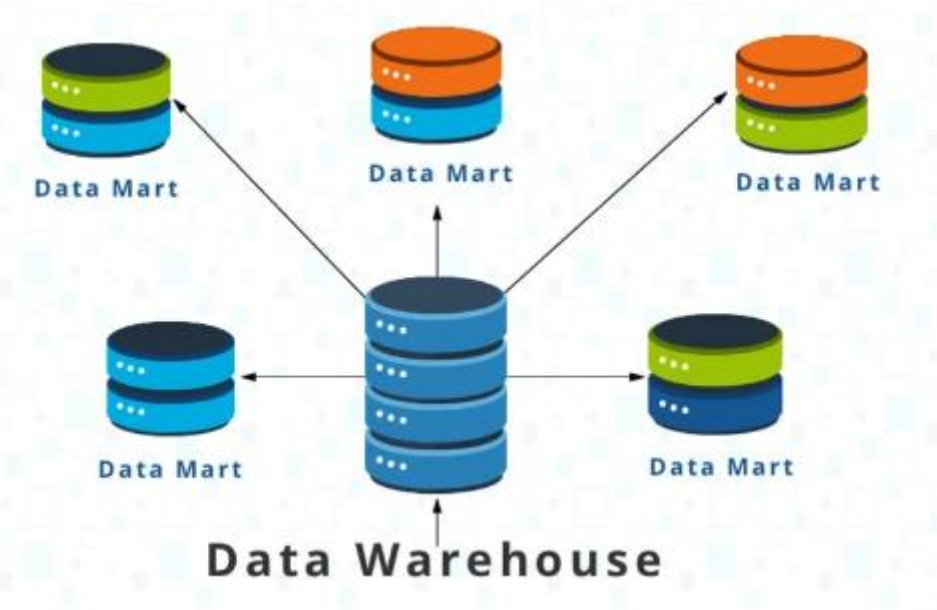

Figure 5 : Data Mart
https://www.talend.com/fr/resources/what-is-data-mart/

6. ETL (Extract, Transform, Load)

Les termes ETL (Extract, Transform, Load) décrivent une série d'actions sur les données : elles sont collectées à partir de diverses sources, puis structurées et enfin centralisées dans un référentiel unique.

Figure 6 : ETL
https://www.talend.com/fr/resources/guide-etl/

7. HDFS (Hadoop Distributed File System)

HDFS (Hadoop Distributed File System) est un système de fichiers distribué conçu pour stocker de très grandes quantités de données sur des clusters de serveurs. Il répartit les données sur plusieurs nœuds pour assurer la redondance et la fiabilité, tout en permettant un accès rapide aux données pour les applications traitant de gros volumes de données, telles que les analyses de données distribuées effectuées avec Hadoop.

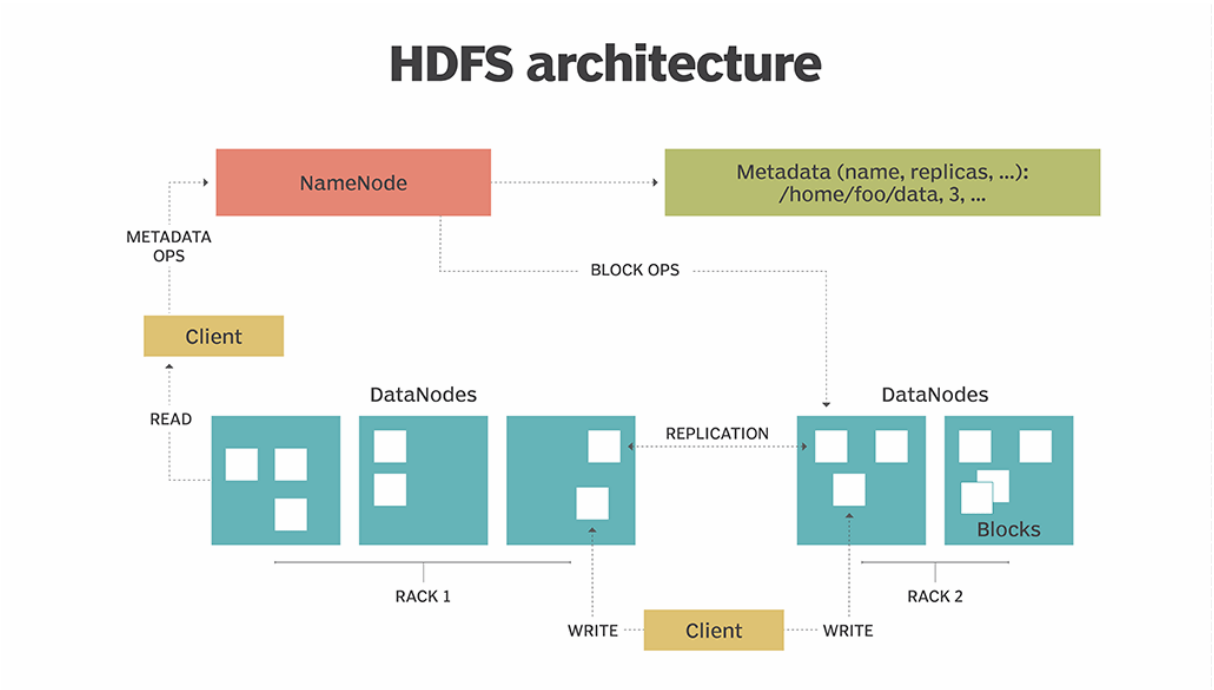

Figure 7 : HDFS architecture
https://www.techtarget.com/searchdatamanagement/definition/Hadoop-Distributed-File-System-HDFS

8. Hadoop

Hadoop est un framework open source pour le stockage et le traitement distribué de gros volumes de données sur des clusters de serveurs. Il comprend des composants comme HDFS pour le stockage et MapReduce pour le traitement parallèle des données.

9. MapReduce

MapReduce est un modèle de programmation pour le traitement parallèle de gros volumes de données sur des clusters distribués. Il divise les tâches en deux phases : "map" pour traiter et filtrer les données, et "reduce" pour agréger les résultats.

10. Hive

Hive est une infrastructure de traitement de données construite au-dessus d'Hadoop qui permet d'écrire des requêtes similaires à SQL pour interroger et analyser les données stockées dans Hadoop HDFS. Il fournit une couche d'abstraction qui permet de manipuler les données sans avoir à écrire des programmes MapReduce complexes. Hive utilise une syntaxe similaire à SQL et traduit les requêtes en tâches MapReduce exécutées sur le cluster Hadoop. Cela rend l'analyse de données plus accessible aux analystes et aux développeurs familiers avec SQL, mais pas nécessairement avec le développement MapReduce.

11. Power BI

Power BI est un outil d'analyse d'affaires de Microsoft qui permet de visualiser des données et de partager des insights à travers une organisation, ou de les intégrer dans une application ou un site web. Il transforme des données brutes en tableaux de bord et rapports interactifs, offrant une compréhension approfondie et une prise de décision basée sur les données. Power BI se connecte à une large gamme de sources de données, facilite la modélisation de données, et permet la création de visualisations personnalisées. Son service cloud, Power BI Service, et ses applications de bureau et mobiles soutiennent une collaboration et une analyse de données efficaces, quel que soit le lieu.

12. MongoDB

MongoDB est un système de gestion de base de données NoSQL orienté documents, conçu pour stocker de grandes quantités de données sous forme de documents JSON-like. Il se distingue par sa flexibilité structurelle, permettant aux développeurs d'ajuster les schémas de données sans downtime. MongoDB est idéal pour les applications nécessitant un stockage de données non relationnelles, comme le big data ou les applications mobiles et web, grâce à sa scalabilité, sa haute performance, et sa facilité d'utilisation. Il supporte des fonctionnalités telles que l'indexation, les agrégations complexes, et la réplication pour la haute disponibilité.

13. PostgreSQL

PostgreSQL est un système de gestion de base de données relationnelle et objet (SGBDRO) open-source. Reconnu pour sa robustesse, sa flexibilité, et son respect des standards SQL, il permet de stocker et de manipuler des données structurées. PostgreSQL supporte des fonctionnalités avancées telles que les transactions ACID (Atomicité, Cohérence, Isolation, Durabilité), les clés étrangères, les vues, les index, et le stockage de données de types variés comme le JSON et XML. Il est extensible par l'ajout de types de données, fonctions, et opérateurs personnalisés, rendant PostgreSQL adapté pour une large gamme d'applications, de l'analytique complexe aux systèmes web interactifs.

14. RGPD (Règlement Général sur la Protection des Données)

Le RGPD, ou Règlement Général sur la Protection des Données, est une réglementation adoptée par l'Union européenne qui s'applique à toutes les entreprises et organisations opérant dans l'UE, ainsi qu'aux entités hors de l'UE qui traitent des données de résidents de l'UE. Mis en vigueur le 25 mai 2018, il vise à renforcer la protection et la confidentialité des données personnelles. Les principes clés incluent le consentement explicite pour le traitement des données, le droit à l'oubli, la portabilité des données, et l'obligation de notifier les violations de données dans un délai strict. Les entreprises doivent également intégrer la protection des données dès la conception de leurs services et peuvent être tenues de nommer un Délégué à la Protection des Données (DPO). Les infractions au RGPD peuvent entraîner des amendes allant jusqu'à 20 millions d'euros ou 4% du chiffre d'affaires mondial annuel de l'entreprise, selon le montant le plus élevé.

15. Cloudera

Cloudera est une plateforme de gestion de données et d'analyse basée sur le cloud, qui permet aux entreprises d'exploiter le potentiel du Big Data. Elle est construite autour de l'écosystème Apache Hadoop, une framework open-source conçue pour le stockage et le traitement de grandes quantités de données de manière distribuée sur des clusters de serveurs. Cloudera étend les capacités de Hadoop en offrant des outils supplémentaires et des services de gestion pour faciliter l'implémentation, la gestion, et l'optimisation des applications Big Data.

Les fonctionnalités principales de Cloudera incluent :

- Gestion de données : Cloudera offre un stockage sécurisé et efficace pour de vastes quantités de données, soutenu par HDFS (Hadoop Distributed File System) et diverses options de bases de données comme Apache HBase.
- Traitement et Analyse : La plateforme supporte une large gamme de frameworks de traitement de données, y compris Apache Spark pour le traitement rapide en mémoire, Apache Hive pour des requêtes SQL sur des données Big Data, et Apache Impala pour l'analyse en temps réel.
- Sécurité et Gouvernance : Cloudera fournit des outils robustes pour la gestion de la sécurité, y compris l'authentification, l'autorisation, le chiffrement des données en repos et en mouvement, ainsi que la gestion des politiques de données pour assurer la conformité réglementaire.
- Machine Learning et Analytique Avancée : La plateforme permet aux utilisateurs de construire et de déployer des modèles de machine learning à grande échelle, utilisant Cloudera Data Science Workbench pour une collaboration et une expérimentation facilitée.
- Gestion et Monitoring : Cloudera Manager offre une interface utilisateur intuitive pour la configuration, la gestion, et le monitoring des clusters Hadoop, simplifiant la tâche complexe de gérer un environnement Big Data.

Note WORKSHOP 1 :

En règle générale, les tables de dimension contiennent un nombre relativement petit de lignes. En revanche, les tables de faits peuvent contenir un très grand nombre de lignes et croître au fil du temps.

Conclusion

En concluant, le projet CHU ambitionne d'exploiter le Big Data pour améliorer la prise de décision dans le domaine de la santé. À travers l'élaboration d'un entrepôt de données et l'intégration de diverses sources d'informations, notre objectif est de faciliter l'accès à des données exploitables. En respectant les contraintes de temps et en veillant à la sécurité des données, nous nous dirigeons vers la création d'un système qui répondra aux besoins analytiques du secteur de la santé, marquant un pas significatif vers l'optimisation des soins et des opérations au sein du CHU.

Webographie

https://www.talend.com/fr/resources/architecture-big-data/

https://datascientest.com/les-metiers-de-la-data

https://www.custup.com/big-data-introduction/

https://blog.octo.com/levolution-des-architectures-decisionnelles-avec-big-data/

https://www.cetic.be/Comment-deployer-avec-succes-un-projet-Big-Data

https://www.solution-bi.com/fr/blog/la-modernisation-du-data-warehouse-a-lere-du-big-data

https://docs.microsoft.com/fr-fr/azure/architecture/guide/architecture-styles/big-data

https://www.lebigdata.fr/data-lake-definition

https://fr.myservername.com/top-15-big-data-tools-2021

https://www.cyres.fr/blog/plateforme-cloudera-a-qui-sert-elle/

https://docs.microsoft.com/fr-fr/power-bi/fundamentals/power-bi-overview

https://chrtophe.developpez.com/tutoriels/filesystems-distribues/

https://waytolearnx.com/2018/07/difference-entre-san-et-nas.html