ANALYSE ET VISUALISATION DES DONNÉES

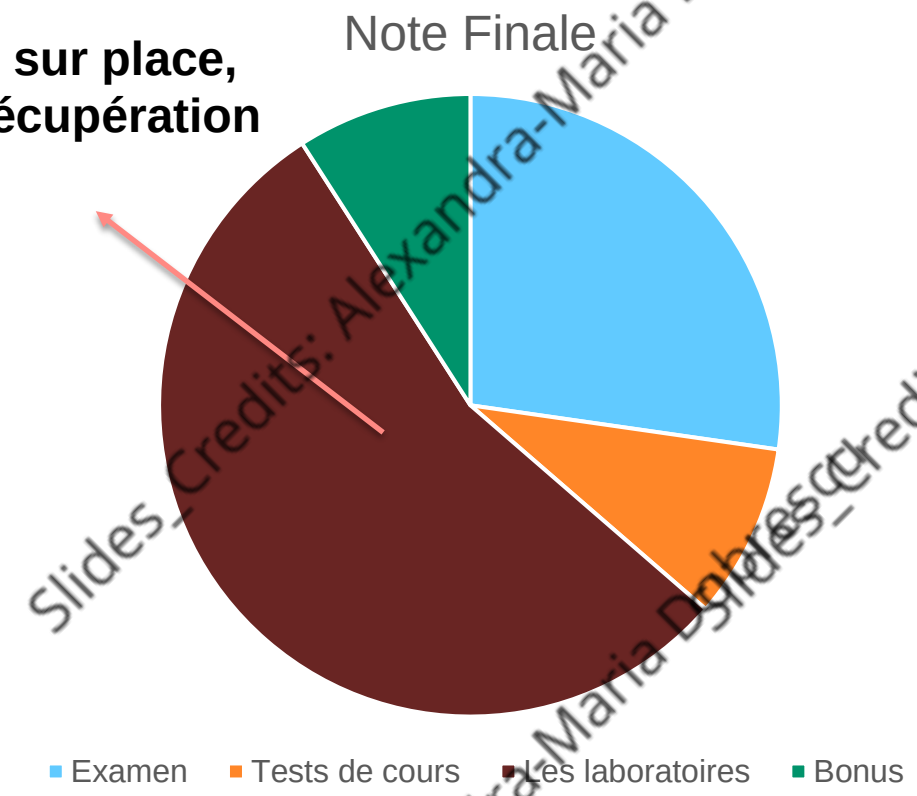
Alexandra-Maria DOBRESCU

| Cours 2 |

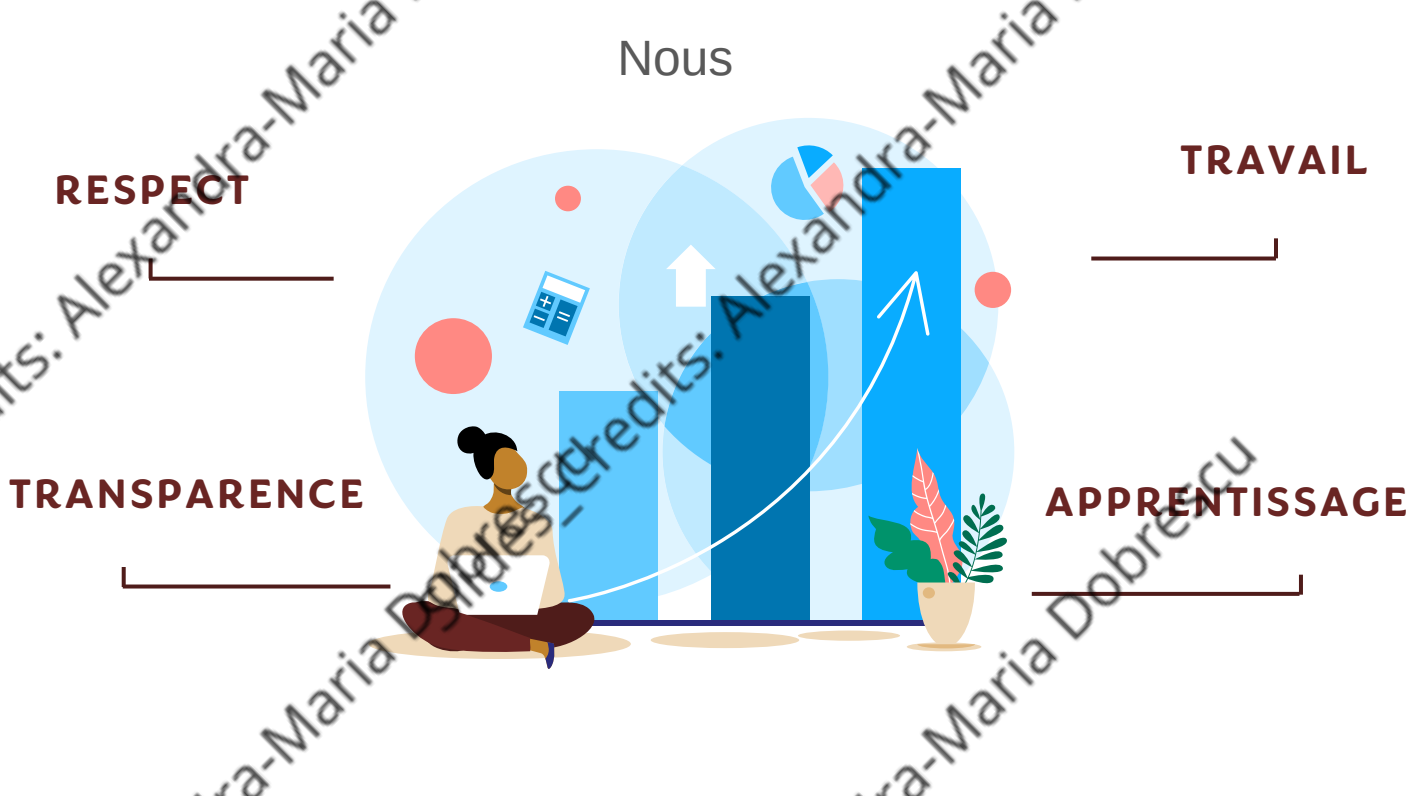
Dis-m'en plus sur les notes

QUANTITATIVE

Travail sur place, sans récupération



QUALITATIVE



Introduction

AVD | Cours 2 | AMD

- DÉFINIR L'ANALYSE DES DONNÉES
- DÉFINIR LA VISUALISATION DES DONNÉES
- APROFONDIR
- COMPARAISON
- POURQUOI L'ANALYSE DES DONNÉES EST IMPORTANTE DANS DIFFÉRENTS DOMAINES
- TYPES DE DONNÉES (STRUCTURÉES, NON-STRUCTURÉES, SEMI-STRUCTURÉES)

Comparaison

OBJECTIF

PROCESSUS

RÉSULTATS

OUTILS

RÔLE

INTERDÉPENDANCE

OBJECTIF

Analyse des données

Implique l'examen, le nettoyage, la transformation et l'interprétation des données afin d'en extraire des informations significatives, d'identifier des modèles, de faire des prédictions ou de répondre à des questions de recherche spécifiques.

Vise à extraire des connaissances et à tirer des conclusions à partir des données.

Visualisation des données

Consiste principalement à représenter visuellement les données au moyen de tableaux, de graphiques, de cartes et d'autres éléments graphiques.

Son objectif principal est d'aider les utilisateurs à comprendre et à interpréter les données rapidement et efficacement.

PROCESSUS

Analyse des données

Comprend des étapes telles que le nettoyage des données, l'analyse exploratoire des données (AED), la modélisation statistique, la vérification des hypothèses et la formulation de conclusions.

Il fait souvent appel à des techniques statistiques et informatiques.

Visualisation des données

Implique la sélection de techniques de visualisation appropriées, la conception de visuels et la création de tableaux ou de graphiques pour représenter visuellement les données.

On met l'accent sur la communication des informations.

Analyse des données

Le résultat de l'analyse des données se présente généralement sous la forme de résumés statistiques, de rapports, de modèles ou de résultats numériques.

Il fournit des informations et des réponses à des questions de recherche spécifiques.

Visualisation des données

Le résultat de la visualisation des données consiste en des diagrammes, des graphiques, des tableaux de bord et des représentations visuelles interactives qui permettent aux individus de saisir plus facilement les tendances, les modèles et les relations au sein des données.

OUTILS

Analyse des données

L'analyse des données utilise souvent des outils logiciels et des langages de programmation tels que R, Python, SQL et des progiciels statistiques pour effectuer des analyses et des calculs.

Visualisation des données

Les outils et bibliothèques de visualisation des données, tels que Tableau, D3.js, Matplotlib et ggplot2, sont utilisés pour créer des représentations visuelles des données.

RÔLE

Analyse des données

Les analystes et les scientifiques des données sont principalement responsables de l'analyse des données. Ils se concentrent sur l'exploration des données, la modélisation statistique et l'obtention d'informations.

Visualisation des données

Les spécialistes de la visualisation des données ou les analystes de données se spécialisent souvent dans la création de visualisations efficaces pour transmettre des informations sur les données à un public plus large.

Analyse des données

L'analyse des données peut générer des idées qui éclairent la conception des visualisations.

Visualisation des données

Les visualisations peuvent aider à présenter les résultats de l'analyse des données d'une manière claire et compréhensible.

GÉNÉRATION DE DONNÉES PARTOUT

L'INTERNET DES OBJETS (IOT)

L'ÈRE DU BIG DATA

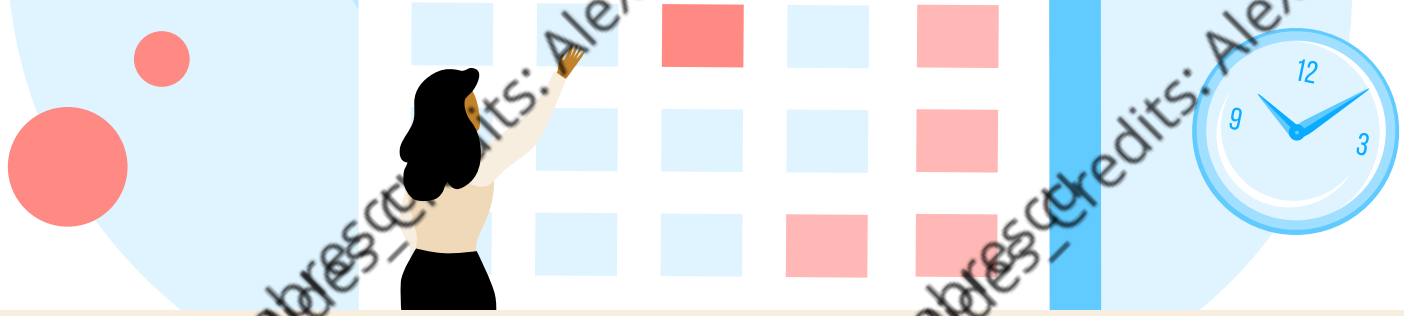
LES DONNÉES DANS LA RECHERCHE ET L'UNIVERSITÉ

PRISE DE DÉCISION FONDÉE SUR DES DONNÉES

CONFIDENTIALITÉ ET SÉCURITÉ DES DONNÉES

VISUALISER LA SURCHARGE DE DONNÉES

SENSIBILISATION AUX DONNÉES PERSONNELLES



Omniprésence des données

Génération de données partout

LES DONNÉES SONT GÉNÉRÉES EN PERMANENCE, TOUT AUTOUR DE NOUS, SOUS DIVERSES FORMES ET À PARTIR DE DIVERSES SOURCES:

Les plateformes de médias sociaux collectent des données sur les interactions des utilisateurs et leurs préférences en matière de contenu [1].

Les sites de commerce électronique enregistrent le comportement des clients, leurs achats et leur historique de navigation.

Les senseurs des appareils intelligents recueillent des données sur la température, l'emplacement et les habitudes d'utilisation [2].

L'internet des objets (IoT)

LES DISPOSITIFS DE L'IOT PRODUISENT DE GRANDES QUANTITÉS DE DONNÉES.

Flux de données continus et en temps réel : Permet de détecter immédiatement les anomalies, les irrégularités ou les événements critiques.

Example 1: Les thermostats intelligents surveillent en permanence la température et l'humidité [3].

Example 2: Les trackers de fitness suivent le rythme cardiaque et les pas.

Permettre d'améliorer considérablement l'efficacité, la réduction des coûts et la prise de décision dans différents domaines.

Étude de cas – L'IdO et la surveillance de l'environnement

EXERCICE : PRÉVISION DE L'INDICE DE QUALITÉ DE L'AIR (IQA)

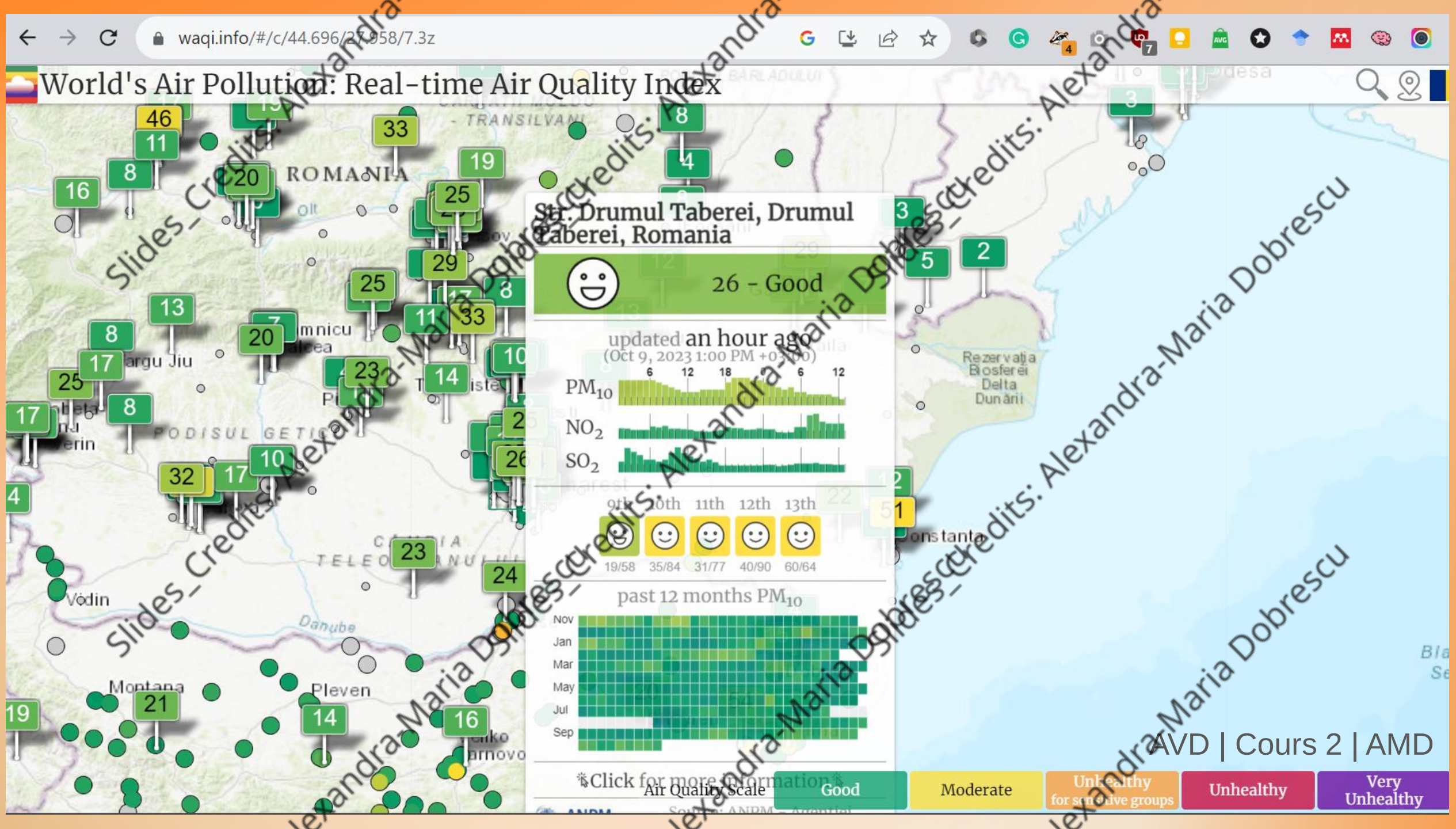
Scénario: Imaginez que vous disposiez d'un réseau de capteurs IoT placés dans une ville pour surveiller la qualité de l'air. Ces capteurs mesurent divers polluants:

- les particules (« Particle Matter » PM2, .5),
- le monoxyde de carbone (CO)
- l'ozone (O3).

Vous souhaitez utiliser les données collectées par ces capteurs pour prédire l'indice de qualité de l'air (IQA) de la ville, qui est une mesure critique de la qualité de l'air.

Description:

1. Vous avez collecté des données au cours de l'année précédente, y compris des mesures quotidiennes de PM2.5, CO, O3, température, humidité et vitesse du vent.
2. Votre tâche consiste à créer un modèle prédictif capable d'estimer l'IQA sur la base de ces variables.


waqi.info/#/c/44.696/27.958/7.3z
World's Air Pollution: Real-time Air Quality Index
ROMANIA
Str. Drumul Taberei, Drumul Taberei, Romania
26 - Good
updated an hour ago
(Oct 9, 2023 1:00 PM +03:00)
PM10
NO2
SO2
9th
10th
11th
12th
13th
19/58
35/84
31/77
40/90
60/64
past 12 months PM10
Nov
Jan
Mar
May
Jul
Sep
Click for more information
Air Quality Scale
Good
Moderate
Unhealthy for sensitive groups
Unhealthy
Very Unhealthy
Rezervatia Biosferei Delta Dunării
Danube
Montana
Pleven
Vidin
AVD | Cours 2 | AMD

Étude de cas – L'IdO et la surveillance de l'environnement

EXERCICE : PRÉVISION DE L'INDICE DE QUALITÉ DE L'AIR (IQA)

Étapes:

1. Divisez votre ensemble de données en un ensemble de formation et un ensemble de test (par exemple, 80 % pour le « training » et 20 % pour le test).
2. Utilisez la régression linéaire ou une autre technique de régression appropriée pour construire un modèle prédictif. Vous pouvez utiliser des outils tels que la bibliothèque scikit-learn de Python [4].
3. Entraînez votre modèle à l'aide de l'ensemble de données d'entraînement. Utilisez des variables telles que les PM2,5, le CO, l'O3, la température, l'humidité et la vitesse du vent comme caractéristiques d'entrée et l'IQA comme variable cible.
4. Évaluez les performances du modèle à l'aide de l'ensemble de données de test. Calculez des mesures telles que l'erreur absolue moyenne (MAE) et le R au carré (R2) pour évaluer la capacité de votre modèle à prédire l'IQA.
5. Interprétez les résultats. Discutez de l'importance de chaque variable d'entrée dans la prédiction de l'IQA. Par exemple, une concentration plus élevée de PM2.5 entraîne-t-elle un IQA plus élevé ?
6. Examinez les possibilités d'optimisation du modèle. Pourriez-vous améliorer sa précision en incluant des variables supplémentaires ou en utilisant une technique de régression différente ?

L'ère du Big Data

LES BIG DATA DÉSIGNENT DES ENSEMBLES DE DONNÉES EXTRÊMEMENT VOLUMINEUX ET COMPLEXES QUI NE PEUVENT ÊTRE GÉRÉS OU ANALYSÉS EFFICACEMENT À L'AIDE D'OUTILS ET DE MÉTHODES DE TRAITEMENT DE DONNÉES TRADITIONNELS.

Dans le traitement traditionnel des données, les systèmes RDBMS (Relational Database Management System) tels qu'Oracle, MySQL ou Microsoft SQL Server ont été les solutions privilégiées pour le stockage et l'analyse **des données structurées**.

Ces systèmes sont excellents pour:

- traiter des données structurées avec des schémas bien définis
- des volumes de données relativement modérés.

Big Data: comprennent généralement de grandes quantités de **données non structurées** ou **semi-structurées,** en plus des données structurées.

L'ère du Big Data

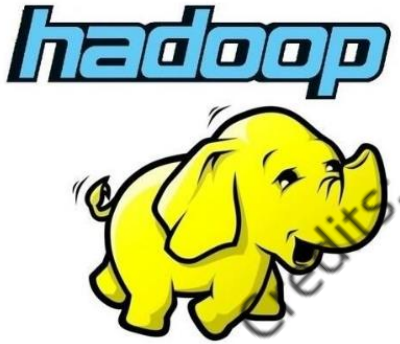
LES LIMITATIONS DU RDBMS

1. **Scalabilité:** Les bases de données traditionnelles peuvent avoir du mal à évoluer horizontalement pour gérer les volumes massifs de données générés dans les scénarios de big data.
2. **Performance:** Au fur et à mesure que les données augmentent, les requêtes sur un RDBMS traditionnel peuvent devenir lentes et inefficaces.
3. **Complexité:** La structuration et la gestion des données dans une base de données relationnelle traditionnelle peuvent devenir complexes, en particulier lorsqu'il s'agit de divers types de données.
4. **Coût**: La mise à l'échelle des bases de données traditionnelles peut être coûteuse, tant en termes de matériel que de licences logicielles.

L'ère du Big Data

POUR GÉRER L'ÉCHELLE ET LA VARIÉTÉ DES DONNÉES VOLUMINEUSES DE MANIÈRE PLUS EFFICACE, ON A:

Des technologies Big Data telles que:



Des bases de données (BD) NoSQL:

Quelle technologie choisir ?

HADOOP

- Traitement de grands ensembles de données dans des environnements où la taille des données dépasse la mémoire disponible.
- Mise en place d'une infrastructure d'analyse de données avec un budget limité.
- Exécution de tâches ne nécessitant pas de résultats immédiats et pour lesquelles le temps n'est pas un facteur limitant.
- Traitement par batch avec des tâches exploitant les opérations de lecture et d'écriture sur disque.
- Analyse de données historiques et d'archives.

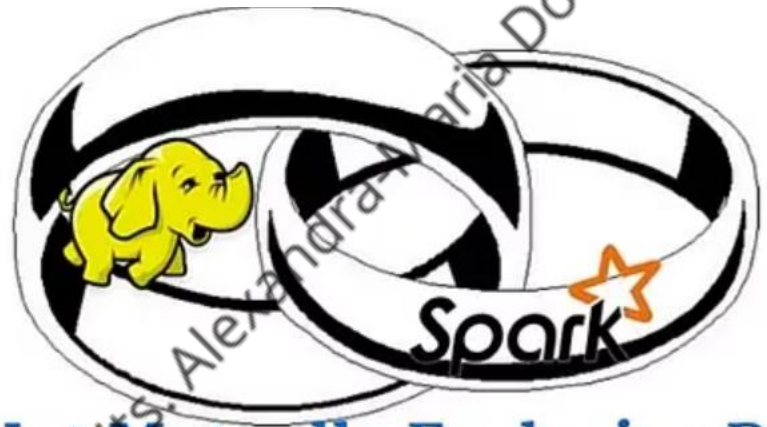
SPARK

- L'analyse des données de flux en temps réel.
- Lorsque le temps est compté, Spark fournit des résultats rapides grâce à des calculs en mémoire.
- Traitement des chaînes d'opérations parallèles à l'aide d'algorithmes itératifs.
- Le traitement parallèle des graphes pour modéliser les données.
- Toutes les applications d'apprentissage automatique.

Quelle technologie choisir ?

Hadoop et Spark fonctionnent mieux ensemble. En tant que successeur, Spark n'est pas là pour remplacer Hadoop, mais pour utiliser ses fonctionnalités afin de créer un nouvel écosystème amélioré.

En combinant les deux, Spark peut tirer parti des fonctionnalités qui lui manquent, telles qu'un système de fichiers.



Not Mutually Exclusive But Better Together

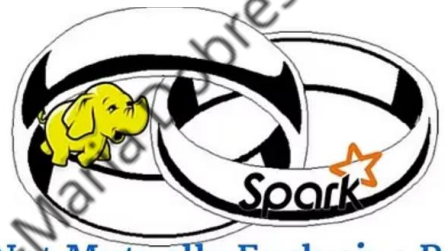
DeZyre

Quelle technologie choisir ?

En combinant les deux, Spark peut tirer parti des fonctionnalités qui lui manquent, telles qu'un système de fichiers.

Hadoop stocke une grande quantité de données à l'aide d'un matériel abordable et effectue ensuite des analyses, tandis que Spark apporte un traitement en temps réel pour traiter les données entrantes.

Sans Hadoop, les applications professionnelles risquent de passer à côté de données historiques cruciales que Spark ne peut pas traiter.



Not Mutually Exclusive But Better Together

Quelle technologie choisir ?

Spark tire également parti des avantages de Hadoop en matière de sécurité et de gestion des ressources.

Avec YARN, la mise en cluster de Spark et la gestion des données sont beaucoup plus faciles.

Vous pouvez exécuter automatiquement les charges de travail Spark en utilisant toutes les ressources disponibles.

Quelle technologie choisir ?

Spark tire également parti des avantages de Hadoop en matière de sécurité et de gestion des ressources.

Vous pouvez exécuter automatiquement les charges de travail Spark en utilisant toutes les ressources disponibles.

Un composant essentiel de l'écosystème Hadoop qui permet une gestion efficace des ressources, une planification des tâches et une gestion du cycle de vie des applications.

Avec ***YARN***, la mise en cluster de Spark et la gestion des données sont beaucoup plus faciles.

Quelle technologie choisir ?

CETTE COLLABORATION PERMET D'OBTENIR LES MEILLEURS RÉSULTATS EN MATIÈRE:

- **D'ANALYSE DE DONNÉES TRANSACTIONNELLES RÉTROACTIVES,**
- **D'ANALYSE AVANCÉE**
- **DE TRAITEMENT DES DONNÉES IOT/IDO.**

TOUS CES CAS D'UTILISATION SONT POSSIBLES DANS UN SEUL ENVIRONNEMENT.

Quelle BD choisir [5] ?

- Système de gestion de base de données **NoSQL** gratuit et open-source, distribué, à large stockage de colonnes.
- Prévu pour traiter de grandes quantités de données sur de nombreux serveurs de base, offrant une haute disponibilité sans point de défaillance unique.

- MongoDB est un programme de base de données non relationnelle (c'est-à-dire **NoSQL**) orienté vers les documents (document-oriented) et multiplateforme (cross-platform).
- Offre une grande vitesse, une grande disponibilité et une grande évolutivité.

Caractéristiques [5]	Cassandra	MongoDB
Modèle de données	**Base de données NoSQL** à grand nombre de colonnes qui utilise un format tabulaire pour le stockage des données. Elle est conçue pour traiter des quantités massives **de données structurées.**	**Base de données NoSQL** basée sur des documents qui stocke les données dans des documents BSON flexibles de type JSON.
Langage de requête (Query Language)	CQL (Cassandra Query Language) pour les requêtes, qui est similaire à SQL mais avec quelques différences dues à sa nature NoSQL.	Langage de requête puissant pour des requêtes flexibles et complexes qui peuvent être exprimées dans une syntaxe de type JSON.

Caractéristiques [5]	Cassandra	MongoDB
Évolutivité (Scalability)	Connu pour son excellente évolutivité horizontale, ce qui le rend adapté aux déploiements distribués à grande échelle.	Prend également en charge l'évolutivité horizontale, mais il se peut qu'il n'évolue pas de manière aussi transparente que Cassandra dans des scénarios de données extrêmement volumineuses et à grande vitesse.
Modèle de Consistance (Consistency Model)	Offre des niveaux de cohérence réglables, ce qui vous permet de choisir entre une consistance forte et une consistance éventuelle en fonction des besoins de votre application.	Offre généralement une cohérence forte au sein d'un document unique, mais permet une consistance éventuelle lorsqu'il est distribué.

Définir de nouveaux mots-clés

CONSISTANCE FORTE

CONSISTANCE ÉVENTUELLE

Définir de nouveaux mots-clés

CONSISTANCE FORTE

Imaginez que vous disposez d'un compte bancaire dont le solde est de 100 dollars. Vous effectuez deux transactions simultanées: l'une pour retirer 30 dollars et l'autre pour vérifier le solde de votre compte. Dans un système à forte cohérence, le système garantit que lorsque vous vérifiez votre solde immédiatement après avoir retiré 30 $, vous verrez le solde mis à jour de 70 $.

En effet, **la cohérence forte garantit que toutes les opérations semblent se produire instantanément**, comme s'il n'existait qu'une seule copie de votre compte.

CONSISTANCE ÉVENTUELLE

Dans un autre scénario, disons que vous avez publié une photo sur une plateforme de médias sociaux. Vous la consultez ensuite immédiatement. Il est possible que vous ne voyiez pas immédiatement votre propre publication lorsque vous la consultez juste après l'avoir publiée. Toutefois, au bout d'un certain temps (quelques secondes ou quelques minutes), toutes les répliques des données du système distribué se synchroniseront et vous finirez par voir votre propre message.

En effet, **la cohérence éventuelle autorise les incohérences temporaires car elle privilégie la disponibilité et la faible latence par rapport à la cohérence immédiate.**

Définir de nouveaux mots-clés

CONSISTANCE FORTE

CONSISTANCE ÉVENTUELLE

Dans les deux exemples, la cohérence forte garantit que vous voyez l'état immédiat et cohérent des données, tandis que la cohérence à terme autorise des incohérences temporaires mais garantit que les données convergeront vers un état cohérent au fil du temps. Le choix entre les deux dépend des exigences spécifiques de l'application et des compromis entre la cohérence, la disponibilité et la latence.

Caractéristiques [5]	Cassandra	MongoDB
Flexibilité du modèle de données (Data Model Flexibility)	Bien adapté aux données de séries temporelles et à l'enregistrement d'événements, où la structure des données reste relativement constante.	La conception sans schéma (schema-less design) est plus flexible et polyvalente, ce qui la rend adaptée à un large éventail d'applications et de données dynamiques.
Cas d'utilisation (Use Cases)	Souvent utilisé dans des scénarios où la haute disponibilité, la tolérance aux pannes et les charges de travail à forte intensité d'écriture sont cruciales, telles que les données chronologiques, les données de capteurs et les analyses en temps réel.	Populaire pour les applications nécessitant de la flexibilité, des requêtes complexes et un développement rapide, telles que les systèmes de gestion de contenu, les catalogues et les profils d'utilisateurs.

Who Uses Cassandra

Who Uses MongoDB

Bibliographie

[1] https://www.freepik.com/premium-vector/round-social-media-icons-network-platforms-logos_12628744.htm
[2] https://www.taggdigital.com/blog/what-are-smartwatch-sensors-and-how-do-they-function
[3] https://www.goodhousekeeping.com/uk/product-reviews/tech/g685723/best-smart-heating-thermostats/#product-e8b0b9d6-3938-4403-8422-24f8da27ff43
[4] Kramer, O., & Kramer, O. (2016). Scikit-learn. Machine learning for evolution strategies, 45-53.
[5] https://www.geeksforgeeks.org/difference-between-cassandra-and-mongodb/
[6] https://www.image-net.org/about.php
[7] G. Miller, R. Beckwith, C. Felbaum, D. Gross, K. Miller, Introduction to WordNet: An On-line Lexical Database (1993)
[8] Knaflic, C. N. (2015). Storytelling with data: A data visualization guide for business professionals. John Wiley & Sons.