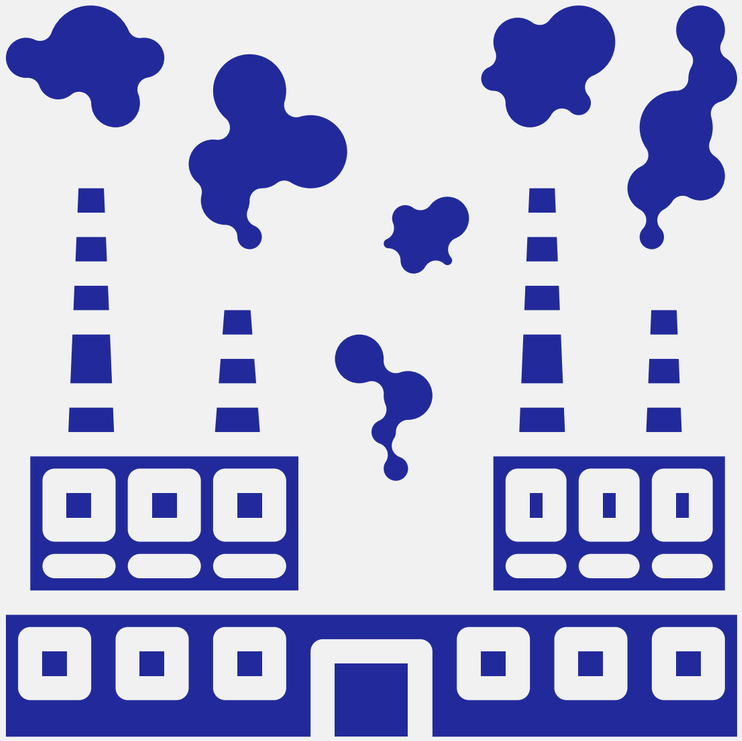
Ecole Polytechnique de Thies



Mise en place d'un ETL: Étude de la qualité de l'air au Sénégal

Cours: Data Engineering

Enseignante: Mme Mously DIAW

Année scolaire:2022/2023

Membres du groupe

Oumar Sahaba NDIAYE

Cheikh Ahmadou Bamba DIOP

Aïssatou BALDE

Plan de la présentation

Étape 1 Contexte: problématique et architecture du projet

Étape 2 Sources et exploration des données

Étape 3 Stockage et déploiement

Étape 4 Visualisation des indicateurs

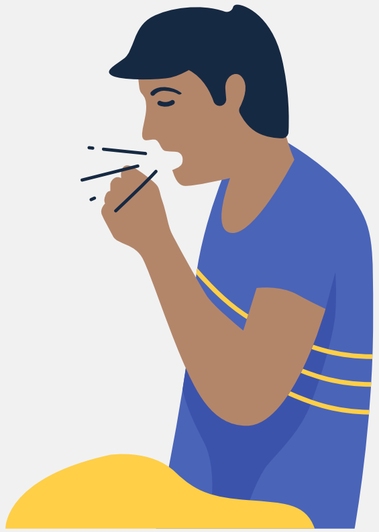
Étape 5 Perspectives

Contexte et architecture

- Problématique
- Objectifs
- Solution et architecture

Contexte

La qualité de l'air



La qualité de l'air fait référence à la mesure de la pureté de l'air que nous respirons. Elle est déterminée par la concentration de différents polluants atmosphériques présents dans l'air. Ces polluants peuvent être d'origine naturelle (comme les émissions volcaniques) ou anthropique (causés par les activités humaines telles que la combustion de combustibles fossiles, les émissions industrielles, le trafic routier, etc.).

L'importance de la qualité de l'air réside dans son impact direct sur la santé humaine. L'un des aspects les plus critiques de la qualité de l'air est son impact sur la santé humaine. L'exposition à des niveaux élevés de polluants atmosphériques, tels que les particules fines (PM2.5), l'ozone, le dioxyde d'azote (NO2) et le dioxyde de soufre (SO2), peut provoquer ou aggraver divers problèmes de santé, notamment les maladies respiratoires, les maladies cardiovasculaires, les allergies, les asthmes, les cancers et même la mortalité prématurée.

Contexte

Problématique

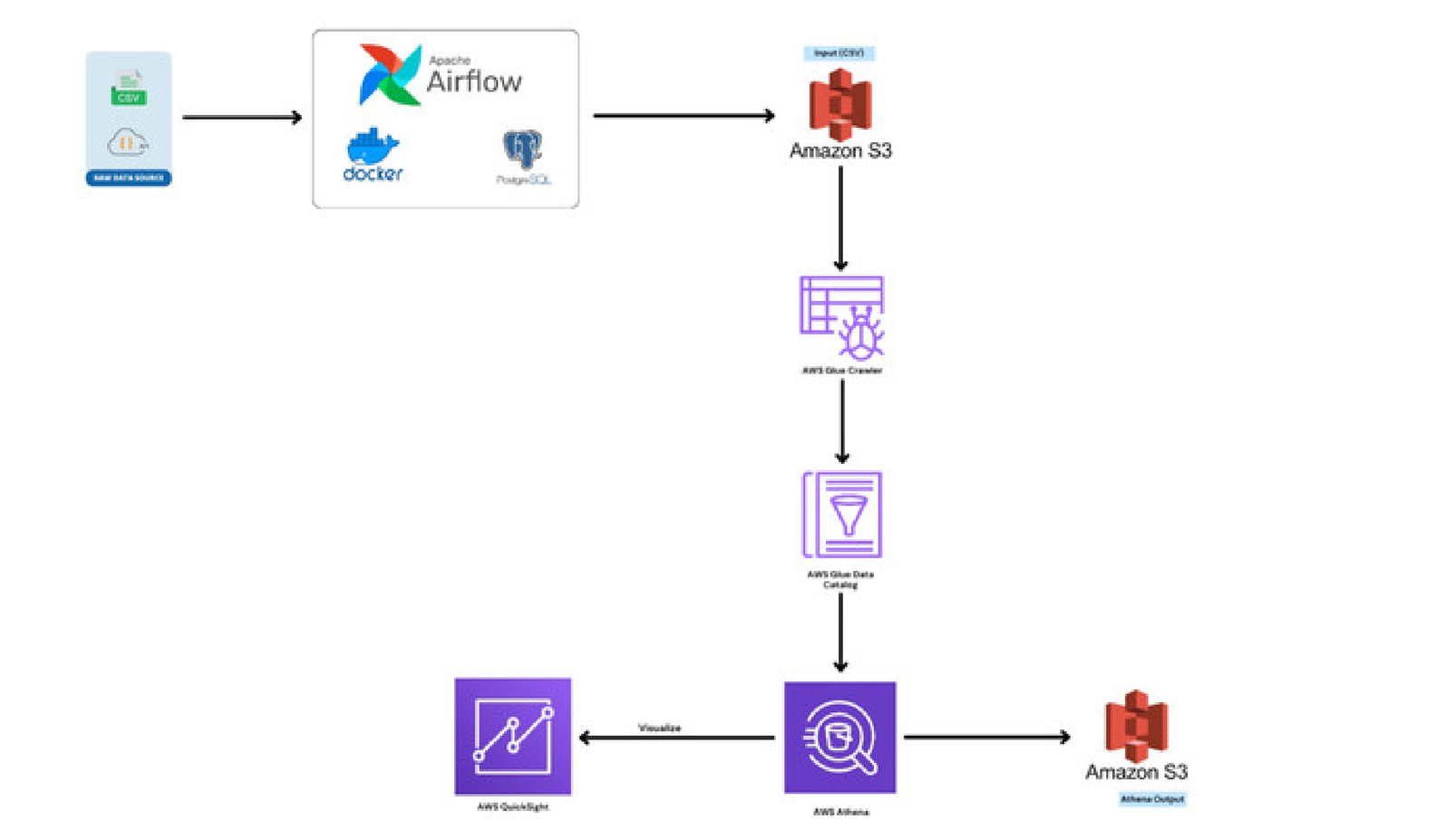
- Au Senegal, on assiste ces dernières années de plus en plus a l'implantation de nouvelles industries très polluantes (port de Sendou) près des populations
- Selon l'OMS, rien que a pollution de l’air à l’intérieur des habitations a été responsable d’environ 3,2 millions de décès par an, en 2020, dont plus de 237 000 décès d’enfants de moins de 5 ans.
- Au Sénégal, il n'existe aucun système d'alerte, ni d'aperçu sur les niveaux journaliers et globaux de la présence des particules dans l'air.

Objectifs

- Permettre au ministère de la Santé d'avoir une vue globale sur la pollution de l'air et de faire des alertes aux populations
- Plus tard avoir une application mobile accessible à tous et qui permet d'avoir également une vue globale des indicateurs de la qualité de l'air

Architecture du project

Pour atteindre les objectifs fixés, nous avons décidé de mettre en place un projet ETL des indicateurs de pollution/qualité de l'air au Sénégal pour le ministère de la santé



ETL pipeline with airflow from ingestion to exploration with Athena and QuickSight

Sources et exploration des données

- Présentation des indicateurs
- Sources de donnees
- Exploration des données

Les indicateurs de la pollution de l'air

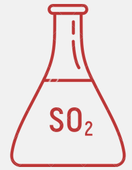
Les indicateurs de la pollution de l'air sont des mesures utilisées pour évaluer la qualité de l'air et les niveaux de pollution atmosphérique dans une région donnée. Ces indicateurs permettent de quantifier et de surveiller la présence de divers polluants atmosphériques, tels que les particules fines, l'ozone, le dioxyde de soufre, le dioxyde d'azote, le monoxyde de carbone et d'autres substances nocives

Particules fines

Les particules fines sont de petites particules solides ou liquides en suspension dans l'air. PM2.5 fait référence aux particules d'un diamètre inférieur à 2,5 micromètres, tandis que PM10 fait référence aux particules d'un diamètre inférieur à 10 micromètres. Ces particules sont produites par des sources telles que la combustion de combustibles fossiles, les émissions industrielles,. Elles peuvent pénétrer profondément dans les voies respiratoires et causer des problèmes respiratoires et cardiovasculaires.

Ozone

L'ozone est un gaz présent naturellement dans l'atmosphère. Cependant, à des niveaux élevés près de la surface de la Terre (l'ozone troposphérique), il peut être nocif pour la santé. L'ozone est formé par des réactions chimiques impliquant des polluants atmosphériques émis par les véhicules, les usines et d'autres sources. L'inhalation d'ozone peut causer des irritations des voies respiratoires, des problèmes respiratoires et aggraver les symptômes des maladies pulmonaires existantes.



Soufre

Le dioxyde de soufre est un gaz produit principalement par la combustion de combustibles fossiles contenant du soufre, tels que le charbon et le pétrole. Il peut provoquer des irritations des voies respiratoires et des problèmes respiratoires chez les personnes sensibles. Le SO2 peut également réagir avec d'autres polluants atmosphériques pour former des particules fines.

Exploration des données de la banque mondiale

- La Banque mondiale (World Bank en anglais) est une institution internationale qui fournit une grande quantité de données économiques, sociales et environnementales provenant du monde entier

- Ces données sont généralement accessibles au public via la plateforme de données de la Banque mondiale appelée World Bank Data (https://data.worldbank.org/).

- **`wbgapi`** est une bibliothèque Python développée par la Banque mondiale (World Bank) pour faciliter l'accès aux données de la Banque mondiale via son API (Interface de Programmation d'Application). Cette bibliothèque permet aux développeurs d'accéder aux ensembles de données de la Banque mondiale et de récupérer des informations économiques, sociales et environnementales à partir de différentes sources.

En exploitant les données de pollution de la banque banque mondiale sur les indicateurs pm2.5 et pm10: on se retrouve avec les indicateurs suivants

- **Mesure annuelle de la pollution de l'air due aux particules fines PM2.5**

Représente la concentration moyenne annuelle de particules fines PM2.5 présentes dans l'air d'une région donnée

- **Proportion de la population d'un pays qui est exposée à des niveaux de pollution atmosphérique dépassant la valeur cible intérimaire 1**

Une de ces valeurs cibles intérimaires spécifiques pour les PM2.5. Elle représente un niveau de concentration des PM2.5 considéré comme acceptable pour la santé publique. Dépasser cette valeur cible peut avoir des effets néfastes sur la santé de la population exposée.

- **Proportion de la population d'un pays qui est exposée à des niveaux de pollution atmosphérique dépassant la valeur cible intérimaire 2**

L'Interim Target-2 est une valeur cible spécifique pour les PM2.5 qui est plus stricte que l'Interim Target-1. Dépasser cette valeur cible peut indiquer une exposition plus élevée aux PM2.5 et peut avoir des conséquences néfastes pour la santé de la population exposée.

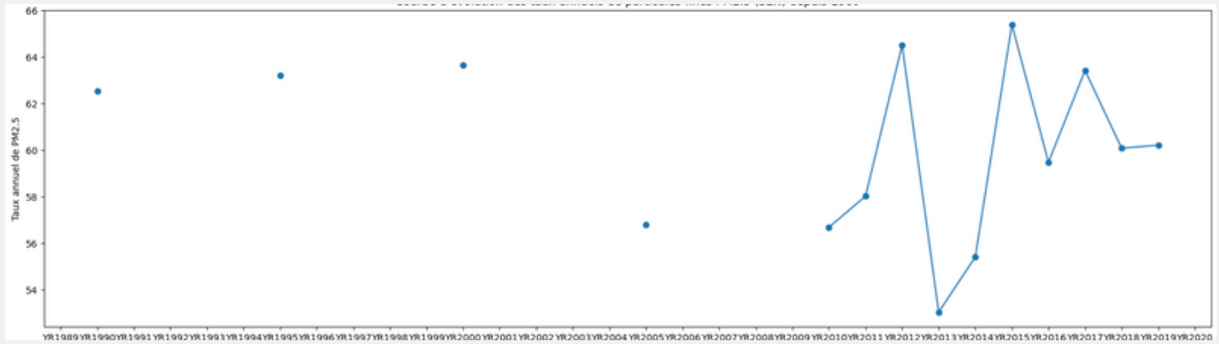
- **Proportion de la population d'un pays qui est exposée à des niveaux de pollution atmosphérique dépassant la valeur cible intérimaire 3**

L'Interim Target-3 est encore plus stricte que l'Interim Target-2

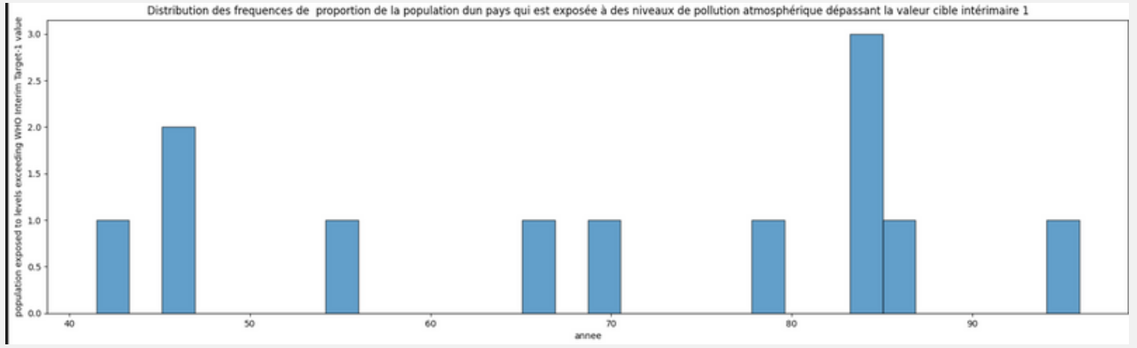
- **Proportion de la population d'un pays qui est exposée à des niveaux de particules fines PM2.5 dépassant la valeur limite (guideline value) établie par l'OMS**

La valeur limite de l'OMS est une valeur cible pour les PM2.5 considérée comme étant sécuritaire pour la santé publique

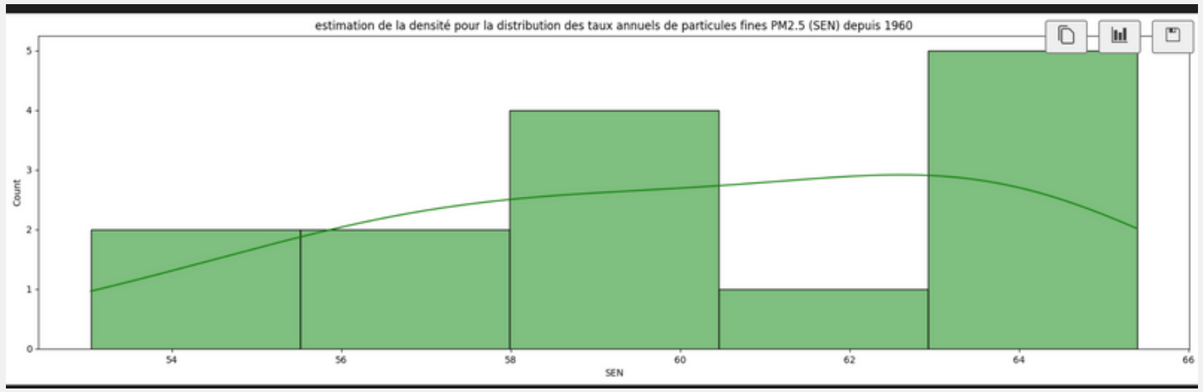
Quelques graphes des données de la banque mondiale



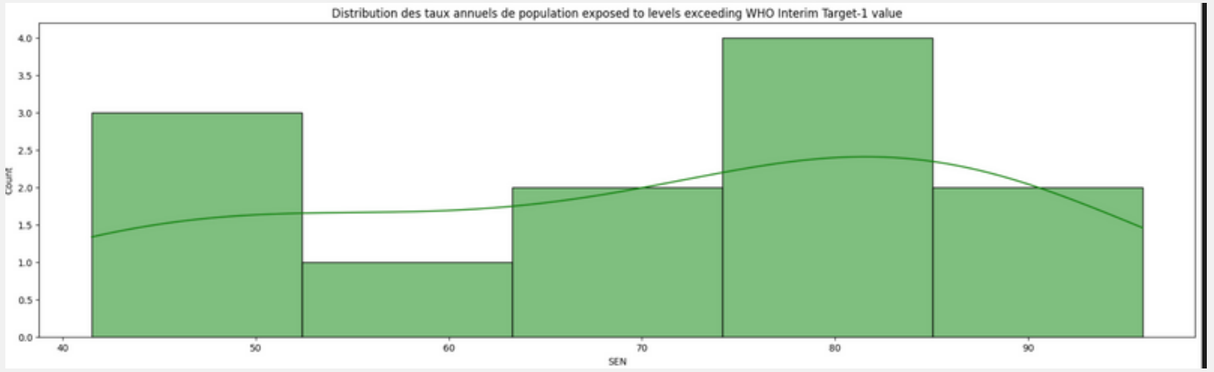
Courbe d'evolution des taux annuels de particules fines PM2.5 (SEN) depuis 1960



Distribution des frequences de proportion de la population du SEN qui est exposée à des niveaux de pollution atmosphérique dépassant la valeur cible intérimaire 1



Estimation de la densité de distribution des taux annuels de population exposed to levels exceeding WHO Interim Target-1 value SEN



estimation de la densité pour la distribution des taux annuels de particules fines PM2.5 (SEN) depuis 1960

Exploration des données du WAQI

- L'API WAQI (World Air Quality Index) est un service en ligne qui fournit des données sur l'indice de la qualité de l'air (AQI) à l'échelle mondiale. L'indice de qualité de l'air (AQI) est un indicateur standardisé utilisé pour communiquer les niveaux de pollution de l'air et les impacts potentiels sur la santé humaine.
- L'API WAQI permet d'accéder à des données en temps réel et historiques sur la qualité de l'air dans différentes villes et régions du monde. Elle fournit des informations sur divers polluants atmosphériques tels que les particules fines (PM2.5 et PM10), l'ozone (O3)

En utilisant l'API WAQI, vous pouvez obtenir des informations telles que :

- L'indice de qualité de l'air global (AQI) pour une ville ou une région spécifique.
- Les concentrations actuelles de différents polluants atmosphériques.
- Les niveaux de pollution de l'air au fil du temps (données historiques).

En exploitant les données de l'api on retrouve es indicateurs pm2.5, pm10 et o3 pour la ville de Dakar sur huit(08) jours à partir d'aujourd'hui avec les valeurs moyenne, maximale et minimale pour chaque indicateur. Les donnéesse presentent comme suit


```
"forecast": {
    "daily": {
        "o3": [
            {
                "avg": 11,
                "day": "2023-07-23",
                "max": 67,
                "min": 1
            },
            {
                "avg": 6,
                "day": "2023-07-24",
                "max": 57,
                "min": 1
            },
            {
                "avg": 6,
                "day": "2023-07-25",
                "max": 28,
                "min": 1
            },
            {
                "avg": 5,
                "day": "2023-07-26",
                "max": 33,
                "min": 1
            },
            {
                "avg": 1,
                "day": "2023-07-27",
                "max": 8,
                "min": 1
            },
            {
                "avg": 2,
                "day": "2023-07-28",
                "max": 14,
                "min": 1
            },
            {
                "avg": 4,
                "day": "2023-07-29",
                "max": 20,
                "min": 1
            },
            {
                "avg": 1,
                "day": "2023-07-30",
                "max": 1,
                "min": 1
            }
        ],
        "pm10": [
            {
```

Traitements des données

Aux termes de l'exploration des données, nous avons défini les indicateurs à extraire et à afficher à savoir les taux de particules fines **pm2.5 et pm10** principalement.

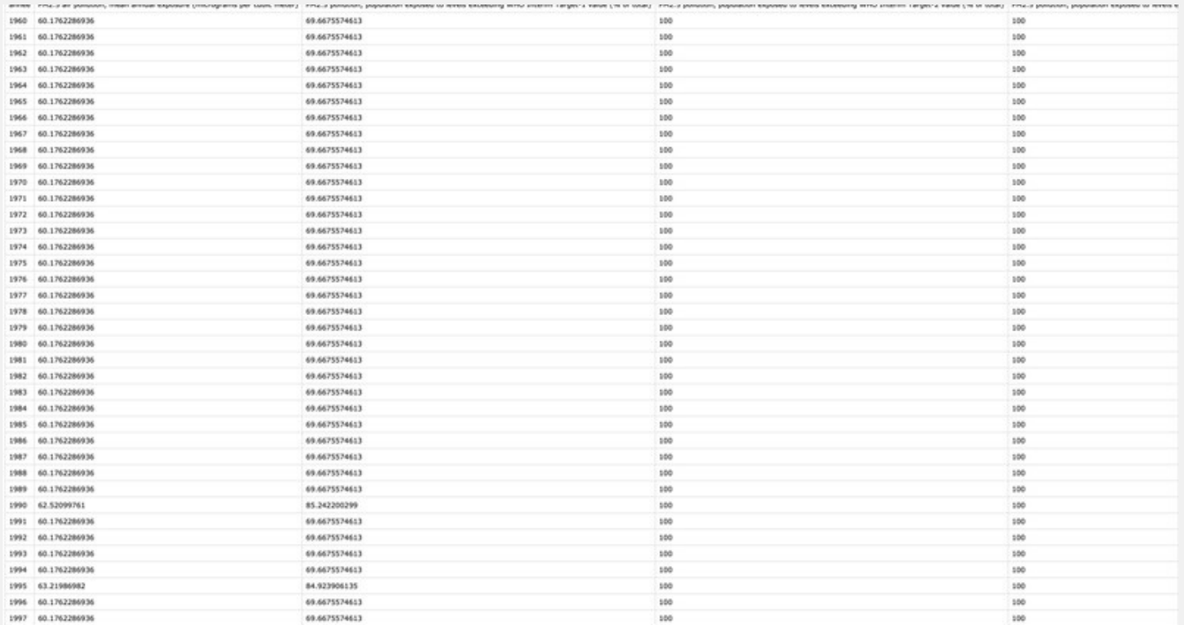
Pour les donnéesde la banque, les traitements suivants ont été effectués:

- Regrouper les cinq(05) datasets en un seul pour une consolidation des données
- Inférence des données manquantes par la moyenne au vu du taux important de valeurs manquantes et de leur distribution
- Pretraitement sur les dates

Pour les données de l'api waqi, les pretraitements effectués sont les suivantes:

- Transformer chaque indicateur en une colonne et un taux (Avg, min, max) en une colonne
- Regrouper l'ensemble des indicateurs dans un seul dataset

Les données obtenues après traitements sont les suivantes



waqi_data2023-07-2617_11.csv — Ouvrir avec Visual Studio Code

	day	avg_o3	max_o3	min_o3	avg_pm10	max_pm10	min_pm10	avg_pm25	max_pm25	min_pm25
0	2023-07-24	9	17	6	11	26	4	22	42	8
1	2023-07-25	8	12	8	9	16	4	18	36	8
2	2023-07-26	8	12	6	6	8	5	11	15	8
3	2023-07-27	10	13	6	24	52	3	26	55	5
4	2023-07-28	13	16	13	124	150	59	85	99	52
5	2023-07-29	13	14	11	73	109	10	58	76	21
6	2023-07-30	9	12	7	5	6	3	8	13	5
7	2023-07-31	10	10	10	6	6	6	13	13	13

Stockage et déploiement

- stockage
- l'orchestration
- automatisation des workflows
- technologies cloud

Airflow

Airflow est un puissant système de gestion de workflow open-source qui joue un rôle central dans notre projet. Il nous permet de planifier, exécuter et surveiller des flux de travail complexes composés de multiples tâches. En utilisant Airflow, nous avons pu définir des dépendances entre différentes tâches, notamment l'extraction, la transformation et le chargement (ETL) des données. Cette coordination automatisée des processus de traitement et de stockage des données assure une exécution fiable et efficace.

Une des caractéristiques clés d'Airflow est la possibilité de planifier des tâches à des horaires récurrents à l'aide de déclencheurs basés sur des événements. PostgreSQL, qui joue un rôle essentiel en tant que base de données de métadonnées pour Airflow, enregistre toutes les informations importantes concernant les workflows, les tâches, les logs et les utilisateurs. De plus, Airflow offre une intégration simple avec divers outils et services couramment utilisés pour le stockage et le traitement des données, tels qu'Amazon S3, Google Cloud Storage, Hadoop et Spark. Cela nous a permis de construire des workflows qui interagissent facilement avec notre solution de stockage, Amazon S3.

Un autre avantage d'Airflow est son déploiement flexible et isolé grâce à l'utilisation de Docker/Docker-compose. Cette approche offre une meilleure isolation des différents composants du système, garantissant ainsi une infrastructure robuste et flexible pour la gestion et le traitement des données. Cela nous a permis de facilement déployé.
En exploitant cette flexibilité, nous avons pu déployer facilement toute l'application sur AWS. Docker a joué un rôle essentiel dans la portabilité de notre projet, nous permettant de créer des conteneurs autonomes contenant toutes les dépendances et configurations nécessaires.

En intégrant Airflow dans notre projet, nous avons considérablement amélioré la planification des tâches, la gestion des workflows et la visibilité de l'état de nos processus. Cela nous permet de concentrer nos efforts sur l'analyse et l'utilisation des données plutôt que sur la gestion manuelle des tâches.

Simple Storage Service

Amazon S3 (Simple Storage Service) est notre service de stockage objet principal dans le projet. Nous avons créé un bucket S3 pour stocker les fichiers au format CSV contenant les données prétraitées. Grâce à l'utilisation de la bibliothèque boto3 et à la planification automatisée d'Airflow, les fichiers CSV sont envoyés de manière transparente dans leurs buckets respectifs sur Amazon S3.

Nous avons organisé les fichiers CSV dans une arborescence pour faciliter leur gestion et leur accès. Par exemple, les données peuvent être stockées dans les chemins suivants :

- s3://airqualitydatastorage/data/input/wb_data/
- s3://airqualitydatastorage/data/input/waqi_data/

En utilisant Amazon S3 comme service de stockage, nous bénéficions d'une solution fiable, évolutive et sécurisée pour stocker nos données prétraitées et prêtes à être traitées.

AWS Glue Crawler & Glue Data Catalog

AWS Glue Crawlet

Pour explorer et découvrir les fichiers CSV dans notre bucket S3 de données, nous avons configuré AWS Glue Crawler. Ce service joue un rôle crucial en analysant les données, en inférant automatiquement le schéma des fichiers CSV et en créant des tables correspondantes dans le catalogue de données Glue.
Le Crawler est configuré pour utiliser un rôle IAM spécifique avec des autorisations appropriées pour accéder aux fichiers CSV dans Amazon S3 et créer des tables dans le catalogue de données Glue. Grâce à cette configuration, nous pouvons facilement mettre à jour et gérer les tables dans le catalogue sans intervention manuelle.

Glue Data Catalog

Pour explorer et découvrir les fichiers CSV dans notre bucket S3 de données, nous avons configuré AWS Glue Crawler. Ce service joue un rôle crucial en analysant les données, en inférant automatiquement le schéma des fichiers CSV et en créant des tables correspondantes dans le catalogue de données Glue.
Le Crawler est configuré pour utiliser un rôle IAM spécifique avec des autorisations appropriées pour accéder aux fichiers CSV dans Amazon S3 et créer des tables dans le catalogue de données Glue. Grâce à cette configuration, nous pouvons facilement mettre à jour et gérer les tables dans le catalogue sans intervention manuelle.

Amazon Athena

Nous avons configuré Amazon Athena pour accéder au catalogue de données Glue et aux tables créées par le Crawler. Amazon Athena est utilisé pour exécuter des requêtes SQL sur les données stockées dans le catalogue de données Glue. En utilisant ce service serverless, nous n'avons pas besoin de provisionner des ressources spécifiques pour exécuter des requêtes, ce qui rend l'interrogation des données plus efficace.
Nous utilisons Amazon Athena pour interroger les tables créées par le Glue Crawler à l'aide de requêtes SQL. Athena se charge alors d'extraire et d'analyser les données en fonction du schéma inféré par le Crawler. Cela nous permet d'obtenir des informations précieuses à partir de nos données stockées dans Amazon S3 sans effort supplémentaire.
Grâce à cette architecture de stockage et de déploiement, nous avons pu construire une solution complète et efficace pour la gestion, le traitement et l'analyse de nos données. L'utilisation d'Airflow pour orchestrer les workflows, Amazon S3 comme stockage objet, AWS Glue Crawler pour l'exploration des données et Amazon Athena pour les requêtes nous offre une infrastructure robuste et évolutive pour tirer le meilleur parti de nos données.

Visualisation

- Objectifs
- Services utilisés et roles
- Résultats

Objectifs

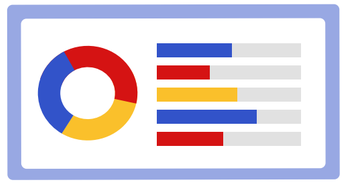
Nous rappelons que l'objectif final du projet était de mettre à disposition des indicateurs clefs sur le niveau de toxicité de l'air. Notamment identifier:

- Le niveaux de toxicité journalier de l'air mesurer par le pm10(Main pollutant)
- La qualité journalière de l'air, indiqué par le pm2.5
- La qualité moyenne annuelle de l'air
- Et la population annuelle exposé à une mauvaise qualité de l'air suivant les niveaux de criticité de l'OMS.

Pour arriver à un tel résultat, il nous faudra mettre en place un dashboard sur la base des schémas enregistrés (depuis AWS Athena dans notre cas).

Quicksight

La visualisation des indicateurs identifiés nécessite un outils de génération de dashboard. Une multitude de solution s'offraient à nous. Pour des raisons multiple, nottament l'intégration avec les autres services utilisés nous avons choisi Amazone QuickSight.



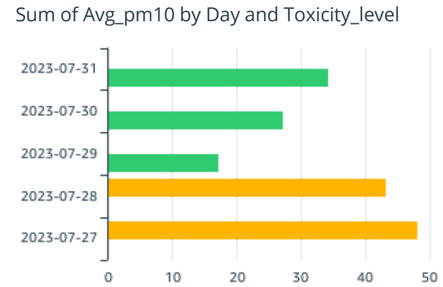
Amazon QuickSight est un service d'analyse de données interactif proposé par Amazon Web Services (AWS). Il permet aux utilisateurs d'explorer, de visualiser et de partager facilement des données à partir de différentes sources de données, telles que des bases de données, des services cloud, des fichiers plats, etc.

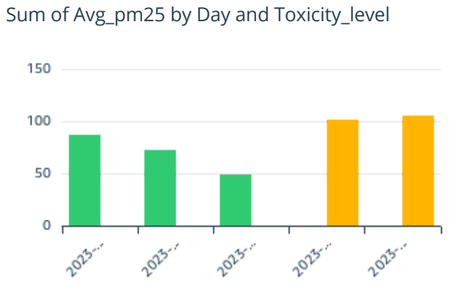
Avec QuickSight, nous peuvons créer des tableaux de bord interactifs, des graphiques et des visualisations en utilisant une interface simple et intuitive. Le service prend en charge diverses options de visualisation, notamment des graphiques à barres, des graphiques linéaires, des diagrammes circulaires, des cartes géographiques, etc.

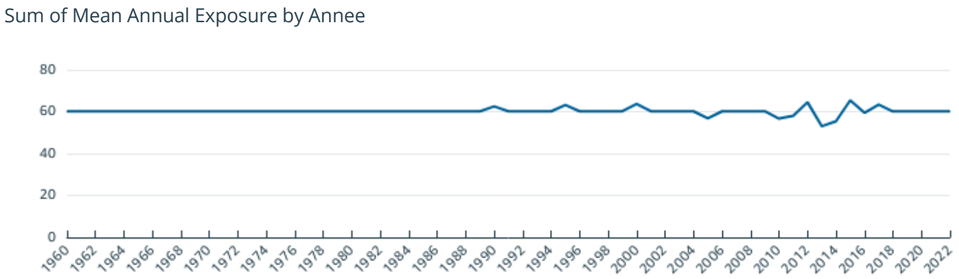
QuickSight est intégré nativement avec Amazon Athena, ce qui facilite la connexion et l'analyse des données stockées dans le Data Lake S3 via Athena. Nous pouvons accéder aux résultats de nos requêtes Athena directement dans QuickSight. En combinant Amazon QuickSight et Amazon Athena, nous disposons d'une solution complète d'analyse de données en libre-service, adaptée à l'exploration et à la visualisation rapides de vastes volumes de données stockées dans le Data Lake S3 avec la puissance du SQL.

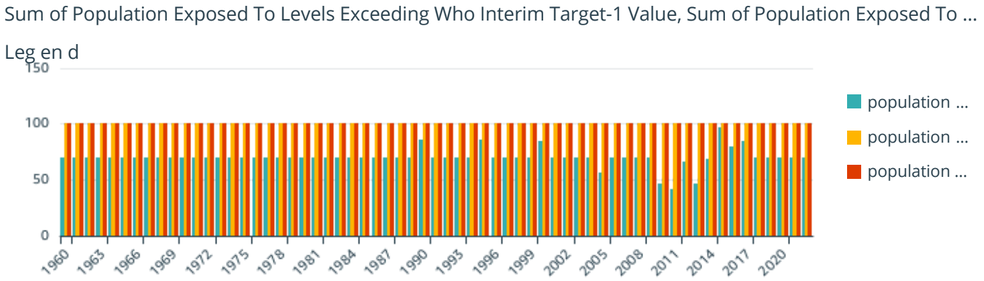
Résultats final

Nous n'avons qu'à définir une fréquence de génération du dashbord édité(Journalier dans nôtre cas).









Perspectives

- Plus de Maturité avec l'intégration continue
- Consolidation des données
- Application mobile

Perspectives du projet

Ce sont des points d'amélioration du projet

Chaine CI/CD

Automatiser encore plus le processus entre les modifications qui seront apportéesau niveau des task pour les deployer automatiquement au niveau de AWS et générer de nouveaux graphes

Consolidation des données

Encore plus avoir des données comme les causes de la pollution, les données météorologiques, les données sur les autres indicateurs comme le soufre ou le monoxyde de carbone, ou encore les données sur les autres pays pour pour avoir plus de digrammes explicites

Application mobile

Mettre en place une application mobile pour les citoyens qui veulent avoir des informations sur le taux de pollution