Projet #2

Mise en Place et Optimisation d'une Infrastructure Data sur le Cloud

Année 1 - **L'École Multimédia**

Sommaire

Contexte

Présentation

Vous êtes embauché en tant qu'ingénieur data dans une entreprise qui souhaite moderniser son infrastructure de gestion de données.

Votre mission consiste à concevoir et mettre en œuvre une infrastructure data sur le Cloud, à collecter et intégrer des données provenant de diverses sources, et à fournir des analyses et visualisations interactives pour soutenir la prise de décision.

Le projet comprend plusieurs phases :

1. Mise en place de l'infrastructure Cloud
2. Création d'un pipeline de collecte et de traitement des données
3. Exploration et visualisation des données
4. Documentation complète du processus.

Contraintes

- **Travail en autonomie** : Vous travaillerez seul sur ce projet, ce qui implique une gestion efficace du temps et des ressources.
- **Utilisation d'AWS Free Tier** : Le projet doit être réalisé en utilisant les services gratuits d'AWS (S3, Glue, Lambda, RDS) afin de garantir une accessibilité sans frais supplémentaires.
- **Respect des normes de sécurité** : L'infrastructure Cloud doit respecter les bonnes pratiques de sécurité, notamment :
    - Configuration des accès et permissions via **IAM (Identity and Access Management)**.
    - Activation du **chiffrement des données** pour le stockage sur S3 et les bases de données.
    - Mise en place d'un système de **monitoring et de journalisation** (AWS CloudWatch, AWS CloudTrail).
- **Documentation complète** : Chaque étape du projet doit être clairement documentée, notamment :
    - La **mise en place du Data Lake et du Data Warehouse**.
    - Les **transformations ETL** et leur impact sur la qualité des données.
    - L'**analyse des données** et les insights obtenus.
- **Utilisation de Git et GitHub** : Tout le projet doit être versionné sur GitHub, avec des **commits réguliers**, un **README détaillé**, et une bonne structuration du code.
- **Respect des délais** : Le projet doit être livré sous la forme d'une **archive ZIP** (nommée votreprenom_votrenom_projet2.zip) et dans les délais imposés sous peine de pénalisation.

Libertés

Choix des outils complémentaires : Même si AWS est recommandé, vous êtes libre d'expérimenter avec d'autres services **Cloud gratuits** (Google Cloud Free Tier, Azure Free) pour **comparer** les performances et les coûts.

Personnalisation du pipeline ETL : Vous pouvez choisir les **étapes de transformation** les plus pertinentes en fonction des données (nettoyage, normalisation, agrégation, enrichissement…).

Approche d'analyse : Vous êtes libre d'explorer les données sous différents angles :

- **Analyses statistiques classiques** (distributions, moyennes, médianes…).
- **Détection d'anomalies** et gestion des valeurs aberrantes.
- **Segmentation et clustering** si pertinent.

Design du Dashboard : Vous pouvez organiser votre tableau de bord comme vous le souhaitez :

- Sélection des **visualisations interactives les plus pertinentes** (graphiques, cartes, indicateurs…).
- Utilisation de **Plotly et Bokeh** pour maximiser l'interactivité et la clarté des résultats.

Organisation du Code et de la Documentation :

- Vous êtes libre de structurer votre **dépôt GitHub** de manière à optimiser la **lisibilité et la réutilisation**.
- Vous pouvez ajouter des **tutoriels ou guides d'utilisation** pour faciliter la compréhension de votre solution.

Objectifs

Vous devrez avoir réalisé les éléments suivant :

1. Concevoir et déployer une infrastructure data sur le Cloud, incluant un Data Lake et un Data Warehouse, en utilisant les services AWS.
2. Mettre en place un pipeline ETL pour la collecte, la transformation, et le chargement des données provenant d'APIs et de sources web.
3. Explorer et analyser les données collectées à l'aide de librairies Python telles que Pandas et Seaborn.
4. Créer des visualisations interactives et des dashboards avec les librairies Plotly et Bokeh pour faciliter la prise de décision.
5. Collaborer efficacement via Git et GitHub pour la gestion des versions du code, la documentation, et le suivi des tâches.

Données :

Le marché de la **Data Science** évolue rapidement, et les entreprises recherchent constamment des profils spécialisés. Ce projet vise à **analyser en temps réel** les tendances des offres d'emploi en Data Science en extrayant et stockant des **données issues de sites d'emploi** (Indeed, Welcome to the Jungle, LinkedIn).

Vous devez impérativement :

- Scraper au moins 1 site.
- Utiliser au moins 1 api.

Objectifs principaux

1. **Scraper et collecter** des annonces d'emploi en Data Science.
2. **Nettoyer et structurer** les données pour un stockage optimal.
3. **Créer un Data Lake** pour conserver les offres d'emploi en format brut.
4. **Stocker et interroger** les données via une base SQL (AWS RDS).
5. **Créer des visualisations interactives** pour analyser le marché de l'emploi.

Scraping Facile (Pages Statiques)

- Welcome to the Jungle
- AI-Jobs.net
- Freelance.com
- Malt

Scraping Modéré (Protection Anti-Bot)

- Indeed (nécessite headers User-Agent et délais entre requêtes)
- Glassdoor

APIs Disponibles (Alternative au Scraping)

- Pôle Emploi API
- Adzuna API

Rendu final

Votre rendu final prendra la forme d'une archive Zip et devra comporter les éléments suivants :

1. Infrastructure Cloud déployée sur AWS :
    a. Documentation sur la mise en place du Data Lake et du Data Warehouse.
    b. Scripts ou configurations utilisés pour la création et la gestion de cette infrastructure.
2. Pipeline ETL opérationnel :

    a. Scripts Python ou configurations utilisés pour le pipeline ETL.
    b. Description détaillée des étapes de collecte, transformation et chargement des données.
3. Exploration et analyse des données :
    a. Jupyter notebooks ou scripts Python utilisés pour l'exploration des données.
    b. Résultats des analyses, incluant des visualisations exploratoires avec Seaborn.
4. Dashboards interactifs :
    a. Dashboards créés avec Plotly et Bokeh, intégrant des visualisations interactives des données analysées.
    b. Documentation sur la création et l'utilisation des dashboards.
5. Documentation et gestion de projet sur GitHub :
    a. Dépôt GitHub contenant tout le code source, les scripts, les configurations, et la documentation complète du projet.
    b. Historique des versions et suivi des contributions via Git et GitHub.
6. Diaporama détaillant toutes les étapes du projet, les décisions techniques, les résultats obtenus, et les recommandations pour l'avenir.

Vous devez livrer une archive de votre livrable avec l’ensemble des éléments

Cette archive aura comme titre votre nom et prénom avec votre classe.

Exemple: `steve_jobs_projet2_AIA01.zip`

Attention : Un rendu non livré ou en retard vous pénalise pour la certification

Evaluations

- Pertinence et efficacité de l'infrastructure Cloud déployée sur AWS, en termes de gestion des données et de sécurité.
- Qualité et fiabilité du pipeline ETL, incluant la capacité à gérer les erreurs, l'efficacité des transformations, et la robustesse du chargement des données.
- Profondeur et rigueur de l'exploration des données, démontrée par l'utilisation adéquate des librairies Python et la qualité des analyses statistiques.
- Créativité et interactivité des dashboards créés avec Plotly et Bokeh, et leur utilité pour la prise de décision.
- Clarté et exhaustivité de la documentation, ainsi que la gestion efficace du projet via Git et GitHub.
- Présentation professionnelle du projet, incluant une explication claire des choix techniques, des résultats, et des perspectives d'amélioration future.

Compétences à valider

B-01	Identifier les besoins architecturaux en enquêtant sur les contraintes techniques, opérationnelles et normes en vigueur, afin d'établir un cadre conforme aux exigences de l'entreprise.
B-03	Élaborer des modèles de données logiques et physiques (entité-relation, les modèles de données en étoile...) qui correspondent au cahier des charges établi.
B-04	Concevoir des structures de bases de données adaptées à divers types de données, en tenant compte des performances, de la sécurité, de l'évolutivité, et du volume des données, pour une gestion optimale du Big Data.
C-02	Établir un pipeline de données à travers des processus ETL/ELT pour le transfert et la transformation des données entre différentes bases, en utilisant des outils de programmation, afin de répondre aux spécifications du cahier des charges.
C-04	Surveiller les flux de données pour assurer la qualité et le respect de la politique de gouvernance, en vue de maintenir les normes, la sécurité et la confidentialité dans les pipelines de données.

Conseils

- Bien prendre le temps d’analyser le brief et comprendre le client
- Organisez-vous et planifiez votre travail : donnez vous des objectifs intermédiaires
- Planifiez des sessions de travail régulière
- Utilisez Git pour versionner votre code dès le départ
- Ne jamais être trop ambitieux
- Faites directement des documents du livrable
- Mettez en oeuvre les bonnes pratiques vues en cours
- Refactoriser pour éviter le code redondant
- Soignez la qualité de votre code (commentaires, indentation)
- N’attendez pas la fin pour commencer à travailler sur votre projet
- Pensez à la qualité du résultat !