Module : Data Pipeline

Examen de validation des compétences

Règles de l'examen

Contexte

En utilisant les outils présentés dans ce module uniquement, vous répondrez aux cas d'usage décrits ci-dessous. Vous devrez constituer des groupes de 2 à 3 personnes. Tous les membres du groupe doivent pouvoir expliquer votre rendu. Vous utiliserez un contexte d'exécution local pour utiliser les outils, tel que nous avons pu les déployer pendant les TPs.

Notation

La même note sera attribuée à l'ensemble du groupe. Vos rendus seront évalués à part égal selon les 3 critères suivants :

- Exploitation au maximum des fonctionnalités des outils.
- Qualité du rendu : clarté du code, reproductibilité, documentation de votre projet.
- Pertinence de la réponse au problème.

Livrables

Vous devrez déposer l'ensemble des fichiers, documents, notes, … sur un repository Github. Puis, vous transmettrez par mail le lien de partage de votre repository ainsi que la liste des membres de votre groupe. Vous pouvez documenter votre projet en anglais ou en français. Si vous utilisez des composants complémentaires (à ceux installé pour le cours), vous devrez documenter leur installation sur un environnement linux, ex : Providers Airflow.

Sujets de l'examen

Contexte

Vous êtes en relation avec un cabinet de conseil en stratégie politique. Le cabinet a pour mission d'effectuer une analyse et des résultats de l'élection présidentielle Française du 10 et 24 Avril 2022. Le cabinet recherche une société partenaire pour lui fournir les données. Vous devez collecter et préparer des données disponibles publiquement sur internet afin de gagner l'appel d'offre :

- **Apache NiFi** : Collecter différentes sources de données pertinentes (élections précédentes, recensement de la population française, données sur l'emploi, autres sources …)
- **Apache Spark** : Préparer les données (nettoyage des données, suppression des doublons, valeurs aberrantes, jointure/croisement des sources de données, changement de type, nom des colonnes, etc …)
- **Apache Airflow** : Orchestrer et automatiser les phases de collecte et de préparation des données
- Vous devez fournir un argumentaire qui valorise travail, explique vos choix et comment utiliser les données que vous avez préparé.

Exemple cas d'usage : https://www.data.gouv.fr/fr/reuses/50-1-dis-moi-ou-tu-habites-je-te-dirai-pour-qui-tu-votes/

Arborescence de votre repository git

README.md	-> Fichier contenant votre argumentaire
airflow	-> Dossier contenant vos fichiers airflow (AIRFOW_HOME), ne pas déposer les fichiers : airflow.db, logs, standalone_admin_password.txt
scripts	-> Dossier contenant vos script pyspark à exécuter avec « spark-submit »
nifi	-> Dossier contenant l'export JSON de vos flux Nifi.
data	-> Dossier de stockage des données (ne pas sauvegarder sur git)
spark	-> Dossier d'installation de Spark (ne pas sauvegarder sur git)

Exemple : https://github.com/apointeau/ece-paris-example-exam