Syllabus projet

Année : **2021-2022**

Enseignant(s)	Email(s)
JUMELLE Maxime	mjumelle2@myges.fr

Projet AWS

1 Matières, formations et groupes

Matière liée au projet :

Formations : -

Nombre d'étudiant par groupe : **3 à 4**

Règles de constitution des groupes: **Libre**

Charge de travail estimée par étudiant : **14,00 h**

2 Sujet(s) du projet

Type de sujet : **Imposé**

Stack Data et ML sur AWS

3 Détails du projet

Objectif du projet (à la fin du projet les étudiants sauront réaliser un...)

L'objectif de ce projet est d'aider une entreprise de VTC basée à New-York à utiliser une partie de ses services sur le Cloud AWS.

Le projet à mettre en oeuvre doit répondre au cahier des charges suivant.

- Automatiser l'alimentation des tables DynamoDB.
- Créer des extracts journaliers de données d'entraînement pour le modèle ML.
- Entraîner automatiquement des modèles ML et les exposer via des pipelines SageMaker.
- Effectuer des inférences du modèle à partir de données ingérées dans une file SQS.

Descriptif détaillé

Une entreprise de VTC basée à New-York dispose de toutes les données liées à de trajets effectués par des utilisateurs. Cette dernière a besoin d'agréger toutes ces données pour en faire ressortir des informations diverses.

Chaque fichier contient tous les trajets réalisés pour un mois donné. Les informations agrégées devront être calculées automatiquement chaque jour. On suppose que chaque fichier est ajouté et/ou complété au fur et à mesure chaque jour dans un bucket S3.

Les données et les référentiels sont accessibles ici : https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page

À noter que dans le cadre de ce projet, nous considérons uniquement les trajets Yellow Taxi effectués après Janvier 2018.

1) Automatiser l'alimentation des tables DynamoDB

est d'automatiser la création et l'ingestion de données dans des tables DynamoDB qui contiennent des agrégations de données brutes.

Les différentes informations nécessaires sont les suivantes :

- Le chiffre d'affaire réalisé par conducteur et par jour.
- Le nombre de trajets effectués par zone et par jour.

2) Créer des extracts journaliers de données d'entraînement pour le modèle ML

À l'aide d'un pipeline identique ou similaire, des extracts des données brutes doivent être exportés vers un bucket S3 afin de permettre aux Data Scientists de construire des modèles ML.

3) Entraîner automatiquement des modèles ML et les exposer via des pipelines SageMaker

Afin de proposer une expérience utilisateur optimale, l'entreprise souhaite estimer automatiquement la durée d'un trajet en fonction de plusieurs paramètres (zone de départ, zone d'arrivée, heure de départ, etc). Afin que cette estimation soit la plus précise possible, un algorithme ML doit utiliser des données récoltées sur un historique de 2 semaines.

4) Effectuer des inférences du modèle à partir de données ingérées dans une file SQS

Une fois le modèle exposé, l'inférence doit être réalisée par un système de file d'attente. On utilisera SQS avec un code test afin de simuler des demandes en situation réelle avec une fréquence proche de la réalité.

Ouvrages de référence (livres, articles, revues, sites web...)

Outils informatiques à installer

4 Livrables et étapes de suivi

1	Rendu final		**vendredi 21/01/2022 14h00**

5 Soutenance

Durée de présentation par groupe :	**15 min**	Audience :	**Devant la promotion**
Type de présentation :	**Présentation / PowerPoint**		
Précisions :			