Examen

Vous devez constituer des équipes de 6 personnes, pensez à répartir les équipes sur les deux sujets

Sujet Partie 1 : Création d'une plate-forme vidéo

Pitch

On souhaite mettre en œuvre un service de partage de vidéos, qui pourrait être alimenté par des producteurs de contenus. Le modèle le plus pertinent est naturellement celui de YouTube (*of curse !*), qui propose toute une panoplie de services de cet ordre.

L'organisation qui gère ce service doit s'occuper de deux problématiques distinctes :

1. un volet technique (l'hébergement des vidéos)

Le service en question héberge différents types de contenus, comme des films, des séries, des vidéos pédagogiques et peut-être d'autres catégories que vous imaginerez. D'autre part, les fournisseurs peuvent eux-mêmes avoir plusieurs statuts. Certaines vidéos dont fournies par des entreprises de l'industrie du divertissement ; d'autres par des auteurs « *amateurs* » profitant de la plate-forme pour diffuser leurs créations.

Ces différents types de contenus entraînent également des questions sur les licences qui leurs sont associées.

2. un volet financier (comment rendre ce service « rentable »).

Naturellement, nous souhaitons que le service puis être viable financièrement. Il faut donc trouver les moyens de de générer des revenus à partir de la diffusion des vidéos. Là aussi, deux pistes peuvent être exploitées, qui sont :

- d'une part un accès libre, qui serait par exemple financé par de la publicité insérée dans les vidéos ;
- d'autre part des abonnemeents payants, qui donneraient accès à des contenus protégés par leurs producteurs ;
- on peut imaginer aussi (cf. YouTube) pouvoir louer des films, pour une durée déterminée.

La plate-forme doit bien sûr se rémmunérer elle-même, mais aussi rétribuer les créateurs de contenus, soit sous forme de droits d'auteurs pour l'industrie cinématographqiue, soit sous forme de « *monétisation* » des contenus, basée sur le nombre de vues (oou une autre méthoe que vous choisiriez), pour les vidéos crées par des utilisateurs.

Données

Pour parvenir à atteindre ses objectifs, l'organisation a besoin de savoir comment la plate-forme est utilisée par le public et par les créateurs de contenus.

L'un des objectifs de l'organisation est de pouvoir analyser les données qui lui permettraient de « fidéliser » davantage d'utilisateurs et/ou d'accroître ses revenus. Par ailleurs, un de ses objerctifs est d'améliorer la qualité du service en mettant à disposition de meilleurs vidéos.

Pour pouvoir tirer des enseignements fiables, les gestionnaires de la plate-forme auront besoin de mesures d'audience les plus précises possible.

Naturellement, la plate-forme s'engage à respecter les données personnelles des utilisateurs.

Les tâches qui vous sont assignées sont :

1. Modéliser le domaine conceptuel du service (de constuire le catalogue/référentiel/ontologie de l'organisation)

Vous ferez en sorte de prendre compte dans votre modèle, non seulement la description des contenus en eux-mêmes, mais aussi naturellement les caractéristiques des utilisateurs, celles des « **ayant-droit** » (ceux que vous devez rémunérer) ainsi que celles des personnes/institutions susceptibles de vous apporter des revenus.

2. Chercher des sources dans lesquelles vous pourrez collecter des données utiles à l'enrichissement de votre jeu de données interne

Les données sont pour partie produites par l'activité de la plate-forme, mais vous chercherez à bénéficier de données déjà existantes, en particulier pour tout ce qui concerne la description les contenus. Votre tâche sera d'identifier sur Internet des sources de données (en particulier des API) qui pourraient vous fournir ces informations.

Jne fois ces sources repérées, vous en examinerez la structure et vous établirez une correspondance avec le modèle/rférentiel/ontologie que vous avez vous-mêmes créé.

3. Proposer une chaîne de traitement de l'information pour harmoniser les formats de données collectées depuis différentes surces

Vous proposerez ensuite une chaîne de traitement de type ETL pour rendre les (ou les) jeu(x) de données externes confrome(s) à vos attentes. Vous imaginerez une solution pour automatiser ce processus en intégrant dans la chaîne de transformation l'ontologie que vous ave conçue.

4. Proposer des critères d'analyse des données pour évaluer dans quelle mesure l'application mise en œuvre permet d'atteindre les objectifs que l'organisation s'est fixée.

L'organisation qui gère la plate-forme a besoin d'analyser les données de la plate-forme pour savoir si elle génère des revenus suffisants. Décrivez quels critères, selon vous, permettraient de décider d'une stratégie à adopter vis-à-vis des utilisateurs. Pour cela, vous êtes libres de faire des hypothèses de travail qui vous arrangent.

5. Offrir un moyen aux personnes de l'organisation d'avoir accès à des analyses de données

Décrivez comment vous mettriez en œuvre des outils permettant aux gestionnaires de la plate-forme d'avoir accès aux données. Cherchez notamment des outils Open Source permettant d'effectuer des requêtes d'analyse sur des bases de données.

Rendu

Les questions qui vous sont posées ne nécessitent pas d'écrire du code.

1. Pour ce qui concerne l'ontologie, vous aurez à livrer le code textuel exporté par Protégé ; nous aimerions aussi une représentation graphqiue de cette ontologie, que vous pourrez produire en installant le plugin VOWL pour Protégé.

- Plugin VOWL

2. Pour les jeux de données externes, nous vous demanderons de nous fournir la liste des sources que vous avez jugées intéressantes et de slélectionner dans ces jeux de données celles qui sont reliées à vos propres intérêts.

3. Pour la chaîne ETL, nous vous demanderons de produire un schéma d'architecture, intégrant notamment l'utilisation de votre ontologie pour transformer les données externes selon votre propre format.

Nous ne vous demanderons pas d'écrire le code correspondant. Toutefois, vous prendrez un échantillon des données externes pour montrer quelles exemples de données seraient éventuellement à corriger.

4. Pour la quatrième question, votre tâche sera de produire des critères d'évaluation reposant sur les données définies par votre ontologie. En partant d'hypothèses arbitraires, vous expliquerez dans quels cas le service fourni par la plate-forme est rentable ou non. Donnez 4 ou 5 exemples de critèes possibles.

5. Pour la dernière question, vous aurez à produire un schéma d'architecture intégrant :

- le(s) outil(s) que vous mettrez à disposition pour effectuer l'analyse des données
- quel(s) outil(s) de communication vous privilégieriez pour implémenter le travail collaboratif et les différents messages

Chaque partie rédactionnelle ne devra pas excéder quelques paragraphes.

Vous priviliégierez les représentations visuelles

Sujet Partie 2 Analyse de données

DataOps avec Python données du Titanic

Description :

Vous êtes responsable de la gestion des données pour une entreprise. Mettez en place le processus de traitement des données.

1. Combien de femme de moins de 18 ont survécu
2. Parmi ces femmes (question 1) déterminer la répartition par classe sur le bateau.
3. Déterminer si le port d'embarquemen a une influence sur la survie (calculer la répartiti des morts et des survivants en fonction du port de départ)
4. Déterminer la répartition par sexe et par âge des passagers du navire
5. Conclusion, écrire un document qui vous présentera votre analyse lors de la soutenance.

Indications : - Utilisez Python pour automatiser certaines tâches répétitives. - Intégrez des bibliothèques telles que Pandas, NumPy, ou d'autres outils modules Python.

Tâches :

- Proposez une méthode de récupération des données pour optimiser l'analyse.
- On aimerait à partir des données bruts du titanic mettre en place un modèle (JSON) plus simple que l'on pourrait conserver (enregistrer) pour nos traitements. Vous utiliserez le modèle suivant :

```
passenger = {
    "sex"
    "class"
    "age"
    "survived"
    "price"
    "embarked"
}
```

1. **Pipeline :**
   - Mettez en place le pipeline de bout en bout pour la récupération des données.

- Créez des fonctions pour mettre en place ce pipeline. Vous pouvez par exemple créer les fonctions suivantes.
- Pensez à nettoyer les données (données aberrantes)

```python
# request data
def requestUrl(url):
    pass

# création d'un modèle à partir des données téléchargées
def extract_model(url):
    pass

# nettoyage/formatage des données
def transform(data):
    pass

# Enregistrement des données
def load(data):
    pass
```

Mise en œuvre des Changements avec Python :

3. **Documentation :**
   - Rédigez une documentation complète expliquant le nouveau pipeline DataOps, y compris les choix technologiques, les dépendances, et les instructions pour les utilisateurs futurs.

Rendu : - Un script Python optimisé pour le pipeline DataOps. - Un rapport documentant les améliorations apportées, les choix technologiques, et vos résultats.