Rapport Technique: Banque de Question

Jean-Thomas Baillargeon

September 2020

1 Stratégie d'acquisition des données

Cette section présente la stratégie d'acquisition des données, soit la source des données et la présentation des données.

1.1 Sources des données utilisées

Afin d'obtenir les données pour l'application banque de question, il est nécessaire de fournir aux créateurs un incensitif pour créer des questions. La stratégie retenue a été de leur permettre de créer des questions pour l'examen. Les données sont entreposées dans monPortail et sont téléchargées manuellement via l'interface web. La périodicité d'extraction n'a pas été décidée.

1.2 Présentation des données

Les données utilisées dans l'application sont sous forme liste d'objet json. Chaque élément de la liste correspond à une question et comprend plusieurs champs. Les champs importants sont les champs `typeQuestion`, `question`, `reponse`. Les autres champs tels `matricule` et `nomComplet` ne sont qu'utilisés pour identifier l'étudiant qui a écrit la question. Des exemples de données sont fournis en annexe.

2 Technologies utilisées

Cette section les choix technologiques qui ont été faits pour ce projet. Dans un premier temps, le langage de programmation est présenté et justifié. Dans un deuxième temps, le choix des technologies de persistance est présenté et justifié.

2.1 Langage de programmation

Le langage de programmation choisi pour ce projet est Python. Python a été choisi, car il s'agit d'un choix rendant l'application fiable, vu l'expérience des (du) développeurs principaux. Python est un langage qui respectera les critères d'extensibilité, car la charge attendue et les garanties de offertes sont facilement atteignables. Fianelement, Python a été choisi pour sa maintenabilité vu l'aisance d'utilisation soutenue par les différents *package* développés et testés par la communauté.

2.2 Technologie de persistance

La technologie de persistance qui a été choisie dans ce projet est la base de données orientée document MongoDB. Ce moteur de base de données a été choisi pour stocker les données qui sont déjà sous forme de document. De plus, comme le cas d'usage principal de l'application est de présenter des questions, la possibilité d'avoir les points de données complètement contigües en mémoire supporte considérablement les considérations d'extensibilité de l'application.

3 Processus d'ETL

Cette section présente les différentes étapes nécessaires afin d'avoir les données prêtes pour l'utilisation dans l'application. Premièrement, l'extraction des données est présentée puis les étapes de transformations. Une vue à plus haut niveau est présentée à l'illustration 1

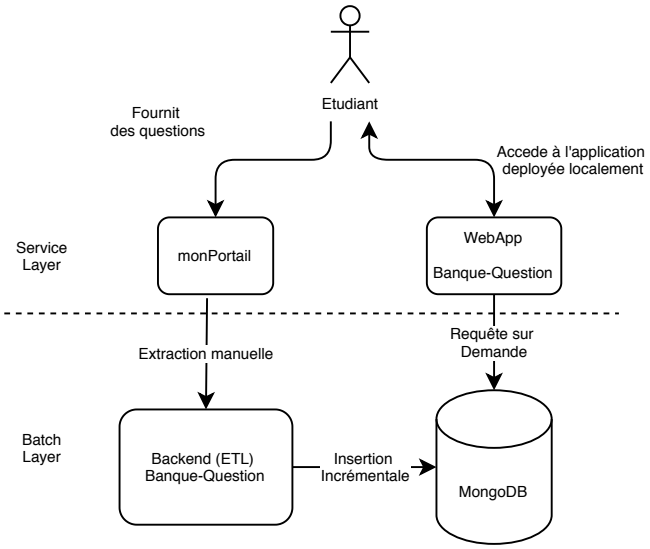

Illustration 1: Architecture du flow des données

3.1 Obtention des données initiales et incrémentales

La première étape est d'obtenir les données de la source et de les intégrer dans la base de données de l'application. L'illustration 2 présente ce processus.

Afin d'obtenir les données pour l'application, un utilisateur extrait manuellement les données du site monPortail et les enregistre dans un disque dur. Chaque fichier représente une liste de questions par étudiant, tel que présenté à la section 1.2

Un processus de batch obtient tous les fichiers sur le disque dur. Afin de s'assurer que les questions ne sont pas ajoutées en double, un script vérifie si le fichier a déjà été ajouté dans la base de données. La comparaison est faite avec le *hash* du contenu du fichier, qui est sauvegardé dans une table `log`. Si le *hash* sauvegardé et celui du fichier courant ne sont pas les mêmes, les questions associées à ce fichier sont supprimées. Cette routine permet d'acquérir des données de façon incrémentales.

Finalement, le processus d'ETL anonymise les données, insère les questions une par une dans la table `questions` et insère le *hash* du contenu du fichier dans la table `log`.

3.2 Transformation des données

Les transformations appliquées sur les données sont faites au niveau des questions individuelles. Le champs `sourceFile` est ajouté afin de connaitre la provenance de chaque question et de l'associer à un hash lors de l'ajout incrémental des données. Le champ `nomComplet` est enlevé afin de protéger la vie privée des étudiants. Un exemple de transformation des données est présenté à l'annexe 2.

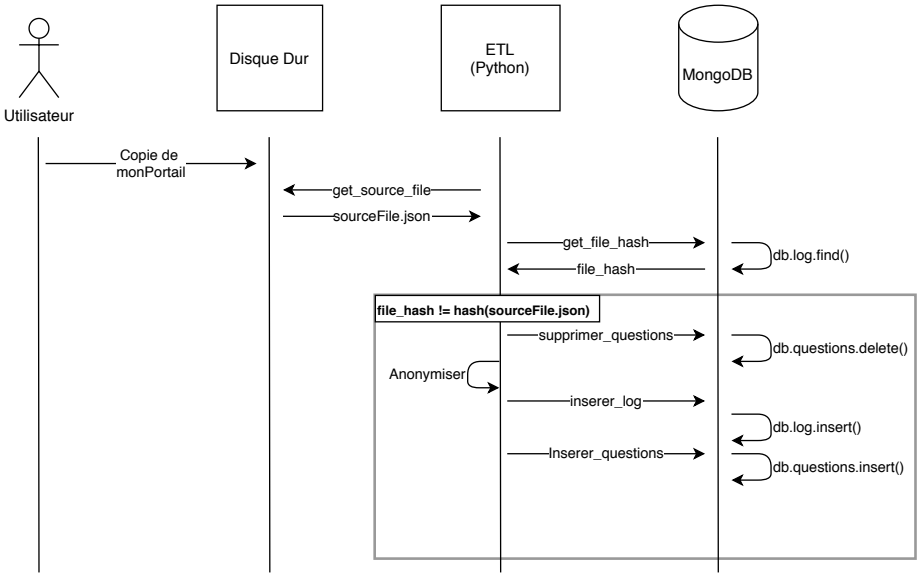

Illustration 2: Processus d'extraction de l'application

Annexe - Exemple de donneés

Présentation de données source qui sera utilisé dans l'application **banque de question**. Ce point comprend 2 questions qui ont été écrites par Jean-Thomas Baillargeon.


```
[{
    "matricule":"000000000",
    "nomComplet": "Jean-Thomas Baillargeon",
    "typeQuestion": "vof",
    "question": "on peut utiliser windows de façon efficace",
    "reponse": "FAUX, il s'agit d'un système d'exploitation inférieur"
},{
    "matricule":"000000000",
    "nomComplet": "Jean-Thomas Baillargeon",
    "typeQuestion": "qrc",
    "question": "Nommez deux avantages d'utiliser une distribution de Linux",
    "reponse": "1 - Pouvoir dire I use linux BTW et 2 - ZSH"
}]
```

Annexe - Exemple de transformation des données

Présentation de transformation appliquée sur une donnée lors de son insertion.

```
{
    "matricule":"000000000",
    "nomComplet": "Jean-Thomas Baillargeon",
    "typeQuestion": "vof",
    "question": "on peut utiliser windows de façon efficace",
    "reponse": "FAUX, il s'agit d'un système d'exploitation inférieur"
}
```

devient

```
{
    "matricule":"000000000",
    "typeQuestion": "vof",
    "question": "on peut utiliser windows de façon efficace",
    "reponse": "FAUX, il s'agit d'un système d'exploitation inférieur",
    "sourceFile": "monfichier_000000000_questions.json"
}
```