# **SIRENE Invader V2**

## **Contexte**

Vous devez réaliser un indexeur mongodb à haute volumétrie pour transitionner des données au format csv vers une base de données NoSQL. Vous serez en charge de la totalité du processus permettant le transfére des données csv vers la base de données. Cette base de donées devra comporter un Cluster de Shard.

## **Consignes**

### 1/ Téléchargement

Vous devez télécharger le fichier csv “**Sirene : Fichier StockEtablissement**” à cette adresse : <https://www.data.gouv.fr/fr/datasets/base-sirene-des-entreprises-et-de-leurs-etablissements-siren-siret/>

### 2/ Décomposition csv

Vous deverez décomposer le fichier principal en une multitude de fichiers d’un poid raisonable pour permettre d’intégrer la totalité des données dans le processus d’indexation.

### 3/ Conception

* Réaliser le schéma de votre processus d’indexation en y intégrant la topographie du cluster nosql que vous utiliserez.
* Réaliser un diagramme d’activité UML de votre processus d’indexation avec [draw.io](https://app.diagrams.net/).
* Réaliser la documentation de votre processus d’indexation avec [markdown](https://www.markdownguide.org/).

### 4/ Processus d’indexation

* Vous devrez utiliser l'utilitaire de gestion des processus [PM2](https://pm2.keymetrics.io/docs/usage/quick-start/) afin de paralléliser vos opérations et accélérer l'enregistrement des informations. Il vous revient de répartir efficacement les fragments de fichier csv par worker pour une indexation optimale et rapide.
* Utiliser l’[API js PM2](https://pm2.keymetrics.io/docs/usage/pm2-api/) avec la gestion des événements dans vos workers pour faire remonter les informations vers votre controller principal qui sera responsable du démarrage de votre programme.
* À vous d'exploiter MongoDB et d'utiliser les méthodes d'indexation les plus efficaces.
* Votre processus doit offrir un mode "PAUSE" pour permettre l'arrêt et la reprise du processus d'indexation en cours de traitement.
* Assurez-vous que vos works soient gérer efficacement en utilisant un fichier de configuration PM2 nommé "process.json", correctement paramétré avec PM2 à la racine devotre projet.
* L'ensemble de votre projet doit être versionné sur Git.
* Veillez à mesurer le temps d'indexation de chaque fichier et établisez un score global à la fin du processus pour évaluer le temps total de l’indexation.
* Monitorer votre processus d’indexation avec [keymetrics](https://id.keymetrics.io/api/oauth/register) ou les outils [cli](https://pm2.keymetrics.io/docs/usage/process-management/) de PM2

### 5/ Modèle

Vous devez collecter les informations suivantes pour tous les établissements dans le fichier CSV. Si une valeur est absente, il est inutile de la stocker dans MongoDB :

* + siren
  + nic
  + siret
  + dateCreationEtablissement
  + dateDernierTraitementEtablissement
  + typeVoieEtablissement
  + libelleVoieEtablissement
  + codePostalEtablissement
  + libelleCommuneEtablissement
  + codeCommuneEtablissement
  + dateDebut
  + etatAdministratifEtablissement
* Toutes les données seront enregistré dans une collection MongoDB.
* Toutes les données requises doivent être collectées.
* Vous devez utiliser [mongoose](https://mongoosejs.com/).

### Notation

Vous serez évalué sur l’utilisation des outils cité dans l’évaluation et sur le respect des consignes strict de ce document.

A vous de faire en sorte d’avoir les meilleurs performances possible avec une indexation mongodb & node.js. Il est interdit d’utiliser d’autres packages que ceux mentionnés dans les consignes.