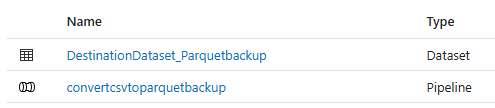
Nous utilisons une Data Factory afin de faire la conversion des fichiers CSV au format Parquet. Data Factory est un service d’Azure permettant de créer et automatiser des flux de données. Dans notre cas, le flux est très simple, étant composé uniquement d’une étape de copie de fichiers.

Pour son fonctionnement, notre Data Factory nécessite deux types de ressources : des datasets et une pipeline.

Les datasets doivent être configurés afin d’être accessibles par les pipelines. Dans notre cas, les datasets pointent sur les deux dossiers du datalake (CSV et Parquet).Une image contenant texte, capture d’écran, nombre, logiciel

Le contenu généré par l’IA peut être incorrect.

Les connections sont paramétrées ainsi :Une image contenant texte, capture d’écran, logiciel, nombre

Le contenu généré par l’IA peut être incorrect.

Une image contenant texte, ligne, Police, nombre

Le contenu généré par l’IA peut être incorrect.

Ces datasets sont déployés avec ces paramètres par le script Terraform.

La pipeline d’une seule étape de copie qui récupère tous les fichiers du dossier CSV et les copie en Parquet dans le dossier correspondant.

Une image contenant texte, logiciel, nombre, Icône d’ordinateur

Le contenu généré par l’IA peut être incorrect.

Cette pipeline est déployée par le script Terraform aussi. Il est assez simple de paramétrer le déploiement Terraform car il est possible de consulter la représentation de la pipeline en JSON depuis le Data Factory Studio.Une image contenant texte, Police, capture d’écran, ligne

Le contenu généré par l’IA peut être incorrect.Le JSON obtenu est à passer en paramètre de la ressource dans le script Terraform.

Ici, la pipeline est déclenchable manuellement en sélectionnant « Add trigger » puis « Trigger now » :

Une image contenant texte, Police, blanc, capture d’écran

Le contenu généré par l’IA peut être incorrect.

A noter que l’exécution d’une pipeline peut coûter des crédits Azure : 