

Mis à jour le 26/10/2026

S'inscrire

Formation Great Expectations

4 jours (28 heures)

Présentation

Great Expectations est un framework de validation de données permettant de fiabiliser les flux métier en entreprise. Avec les Expectations et les Checkpoints vous formalisez vos règles qualité, automatisez les contrôles et produisez des preuves de validation pour vos pipelines.

Notre formation Great Expectations vous permettra de développer des compétences avancées en qualité des données dans des contextes professionnels sécurisés. Vous apprendrez à transformer les exigences métier en contrôles exploitables par vos équipes techniques.

Vous structurerez un projet GX Core et connecterez vos sources pour construire des suites de validation réutilisables. La progression couvre les règles conditionnelles, les résultats, les Data Docs et les décisions de blocage dans un pipeline.

À l'issue de la formation, vous serez en mesure de renforcer le hardening de votre environnement, de protéger les secrets et de limiter les données sensibles exposées. Les intégrations SIEM et EDR reposent sur des exports ou des actions personnalisées, sans remplacer les outils de sécurité.

Cette formation aborde également les tests, le déploiement reproductible et la traçabilité des validations. Les ateliers vous permettront de concevoir des pipelines certifiants au sens de preuves de contrôle, sans certification réglementaire, puis de préparer leur appropriation par les équipes.

Comme toutes nos formations, celle-ci vous présentera **la dernière version stable** de la technologie et ses nouveautés.

Objectifs

- Structurer un projet Great Expectations pour valider des données en environnement professionnel sécurisé
- Construire des Expectation Suites adaptées aux risques métier et aux exigences de qualité
- Automatiser des pipelines certifiants produisant des preuves de validation traçables
- Intégrer des données issues d'un EDR et transmettre des événements qualité à un SIEM par des interfaces personnalisées
- Appliquer une stratégie de hardening aux secrets, aux accès et aux résultats de validation
- Déployer un dispositif de validation documenté et préparer le transfert de compétences aux équipes techniques

Public visé

- Data Engineer
- Développeur Python
- Ingénieur qualité des données
- Ingénieur DevSecOps
- Analyste SOC

Pré-requis

- Savoir écrire des scripts en Python et manipuler des tableaux avec pandas dans un environnement de développement
- Savoir interroger une base relationnelle avec SQL pour explorer des données
- Comprendre les étapes d'un pipeline de données et les principales dimensions de qualité
- Posséder des notions de journalisation, de gestion des secrets et de contrôle des accès

Pré-requis techniques

- Disposer d'un poste avec 4 cœurs, 16 Go de mémoire et 20 Go libres, sans carte graphique dédiée nécessaire
- Utiliser Ubuntu 24.04 LTS directement ou dans une machine virtuelle disposant des ressources nécessaires
- Installer Python 3.12 et GX Core 1.x dans un environnement virtuel, avec les dépendances et versions figées fournies avant la session
- Disposer de Git et de Visual Studio Code ainsi que d'un accès à la base PostgreSQL du laboratoire
- Prévoir une connexion Internet stable avec les accès pédagogiques autorisés, les exports EDR anonymisés et le collecteur SIEM de test étant fournis sans compte cloud personnel requis

Programme de notre formation Great Expectations

[Jour 1 - Matin]

Fondamentaux et environnement de validation

- Dimensions de qualité des données et conséquences des anomalies sur les risques métier
- Périmètre de GX Core et distinction entre validation de données, cybersécurité et certification
- Architecture des objets : Data Context, Data Source, Data Asset, Batch Definition et Batch
- Installation isolée, dépendances figées et organisation d'un projet Python reproductible
- Choix entre Data Context éphémère et persistant selon les usages
- Atelier pratique : Initialiser un projet GX Core et exécuter un premier contrôle sur un jeu de données métier anonymisé.

[Jour 1 - Après-midi]

Sources et premières règles qualité

- Connexion aux données avec pandas et PostgreSQL selon le contexte d'exécution
- Définition des Data Assets et découpage des données en lots reproductibles
- Construction d'une Expectation Suite à partir d'exigences métier explicites
- Contrôles de complétude, d'unicité, de valeurs autorisées et de bornes
- Création d'une Validation Definition et lecture des résultats de validation
- Atelier pratique : Connecter deux sources, construire une suite de contrôles et analyser les anomalies injectées.

[Jour 2 - Matin]

Règles métier et validations approfondies

- Traduction des risques métier en exigences mesurables et seuils d'acceptation
- Contrôles de structure, de volumétrie et de cohérence entre colonnes
- Paramétrage des tolérances avec mostly et gestion explicite des valeurs nulles
- Utilisation de conditions de lignes et de règles SQL adaptées au moteur choisi
- Tests des suites sur des cas limites et distinction entre anomalie de données et erreur d'exécution
- Atelier pratique : Construire et tester une suite de validation métier avec tolérances, contrôles conditionnels et cas limites.

[Jour 2 - Après-midi]

Checkpoints et pipelines certifiants

- Assemblage des Validation Definitions dans des Checkpoints réutilisables
- Automatisation des Actions et publication des Data Docs pour les équipes
- Choix du result_format selon les besoins de diagnostic et de confidentialité
- Gestion des seuils de blocage, des rejets et de la reprise dans le code du pipeline
- Conception de pipelines certifiants par conservation des preuves, sans valeur de certification réglementaire
- Atelier pratique : Réaliser un pipeline bloquant les lots non conformes et archivant les résultats avec les versions des règles et les identifiants de lots.

[Jour 3 - Matin]

Hardening et protection des résultats

- Analyse des surfaces d'exposition du projet, des connexions et des artefacts de validation
- Externalisation des secrets et exclusion des informations sensibles du dépôt de code
- Application du moindre privilège aux comptes de lecture et aux répertoires de stockage
- Réduction des données sensibles dans les résultats, les journaux et les Data Docs
- Stratégie de hardening de l'environnement hôte, des dépendances et des communications, hors fonctions natives de GX
- Atelier pratique : Durcir un projet de laboratoire, supprimer les secrets exposés et vérifier les informations présentes dans les artefacts produits.

[Jour 3 - Après-midi]

Intégration aux flux SIEM et EDR

- Positionnement de Great Expectations comme contrôle qualité des flux de sécurité, sans fonction de détection EDR native
- Préparation d'exports EDR anonymisés avec identifiants, horodatages et champs normalisés
- Validation de la complétude, de l'unicité et de la cohérence temporelle des événements
- Conception d'une Action personnalisée pour transmettre des résultats synthétiques à un collecteur SIEM
- Gestion des échecs de transmission, des doublons et des identifiants de corrélation dans l'intégration
- Atelier pratique : Valider un export EDR et envoyer un événement qualité sans données sensibles vers un collecteur SIEM de test.

[Jour 4 - Matin]

Industrialisation et exploitation fiable

- Organisation des configurations entre développement, test et production
- Intégration des validations dans une chaîne CI/CD avec tests de non-régression
- Optimisation du périmètre des lots, des requêtes et du volume de résultats retournés
- Suivi des durées d'exécution, des taux de conformité et des incidents de validation
- Politique de rétention, sauvegarde des configurations et reprise après incident
- Atelier pratique : Automatiser l'exécution du pipeline en intégration continue et comparer sa fiabilité et ses performances avant et après optimisation.

[Jour 4 - Après-midi]

Projet de synthèse et transfert aux équipes

- Conception d'une architecture cible associant validation métier, hardening et intégration sécurité
- Définition des critères d'acceptation et des preuves attendues pour chaque lot

- Répartition des responsabilités entre équipes data, exploitation et sécurité
- Création d'un kit de formation interne avec jeux de données, exercices progressifs et corrigés
- Préparation d'un guide d'exploitation et d'une grille d'évaluation des compétences techniques
- Atelier pratique : Livrer et présenter un pipeline certifiant durci sur données métier et exports EDR, avec remontée SIEM, preuves de contrôle et support de prise en main.

Sociétés concernées

Cette formation s'adresse à la fois aux particuliers ainsi qu'aux entreprises, petites ou grandes, souhaitant former ses équipes à une nouvelle technologie informatique avancée ou bien à acquérir des connaissances métiers spécifiques ou des méthodes modernes.

Positionnement à l'entrée en formation

Le positionnement à l'entrée en formation respecte les critères qualité Qualiopi. Dès son inscription définitive, l'apprenant reçoit un questionnaire d'auto-évaluation nous permettant d'apprécier son niveau estimé sur différents types de technologies, ses attentes et objectifs personnels quant à la formation à venir, dans les limites imposées par le format sélectionné. Ce questionnaire nous permet également d'anticiper certaines difficultés de connexion ou de sécurité interne en entreprise (intraentreprise ou classe virtuelle) qui pourraient être problématiques pour le suivi et le bon déroulement de la session de formation.

Méthodes pédagogiques

Stage Pratique : 60% Pratique, 40% Théorie. Support de la formation distribué au format numérique à tous les participants.

Organisation

Le cours alterne les apports théoriques du formateur soutenus par des exemples et des séances de réflexions, et de travail en groupe.

Validation

À la fin de la session, un questionnaire à choix multiples permet de vérifier l'acquisition correcte des compétences.

Sanction

Une attestation sera remise à chaque stagiaire qui aura suivi la totalité de la formation.