

Kaischool - programme de formation

Expert en Infrastructure de données massives

v.2.2

La formation vise la certification Simplon, de niveau 7, code NSF 326 enregistrée au RNCP sous le numéro 37638 par décision de France Compétences en date du 31-05-2023. Cette certification est également accessible via la VAE.

Objectifs de la formation

La formation prépare au métier de data engineer spécialisé en infrastructures de données massives. Le professionnel formé aura la capacité de concevoir, automatiser et maintenir les pipelines de données, de mettre en œuvre des architectures distribuées pour le traitement, le stockage et l'analyse de données à grande échelle, et d'assurer leur qualité, sécurité et performance en environnement industriel.

Les objectifs en compétences sont les suivants :

- Piloter un projet data complexe, de l'expression du besoin à la restitution, en intégrant les contraintes techniques, réglementaires et organisationnelles.
- Concevoir, déployer et sécuriser des infrastructures techniques adaptées à la collecte, au traitement et au stockage de données massives, en garantissant performance, résilience et scalabilité.
- Automatiser la collecte et l'intégration de données multi-sources par des chaînes de traitement robustes (API, flux, scripts, ETL).
- Structurer, modéliser et exploiter des bases de données relationnelles et non relationnelles, en assurant leur qualité, leur traçabilité et leur gouvernance.
- Développer et maintenir un entrepôt de données alimenté par des flux contrôlés, modélisé selon les usages métiers et exposé via des outils de visualisation adaptés.
- Intégrer et orchestrer des composants Big Data (Hadoop, Spark, Kafka, etc.) dans des architectures évolutives, en s'appuyant sur des standards de supervision, de journalisation et de sécurité.
- Mobiliser des techniques d'analyse statistique ou d'apprentissage automatique pour améliorer les processus, automatiser certains traitements ou générer de la valeur à partir des données.
- Collaborer dans des environnements agiles et pluridisciplinaires, en assurant la documentation, le contrôle de version, le test et l'industrialisation des solutions déployées.

Prérequis

Les candidats devront démontrer, au moment de la sélection :

- Une formation initiale de niveau Bac+3 minimum dans un domaine en lien avec l'informatique, les statistiques, les mathématiques appliquées, les télécommunications, les systèmes d'information ou les sciences des données.
- Une maîtrise opérationnelle des fondamentaux de la programmation (notamment en Python, SQL ou Java) et des structures de données simples.
- Une capacité à lire et exploiter une documentation technique, y compris en anglais, dans un contexte projet.
- Une appétence pour le domaine, justifiée par des attestations de formation, certifications, projets réalisés, emploi occupé, passion avérée.

Programme de formation

Nom du Module (en plus du nom du module, merci de préciser les compétences contenues dans le module)	La/ les Compétence(s) du référentiel visée(s)	Contenu (sujets traités, thématiques abordées, savoir faire mobilisés, objectifs pédagogiques du module)	Méthodes pédagogiques, types de ressource et support utilisés (ex : étude de cas, mise en situation, exercices, articles, vidéos tutoriels, stack technique etc..)	Durée (h)
Gestion de projet Data (Scrum, Jira, Trello, cartographie des données, veille technologique)	C1, C2, C4, C5, C6, C7	Analyse de besoins, faisabilité, cartographie des données, veille technologique, gestion de projet et communication projet	Cours théoriques, études de cas réels, ateliers collaboratifs, outils de gestion de projet (Jira, Trello), ressources documentaires	50
Extraction et programmation des données (Python, Web Scraping, APIs, automatisation de pipelines)	C8, C10	Extraction automatisée (API, Web scraping), scripts Python avancés, automatisation des pipelines de données	Exercices pratiques, projets d'application, scripts Python, tutoriels, stack technique (Python, BeautifulSoup, Requests, Pandas)	80
Stockage et bases de données	C9, C11	SQL avancé, modélisation de	TD pratiques, mise en situation réelle,	90

avancées (SQL avancé, PostgreSQL, MySQL, NoSQL, MongoDB, Cassandra)		bases relationnelles, introduction aux bases NoSQL (MongoDB, Cassandra), création et optimisation des bases de données	tutoriels techniques, stack (PostgreSQL, MySQL, MongoDB, Cassandra)	
Data Cleaning et intégration (Pandas, NumPy, techniques avancées de nettoyage)	C10	Traitement et agrégation des données, nettoyage avancé, gestion des données manquantes, homogénéisation des formats	Exercices, travaux pratiques, cas pratiques réels, utilisation de bibliothèques Python (Pandas, NumPy)	70
Data Warehouse & ETL (Modélisation dimensionnelle, Talend, Airflow, ETL, SQL)	C13, C14, C15, C16, C17	Modélisation dimensionnelle, création d'entrepôts de données, conception et gestion des ETL, gestion des évolutions historiques des données, en intégrant les variations dimensionnelles de type 1, 2 et 3	Projets d'application, travaux dirigés, tutoriels techniques, outils ETL (Talend, Airflow, SQL), Ateliers pratiques avancés sur la mise en œuvre de la traçabilité et des règles de gestion.	120
Architecture Big Data & Data Lake (Hadoop, Spark, Kafka, Data Lake, gouvernance des données)	C18, C19, C20, C21	Architecture Data Lake, intégration technique (Hadoop, Spark, Kafka), gestion et gouvernance des données, mise en place de catalogues de données et de la gestion des accès	Ateliers pratiques, cas réels d'intégration, tutoriels, stack Big Data (Hadoop, Spark, Kafka), ateliers sur la sécurisation des infrastructures (authentification/autorisation) et le référencement des sources (métadonnées, lignée)	150

Data Analyse & Business Intelligence (Contextualisation métier : visualisation et reporting)	C12	Visualisation et reporting automatisé, outils BI (Tableau, Power BI)	Études de cas, travaux pratiques sur outils BI, tutoriels.	40
Machine Learning & Intelligence Artificielle (Contextualisation métier : Déploiement de modèles prédictifs en production)	C12	Intégration d'outils modernes de modélisation basés sur le Machine Learning	Projets pratiques, TP sur données réelles (Python, Docker, Kubernetes), tutoriels	50
Gouvernance, sécurité et RGPD (RGPD, cybersécurité, gouvernance des données, éthique)	C3, C11, C16, C20, C21	RGPD, gouvernance des données massives, sécurité des infrastructures, cadre juridique et éthique, gestion du catalogue des données (Data Catalog), mise en œuvre d'outils de qualité et de traçabilité (Data Lineage).	Séminaires, études de cas juridiques, ressources documentaires, ateliers RGPD, travaux pratiques sur des outils de Data Quality et de Data Lineage.	80
Mise en situation professionnelle – Bloc A1 (Gestion de projet appliqué)	A1 : C1 à C7	Application concrète du pilotage d'un projet data, de l'étude de faisabilité à la communication des résultats, dans un cas simulé ou réel.	Mise en situation reconstituée, projet tutoré, soutenance intermédiaire	25
Mise en situation professionnelle – Bloc A2 (Collecte, stockage, mise à disposition des données)	A2 : C8 à C12	Conception et automatisation d'un pipeline de collecte, traitement et exposition de données multi-sources.	Projet technique sur cas réel/simulé, présentation technique, évaluation du livrable	25
Mise en situation professionnelle – Bloc A3 (Entrepôt de données et analytique)	A3 : C13 à C17	Construction d'un entrepôt de données, alimentation via ETL et exploitation analytique pour reporting.	Projet de modélisation, démonstration d'analyse, rapport et soutenance	25

Mise en situation professionnelle – Bloc A4 (Architecture Big Data & gouvernance)	A4 : C18 à C21	Déploiement d'un mini data lake, ingestion de données, catalogue, gestion des accès et gouvernance.	Mise en œuvre technique, rapport RGPD/gouvernance , soutenance	25
Préparation à la certification RNCP	Toutes	Formalisation du dossier de preuves, préparation à la soutenance orale, méthodologie de présentation des projets, consolidation des connaissances sur le référentiel métier.	Ateliers de formalisation, simulations de soutenance, coaching personnalisé, revue de projets , grilles d'évaluation et critères du référentiel.	40
Fondamentaux de la Data	C8, C9	Logique algorithmique et structures de données simples. Complexité algorithmique. Environnement système (Linux, commandes de base, scripts shell). Administration système de base. Fondamentaux des réseaux et de l'architecture des ordinateurs.	Cours théoriques, exercices dirigés sur la résolution de problèmes algorithmiques, travaux pratiques sur la ligne de commande Linux, tutoriels de scripting shell.	100
TOTAL				970