

IBM Certified Data Engineer - Big Data

RS3127 · RS · Publiée

L'essentiel

Code(s) NSF	326
Formacode(s)	31054
Type d'enregistrement	Enregistrement sur demande
Date d'échéance	31/12/2021

Certificateur(s)

CIE IBM FRANCE · SIRET : 55211846503644

Partenaires habilités

5 organismes habilités, d'après la liste publiée par France compétences : 4 à former et organiser l'évaluation et 1 à organiser l'évaluation.

Organisme	SIRET	Habilitation
ARROW ECS	38416992600027	Former et organiser l'évaluation
GLOBAL KNOWLEDGE NETWORK FRANCE	40394420000142	Former et organiser l'évaluation
LEARNQUEST FRANCE SAS	80073704100015	Former et organiser l'évaluation
PEARSON FRANCE	68201927800055	Organiser l'évaluation
TECH DATA FRANCE	72206563800065	Former et organiser l'évaluation

Résumé de la certification

OBJECTIFS ET CONTEXTE

Tout type d'entreprise ou association ayant plusieurs postes informatisés, et entreprises du secteur informatique à partir d'une personne

Nous évoluons actuellement dans un monde connaissant une croissance exponentielle des données et, parallèlement, des flux de données.

Ces données proviennent de partout : de messages sur les sites de médias sociaux, d'images numériques et de vidéos publiées en ligne, d'enregistrements transactionnels d'achats en ligne, de signaux GPS de téléphones mobiles, de capteurs utilisés pour collecter les informations climatiques, pour ne citer que quelques sources.

Le concept de Big Data représente un ensemble de techniques concernant des données qui correspondent à plusieurs des caractéristiques suivantes : volume massif, grande variété ainsi que grande vélocité.

Les données, structurées (ex. : fichiers d'un progiciel), ou non structurées (ex. : galeries de photos), jouent un rôle prédominant auprès des entreprises qui les capturent, les gèrent et les analysent. Ces connaissances leur permettent de prendre rapidement des décisions métier bien informées.

La certification IBM Certified Data Engineer - Big Data s'adresse aux ingénieurs Big Data. L'Ingénieur Big Data travaille en étroite collaboration avec l'Architecte Data et les développeurs qui ont une expérience pratique. Grâce à lui, les données définies pour le système-cible et modélisées par l'Architecte Big Data se concrétisent. L'Ingénieur Data maîtrise parfaitement au niveau technique, sur le plan théorique et pratique, une large gamme de produits et de technologies.

Il résout, en outre, les problèmes Big Data en appliquant les technologies adéquates et détient l'aptitude requise pour construire des systèmes de traitement des données à grande échelle pour les entreprises.

Les ingénieurs Data développent, testent et qualifient des solutions Big Data au sein des organisations, tout en

s'assurant que l'infrastructure reste fluide et opérationnelle.
Ils fournissent des indications aux architectes sur les besoins en matériels et logiciels.

COMPÉTENCES ATTESTÉES

A l'issue de sa certification IBM Certified Data Engineer - Big Data, le spécialiste sera en mesure de démontrer les compétences suivantes :

- Réunir, analyser, administrer de vastes ensembles de données, afin de fournir aux Data Scientists les données adéquates et les outils visuels qui leur serviront pour les analyses.
- Saisir la complexité des données et être en mesure de traiter différentes sortes de données (structurées, semi-structurées, non structurées), ainsi que différents volumes, vitesses (y compris traitement des flux) et véracités de données.
- Répondre aux problèmes posés par la gouvernance de l'information et aux défis associés aux données dans le domaine de la sécurité.
- Posséder de solides connaissances en génie logiciel et une grande expérience de la programmation et du cryptage des données.

La journée type de l'Ingénieur Big Data est constituée de 50% de développement, 20% d'étude et spécification, 20% d'accompagnement et résolution de bugs et 10% de veille.

Les compétences particulières suivantes, spécifiques au développement (50% de son activité), seront évaluées :

Architecture et Intégration

- Implémenter MapReduce
- Opter pour Hive, Big SQL ou HBase en fonction des scénarii
- Créer un index Solr ou en faire la requête
- Sélectionner les formats de fichiers appropriés (JSON, CSV, Parquet, Sequence)
- Visualiser les recherches au travers d'Apache View

Chargement des données

- Charger des données non structurées dans InfoSphere BigInsights, importer des données de streaming dans Hadoop avec InfoSphere Streams
- Créer un classeur BigSheets
- Importer des données dans Hadoop et créer des définitions de tableaux Big SQL
- Importer des données dans Hbase et dans Hive
- Charger des données depuis des sources relationnelles dans InfoSphere BigInsights avec Data Click et en mode libre-service
- Extraire des données d'une source relationnelle en utilisant Sqoop
- Charger des données de fichiers log dans Hadoop avec Flume
- Insérer des données via l'API du système de fichiers Posix IBM General Parallel File System (GPFS)
- Charger des données avec l'utilitaire de ligne de commande de Hadoop

Préparation, Transformation et Exportation des données

- Transformer les données dans InfoSphere BigInsights en utilisant les techniques de requêtes Jaql
- Capturer et préparer les données issues des médias sociaux en vue de leur analyse
- Intégrer le scoring de modèle d'IBM SPSS dans InfoSphere Streams
- Mettre en oeuvre une résolution d'entité dans une plate-forme Big Data, par ex. Big Match
- Transformer et manipuler des données à partir de PIG
- Transformer des données dans InfoSphere BigInsights à partir de Big SQL
- Exporter de Hadoop les résultats du traitement des données (DataClick, DataStage)
- Utiliser des zones homogènes dans InfoSphere Streams afin de permettre au moins un traitement

Sécurité des données

- Sécuriser les données dans les standards PCI
- Protéger les données sensibles avec des masquages (Optim, Big SQL)

Performance et Evolutivité

- Améliorer les performances de MapReduce à l'aide de RDD (Resilient Distributed DataSet)
- Sélectionner les formats de fichiers les plus adaptés pour optimiser les performances de Big SQL et de JAQL
- Prendre des décisions spécifiques visant à optimiser les performances de Hive et de HBase
- Analyser les facteurs de performance lors de l'utilisation d'Apache Spark

Voies d'accès

PRÉREQUIS À L'ENTRÉE EN FORMATION

Préalablement à sa préparation à la certification, le professionnel devra détenir les connaissances et aptitudes suivantes :

- Couche de données et ses zones d'enjeux/de risques potentiels
- Capacité à transformer les exigences fonctionnelles en spécifications techniques
- Capacité à aboutir à une architecture physique en partant d'une solution globale ou d'une architecture logique
- Gestion des clusters
- Exigences réseau
- Principales interfaces
- Modélisation des données
- Capacité à identifier pour la solution les exigences non fonctionnelles et à apporter un soutien technique
- Latence, flexibilité, haute disponibilité, réplication et synchronisation des données, restauration après sinistre
- Performance globale (performance des requêtes, gestion de la charge de travail, optimisation de la base de données)
- Proposer les meilleures pratiques ou les plus recommandées en matière de déplacement, manipulation et stockage des données dans une solution Big Data
- Options techniques d'ingestion des données
- Options et ramifications du stockage de données, dont exigences supplémentaires et enjeux imposés par les données dans le Cloud
- Techniques de requêtes de données et leur disponibilité en matière de soutien à l'analyse.
- Généalogie des données et gouvernance des données
- Diversité des données et volume des données
- Mettre en place la sécurité des données et fournir des orientations dans ce domaine, conseiller en particulier sur la sécurité LDAP, les fonctions/la sécurité des utilisateurs, le monitoring des données, la sécurité des données à caractère personnel
- Logiciels à connaître en priorité : BigInsights - BigSQL - Hadoop - Cloudant (NoSQL)
- Logiciels à connaître accessoirement :

Information Server - Intégration à partir de BigInsights - Balanced Optimization pour Hadoop - Fonction Push Down de JAQL - Gouvernance des données - Spécificités de BigInsights liées à la sécurité - Metadata Workbench - Intégrer Optim avec BigInsights - DataClick pour BigInsights - BigMatch - Guardium (monitoring) - Outils d'analyse des données (SPSS) - BigSheets - Support technique de Hadoop/BigInsights - Disponibilité des données et support dans leur requête - Flux de données - Interface/intégration avec BigInsights - Données de streaming - Analyse de données intégrée - Netezza - DB2 BLU - Bases de données Graph - Machine Learning (système ML)

Base légale

Date d'échéance de l'enregistrement 31/12/2021

Pour plus d'informations

LIEN VERS LE DESCRIPTIF

Lien de la certification :

Fiche officielle (France Compétences)

<https://www.francecompetences.fr/recherche/rs/3127/>