Intégration Big Data

Filière Génie Logiciel ISIKA – 2ème Cycle

Présentation du cursus

Le parcours « **Intégration Big Data** » est un des parcours de spécialisation de la filière « **Génie Logiciel et métiers du Numérique** » de ISIKA. Il peut s'inscrire dans le prolongement du tronc commun « Concepteur Développeur Informatique » de ISIKA.

Il peut aussi s'inscrire en parcours d'actualisation de compétences ou d'élargissement de savoir et savoir-faire et de spécialisation destiné :

- soit à des informaticiens confirmés issus du domaine des Etudes et Développement et/ou du domaine de l'informatique de production de données justifiant de compétences réelles en développement en environnement Java EE et Web,
- soit à des auditeurs issus, a minima, de cursus de niveau II ou de niveau III qui souhaiteraient développer une connaissance approfondie des logiques de qualification et gestion, de capture des données (structurées et non structurées), de structuration et de nettoyage des données en vue de leur exploitation en environnement de Big Data.

Ce parcours de haut niveau technique offre aussi l'opportunité de développer une bonne connaissance des logiques d'intégration technique poussée mettant en oeuvre des environnements hétérogènes.

Capacités visées

Les auditeurs de ce parcours ont vocation à occuper des postes à intitulé : **Data Engineer, Ingénieur Développement Big Data, Développeur Hadoop.** Ils auront développé au sein de ce parcours les savoir et savoir-faire nécessaires pour

- Participer à la construction et à l'intégration du socle technique Big Data (librairies de fonctions et d'outils mis à la disposition des data scientists…) ;
- Mener des projets Big Data pour le compte de différents départements métier afin de créer de la valeur autour de leurs données, et en leur assurant comment passer de la donnée brute à de la donnée propre, exposée sous forme de tables ;
- Modéliser les schémas de données, nettoyer et normaliser les données, publier les données ;
- Consolider ces données au fur et à mesure de leur alimentation récurrente dans l'espace Big Data.

Axes de Contenu

Module 1

Les fondamentaux du Big data. Mise en place de l'environnement

(7 jrs – 49 heures)

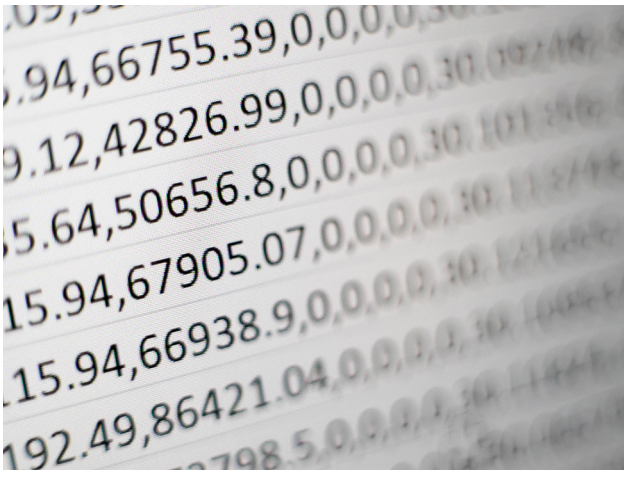
- Etat de l'art, enjeux et perspectives ; Architecture Big Data ; Big Data et Temps réel. ; Les solutions et stratégies Batch et Interactive ; Solutions lambda : Temps réel + Batch.
- Mise en place de la plate-forme de développement et d'intégration : Spring boot et microservices : centralisation et enregistrement des services ; La solution open-source Eureka de Netflix et l'API Spring-Cloud.

Module 2

Le socle de développement Java / Javascript
(10 jrs – 70 heures)

Rappels des Fondamentaux Javascript :

- Types de données, gestion des tableaux, opérateurs et boucles, gestion d'erreur et exceptions, objets et JavaScript, JSON, gestion des événement ;
- JavaScript côté serveur et Node.js ;
- Java, programmation fonctionnelle et expressions lambda.



Module 3

La Distribution Hadoop pour développeur : Cloudera
(8 jrs – 56 heures)

- Mise en oeuvre : HDFS (Stockage de masse), Sqoop (Import/ Export vers SGBD) ; Haddop, le modèle MapReduce et le testing de job Hadoop avec MrUnit ;Le Requêtage sur système de fichier : HIVE ; Langage de script d'interrogation de données : Pig ; l'Orchestration : Oozie ; l'Acquisition de données : Flume ;
- Projet d'intégration : Définition du besoin, cahier des charges, conception, formalisation, développements des bases de données, développement des outils d'alimentation ; mise en oeuvre de traitements batch.

Module 4

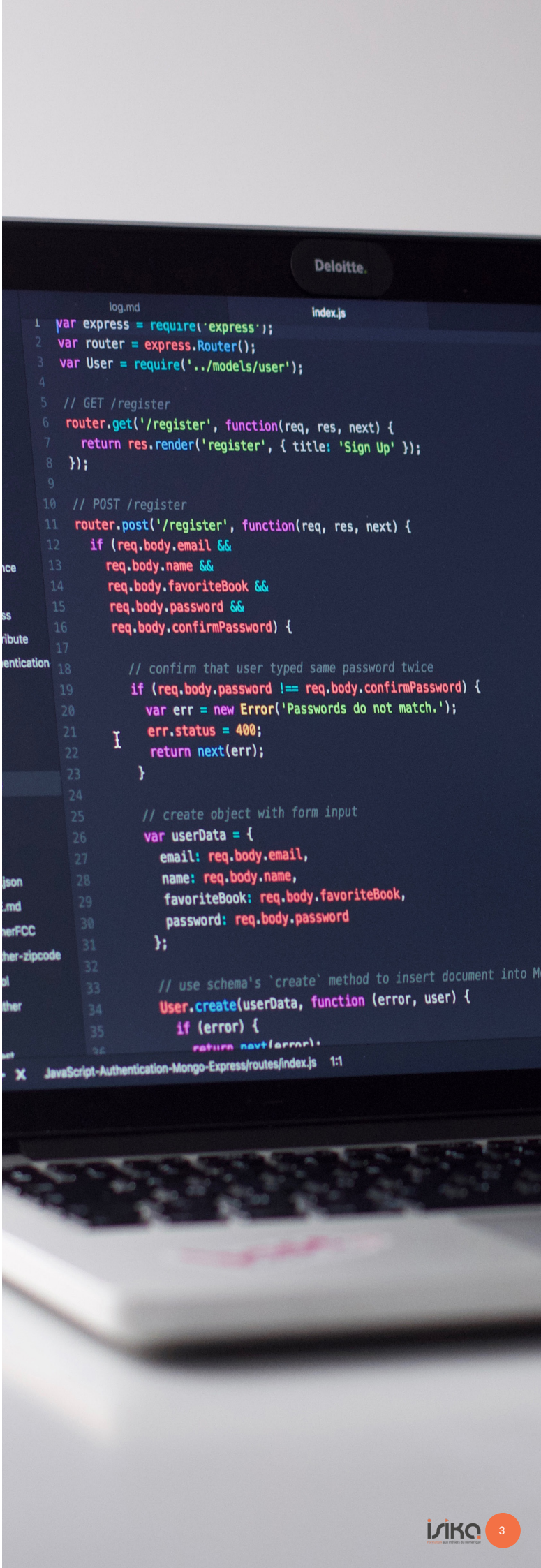
Mise en oeuvre plate-forme Big Data en environnement Scala et Python
(8 jrs – 56 heures)

- Le langage Scala : éléments de base, syntaxe et grammaire du langage ;
- Le langage Python : le langage et sa mise en oeuvre ;
- Le moteur de traitement de données Spark sous Scala et Python : Spark core (Traitement Batch) ;
- Spark Streaming (Traitement temps réel) ;
- Spark SQL (Traitement avec requête SQL like) ;
- Spark ML (Machine Learning) ;
- Gestion de la distribution des messages d'échanges de données : Apache Kafka, mise en oeuvre.

Module 5

Conception et développement NoSQL ; la plate-forme ELK
(10 jrs – 70 heures)

- Les outils et Bases de données NOSQL : MongoDB, Cassandra. Implémentation et mise en oeuvre ;
- La plate-forme ELK : Le moteur de Recherche et d'Analyse ElasticSearch ; l'Acquisition de données : LogStash ; la Visualisation de données : Kibana ;
- Projet d'intégration ELK et traitement temps réel.



Objectifs pédagogiques

A travers une formation privilégiant la mise en oeuvre en mode projet, les auditeurs de ce parcours auront acquis, développé et consolidé une pratique et une maîtrise :

- Du **déploiement d'architectures complexes** mettant en oeuvre des technologies BigData hétérogènes ;
- Des **technologies et du développement d'applications de capture des données** (structurées et non structurées) produites dans les différentes applications, et des démarches d'intégration des éléments ;
- Des **méthodes et outils de structuration** (sémantique, etc.) et de **cartographie de la donnée** ;
- Des **outils et démarches de dédoublonnage**, validation, présentation et mise en forme de la donnée à travers la mise en oeuvre des outils et protocoles de référence : HDFS, Sqoop , Haddop, MapReduce, HIVE, Pig, Oozie, Flume, Scala, Python, Java et Javascript, des bases de données de référence (MongoDB, Cassandra) et du moteur de traitement de données SPARK.

Public et pré-requis

Chercheurs d'emploi de plus de 26 ans ou salariés d'entreprise de niveau II (Bac+3/4). Expérience du projet Web, pratique des environnements distribués. Maîtrise de Java 2 EE.

Durée

En équivalent présentiel, la durée de la formation est de 300 heures, soit 50 jours, soit 10 semaines

Méthodes Pédagogiques

La formation peut être dispensée : en mode full présentiel ou en mode full distanciel avec tutorat asynchrone, tutorat peer-to-peer, en mode blended e-learning où sont accessibles :

- en regroupement, les contenus de cours fortement conceptuels
- en regroupement les ateliers projets, les études de cas avancés, les TPs de consolidation
- en distanciel les autres contenus (vidéos de cours, supports de cours, exercices, corrigés)