Architecture BigData

Tenants & aboutissants

Durée : 3 jours

Réf. : BIGDATA

Méthode pédagogique

- La formation se passe en mode présentiel (face à face), et se compose de 70% de travaux pratiques (Mises en situation, débats, exercices).
- Une évaluation quotidienne de l'acquisition des connaissances de la veille est effectuée.
- Une synthèse est proposée en fin de formation.
- Un support de cours sera remis à chaque participant comprenant les slides de la théorie, les exercices et travaux pratiques ainsi que leurs corrigés
- Une évaluation à chaud sera proposée au stagiaire à la fin du cours.
- Une feuille d'émargement par demi-journée de présence est fournie en fin de formation ainsi qu'une attestation de fin de formation si le stagiaire a bien assisté à la totalité de la session.

Présentation

Si le BigData est LE sujet du moment, il est souvent considéré comme une boîte noire dans laquelle il est très difficile de se retrouver.
Large regroupement de pratiques, d'objectifs et de technologies différentes, il demeure l'objet de nombreux questionnements :

- A partir de quand parlons-nous de BigData ?

- Quels outils pour gérer de gros volume de données ?
- Traitement batch ou traitement en continue ?
- La datascience implique t'elle nécessairement une approche Bigdata ?
- Quelle compétence pour un datascientist ?

Objectifs

- Définir les concepts et identifier l'apport du BigData
- Déterminer l'écosystème technologique
- Organiser la collecte des données
- Choisir une technologie de stockage de données
- Connaître les technologies pour traiter les gros volumes de données
- Définir et comprendre le rôle du datascientist

Audience

DSI, architecte SI, chef de projet, développeur, dataminer, datascientist

Prérequis

Les compétences professionnelles suivantes sont souhaitables : la connaissance d'un langage de programmation structuré et les bases du monde relationnel.

Le formateur

Le formateur est un expert du domaine qui intervient sur le sujet depuis plusieurs années en formation mais aussi en conseil. Doté d'une grande qualité d'écoute, sa pédagogie et sa

compétence technique vous permettront d'acquérir les compétences sur les architectures BigData.

Programme

Comprendre les concepts et les enjeux du BigData

- Origines et définition du BigData.
- Les 3 V : Volume, Vélocité et Variété
- Diversité dans les cas d'usage : données chaudes, données froides
- BigData : Une approche réservée aux GAFA ?
- Un exemple d'architecture BigData.
- **Exercice / Démo :** Parcourir différentes sources de données accessibles via le WEB (API)

Expliquer les technologies du BigData

- Définir les outils de collecte de données
- Anticiper les moyens de stockage en fonction des usages
- Le datalake : votre référentiel de données
- Paralléliser ou traiter vos données en continue ?
- S'approprier les données avec des analyses visuelles : la dataviz

Stocker des données

- État de l'art : Le BigData, sonne t'il le glas des bases de données relationnelles ?
- Le triangle de CAP

- Pourquoi le NoSql ?
- Les différentes approches : document / wide column / key-value
- Tour d'horizon des solutions à disposition : MongoDB, Cassandra, HBase…
- **Exercice / Démo** : définir et mettre en place un modèle de stockage de type document avec MongoDB

Collecter les données

- Comprendre les différentes sources de données : iOT / SI / Réseau sociaux / API : D'où viennent les données ?
- Gérer des formats de données différents : JSON, XML, CSV, binaires, …
- De l'importance des connecteurs…
- Tour d'horizon des outils du marché : NIFI / Node Red / Flume / Sqoop
- **Exercice / Démo** : Utiliser NIFI pour collecter les données d'une API publique

Hadoop

- Comprendre le périmètre de Hadoop : Stockage et traitement
- Une plateforme de traitement batch et de stockage de données froides
- Architecture et composants de la plateforme Hadoop.
- HDFS, YARN et Mapreduce : les 3 piliers
- Un écosystème complexe et complet : Hive, HBase
- **Exercice / Démo** : Manipuler des fichiers via Hue, mise en place de tables et requêtes Hive sur une plateforme Hadoop

Spark

- Un framework pour paralléliser des traitements
- Positionnement Spark / Hadoop
- Quelle infrastructure de déploiement ?
- Comprendre la complexité de la parallélisation des traitements
- SparkML : une librairie pour la datascience
- **Exercice / Démo** : Mise en place et analyse d'un traitement simple

Stream processing

- Le besoin de traitement au fil de l'eau
- Streaming ETL
- Streaming analytics
- Prise de décision en temps réel
- Les approches et outils de streaming : Spark Streaming / Kafka Streaming / Flink …
- **Exercice / Démo** : analyse en continue d'un flux de données simple

Transporter vos données : Kafka

- Définir le besoin d'un bus de données
- Les middleware Orienté Messages dans un contexte BigData
- Kafka : Le bus de données performant
- Définir les acteurs : Producers & Consumers
- Comprendre les composants : Messages, Brokers, topics, …
- Un outil taillé pour les performances
- Kafka Connect : Connectez vos outils à Kafka
- **Exercice / Démo** : Mise en place d'un bus Kafka pour permettre à Elasticsearch de manipuler des données extraites via NIFI

BigData et Machine Learning

- Présentation du Machine Learning

- Positionnement de la datascience dans un contexte Bigdata
- Les différentes approches : Clusterisation, classification, régression
- Les implémentations : Scikit Learning / SparkML
- Spark et DASK : des frameworks de distribution des traitements
- Le « Deep learning »
- Le « Online learning » ou machine learning en streaming
- **Démo :** processus complet d'un projet de datascience (analyse des données, sélection de données, apprentissage, scoring)

Datavisualization

- Pourquoi faire ?
- Dataviz pour comprendre les données
- L'écosystème de la Dataviz : outils et API
- **Exercice / Démo :** Analyse visuelle d'un jeu de données