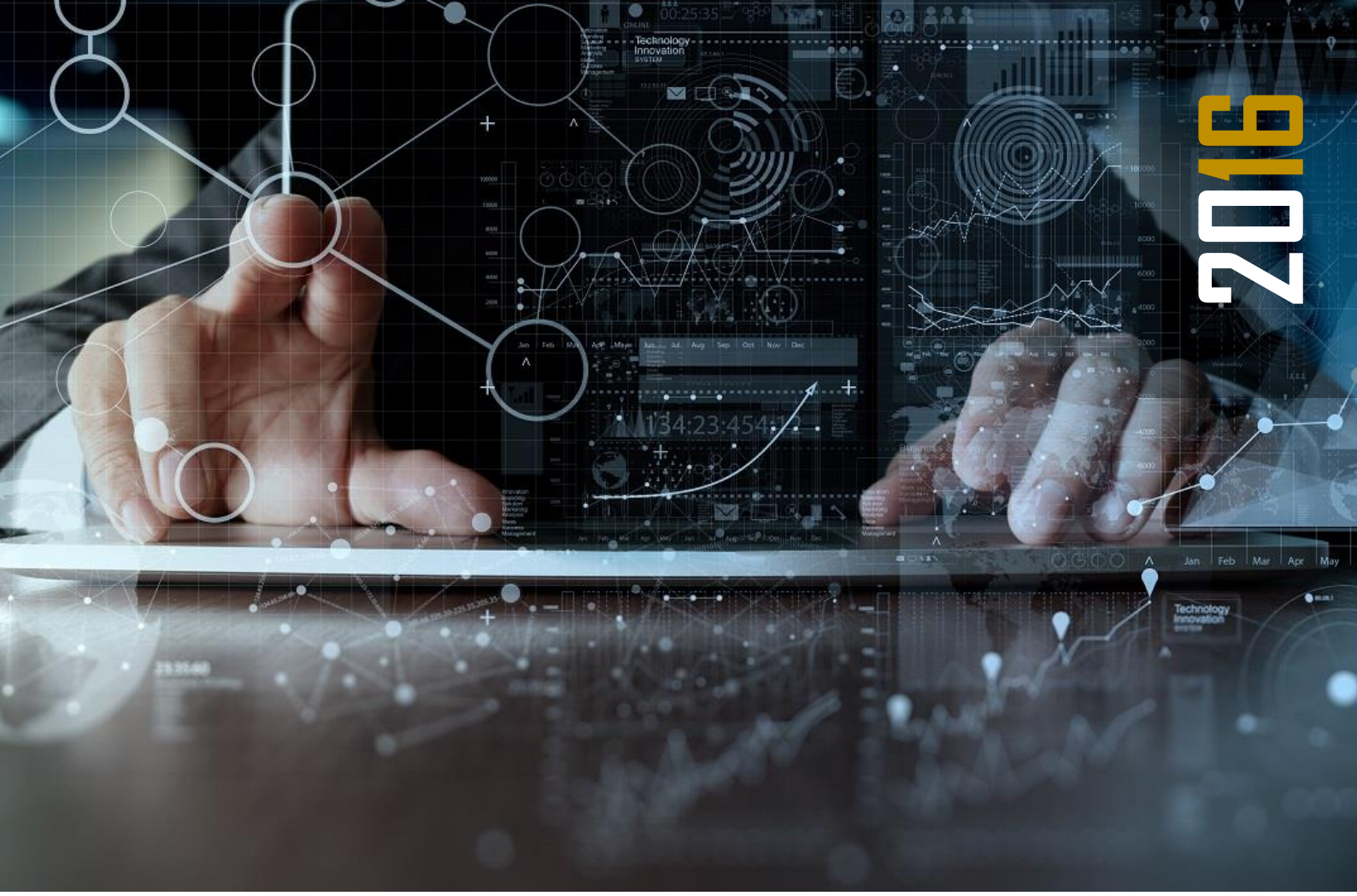


CATALOGUE DE FORMATIONS

BIG DATA

EMGS GROUP

CONSULTING | RECRUITMENT | TRAINING

Sommaire

Formation Data Science

Formations Hadoop

Formations Nosql

FORMATION

Data science

Contact

☎ 01 75 43 05 58
✉ training@emgsgroup.com

Fondamentaux de la Data Science

Code formation : EMFDS
Durée : 3 jours – 21h de cours
Format : Inter-entreprise*

3 jours
21 heures de cours

2300€
Tarif Inter-entreprise*/Hors taxes

Cette formation est également disponible en « Intra-entreprise », nous contacter pour plus d'infos.

| Description

Surfant sur la vague du Big Data, le data scientist joue un rôle clé dans la valorisation de données. Au-delà des paillettes, quel est son rôle, ses outils, sa méthodologie, ses « tips and tricks » ? Venez le découvrir au travers de cette initiation à la Data Science délivrée par des data scientists renommés qui vous apporteront l'expérience des compétitions de Data Science et leurs riches retours d'expérience des modèles réels qu'ils mettent en place chez leurs clients.

| Objectifs pédagogiques

- Découvrir le monde de la Data Science et les grandes familles de problèmes
- Savoir modéliser un problème de Data Science
- Créer ses premières variables
- Constituer sa boite à outils de data scientist

| Publics

Analyste, statisticien, architecte, développeur

| Pré-requis

- Connaissances de base en programmation ou scripting.
- Quelques souvenirs de statistiques sont un plus.

| Méthode pédagogique

Formation avec apports théoriques, échanges sur les contextes des participants et retours d'expérience pratique du formateur, complétés de travaux pratiques et de mises en situation.

| Programme détaillé

Jour 1

- **Introduction au Big Data**
  - Qu'est-ce-que le Big Data ?
  - L'écosystème technologique du Big Data

- **Introduction à la Data Science**
  - Le vocabulaire d'un problème de Data Science
  - De l'analyse statistique au machine learning
  - Overview des possibilités du machine learning

- **Modélisation d'un problème**
  - Input / ouput d'un problème de machine learning
  - Mise en pratique « OCR» (Nous verrons comment modéliser le problème de la reconnaissance optique de caractère)

- **Identifier les familles d'algorithmes de machine learning**
  - Analyse supervisée
  - Analyse non supervisée
  - Classification / régression

- **Sous le capot des algorithmes : la régression linéaire**
  - Quelques rappels : fonction hypothèse, fonction convexe, optimisation
  - La construction de la fonction de coût
  - Méthode de minimisation : la descente de gradient

- **Sous le capot des algorithmes : la régression logistique**
  - Frontière de décision
  - La construction d'une fonction de coût convexe pour la classification

- **La boite à outil du data scientist**
  - Introduction aux outils
  - Introduction à python, pandas et scikit-learn

- **Cas pratique n°1 : « Prédire les survivants du Titanic »**
  - Exposé du problème
  - Première manipulation en python

| Programme détaillé

Jour 2

- **Rappels et révision du jour 1**

- **Qu'est-ce qu'un bon modèle ?**
  - Cross-validation
  - Les métriques d'évaluation : precision, recall, ROC, MAPE, etc.
  - Overview des possibilités du machine learning

- **Les pièges du machine learning**
  - Overfitting ou sur-apprentissage
  - Biais vs variance (Nous verrons comment modéliser le problème de la reconnaissance optique de caractère)
  - La régularisation : régression Ridge et Lasso

- **Data cleaning**
  - Les types de données : catégorielles, continues, ordonnées, temporelles
  - Détection des outliers statistiques, des valeurs aberrantes
  - Stratégie pour les valeurs manquantes
  - Mise en pratique : « Remplissage des valeurs manquantes»

- **Feature engineering**
  - Stratégies pour les variables non continues
  - Détecter et créer des variables discriminantes

- **Cas pratique n°2 : « prédire les survivants du titanic »**
  - Identification et création des bonnes variables
  - Réalisation d'un premier modèle
  - Soumission sur Kaggle

- **Data visualisation**
  - La visualisation pour comprendre les données : histogramme, scatter plot, etc.
  - La visualisation pour comprendre les algorithmes : train / test loss, feature importance, etc.

- **Introduction aux méthodes ensemblistes**
  - Le modèle de base : l'arbre de décision, ses avantages et ses limites
  - Présentation des différentes stratégies ensemblistes : bagging, boosting, etc.
  - Mise en pratique « OCR» (Utilisation d'une méthode ensembliste sur la base du précédent modèle)

- **Apprentissage semi-supervisé**
  - Les grandes classes d'algorithmes non supervisées : clustering, PCA, etc.
  - Mise en pratique : « Détection d'anomalies dans les prises de paris»

  (Nous verrons comment un algorithme non supervisé permet de détecter des fraudes dans les prises de paris)

| Programme détaillé

Jour 3

- **Rappels et révisions**
  - Synthèse des points abordés en journées 1 et 2
  - Approfondissement des sujets sélectionnés avec l'intervenant

- **Mise en pratique**
  - Le dernier jour est entièrement consacré à des mises en pratique

- **Sélection et participation à une compétition**
  - Le formateur sélectionnera une compétition en cours sur Kaggle ou datascience.net qui sera démarrée en jour 3 par l'ensemble des participants

FORMATIONS

Hadoop

EMGS GROUP
CONSULTING | RECRUITMENT | TRAINING

Contact
☎ 01 75 43 05 58
✉ training@emgsgroup.com

Les fondamentaux d'Hadoop

Code formation : EMFDH
Durée : 2 jours – 14h de cours
Format : Inter-entreprise*

2 jours
14 heures de cours

1580€
Tarif Inter-entreprise*/Hors taxes

Cette formation est également disponible en « Intra-entreprise », nous contacter pour plus d'infos.

| Description

Cette formation est une initiation aux fondamentaux d'Hadoop. Elle donne aux participants une connaissance théorique et pratique de la plateforme, au travers de plusieurs exercices pratiques appliqués à des cas réels. A l'issue de la session, les participants seront en capacité d'utiliser les outils de l'écosystème Hadoop pour explorer des données stockées sur un entrepôt Big Data.

| Objectifs pédagogiques

- Appréhender le fonctionnement d'Hadoop
- Identifier l'écosystème : quels outils pour quels usages ?
- Manipuler les principales commandes shell d'interaction avec Hadoop
- Émettre des requêtes SQL avec Hive et HCatalog
- Créer des traitements de données avec Pig

| Publics

Analyste, data scientist, architecte, développeur

| Pré-requis

Connaissances de base en programmation ou en scripting.

| Méthode pédagogique

Formation avec apports théoriques, échanges sur les contextes des participants et retours d'expérience pratique du formateur, complétés de travaux pratiques et de mises en situation.

| Programme détaillé

Jour 1

- **Introduction au Big Data**
  - Qu'est-ce-que le Big Data ?
  - Les grands enjeux métier
  - Paysage technologique : les architectures Big Data

- **Introduction à Hadoop**
  - Historique succinct
  - Le cœur de la plateforme : HDFS et YARN
  - L'écosystème Hadoop :
    - Frameworks et algorithmes
    - Bases de données
    - Traitements des données
    - Intégration
  - Hadoop et la sécurité des données

- **Manipuler la ligne de commande Hadoop**
  - Présentation des principales commandes
  - Mise en pratique « Manipulation et transfert de fichiers en ligne de commande »

- **Une interface utilisateur pour Hadoop : hue**
  - Présentation de Hue et de ses modules
  - Mise en pratique « Manipulation interactive de données »

- **Interroger Hadoop avec du SQL : Hive**
  - Présentation de Hive
  - Mise en pratique « Manipulation de données avec SQL » :
    - Créer un modèle de données
    - Importer des fichiers sources
    - Requête les données

| Programme détaillé

Jour 2

- Transformer des données : le langage Pig
  - Présentation de Pig
  - Mise en pratique « Transformation de données avec Pig » :
    - Charger des données semi-structurées
    - Croiser avec des données Hive
    - Sauvegarder le résultat dans HDFS

- Écriture de traitements avancés
  - Présentation du Framework Hadoop Streaming
  - Présentation express du langage Python et du squelette de programme pour l'exercice
  - Mise en pratique « Ecriture d'un programme de manipulation complexe »

- Composition et ordonnancement de traitements
  - Présentation d'Oozie
  - Mise en pratique « Création d'un pipeline de traitement de données »

Contact

☎ 01 75 43 05 58
✉ training@emgsgroup.com

Développer des applications pour Hadoop 2.X Hortonworks avec Java

Code formation : EMHHJ
Durée : 4 jours – 28h de cours
Format : Inter-entreprise*

4 jours
28 heures de cours

2550€
Tarif Inter-entreprise*/Hors taxes

Cette formation est également disponible en « Intra-entreprise », nous contacter pour plus d'infos.

| Description

Ce stage de formation présente les grands outils de l'écosystème Hadoop d'un point de vue technique et est orienté développement Java. Les objectifs principaux sont d'avoir une parfaite compréhension et pratique du framework d'exécution de calculs MapReduce ainsi que d'être capable de développer des modules d'extensions de Pig et Hive.

| Objectifs pédagogiques

- Identifier et définir les différents composants de l'écosystème Hadoop
- Appréhender l'architecture de Hadoop 2.X
- Mettre en application les techniques avancées MapReduce
- Analyser un use case métier et valoriser les données correspondantes

| Publics

Architecte, développeur, analyste

| Pré-requis

Bonne connaissance du langage Java.

| Méthode pédagogique

Formation avec d'importants apports théoriques, des retours d'expérience du formateur complétés de travaux pratiques sous forme d'exercices d'application et d'analyse de uses cases métier.

| Programme détaillé

Jour 1

- **Comprendre Hadoop 2.x et HDFS**
  - Hadoop et Hadoop 2.X
  - Le système de gestion de ressources et de cluster YARN »
  - Le système de fichiers distribué HDFS
    - Prise en main de l'environnement de développement Hadoop et accès aux fichiers HDFS

- **Écrire des applications Mapreduce**
  - Illustration avec un exemple simple
  - Grands principes du framework MapReduce
  - MapReduce sur YARN
    - Développement de programmes MapReduce

- **Les agrégations avec Mapreduce**
  - Utilisation des combiners
  - Utilisation de l'in-map agrégation
    - Mise en pratique de l'agrégation à travers deux exemples

Jour 2

- **Partitionnement et tri**
  - Le partitioner de MapReduce
  - Analyse et compréhension du Secondary Sort
    - Implémentation de deux types de Partitioner
    - Implémentation du Secondary Sort à travers un cas pratique

- **Input et output formats**
  - Récapitulatifs des formats d'entrée et de sortie standards MapReduce
  - Analyse du fonctionnement d'un input format
    - Implémentations d'un input format et d'un output format

- **Optimiser les jobs Mapreduce**
  - Optimisation des différentes phases d'un programme MapReduce
  - Utilisation et paramétrage de la compression
  - Utilisation des comparateurs de données non sérialisées
    - Illustration du principe de la compression de données
    - Implémentation d'un RawComparator

| Programme détaillé

Jour 3

- **Fonctionnalités avancées de Mapreduce**
  - Localisation partagée des données
  - Les différents types de jointure
  - Les filtres de Bloom
    - Illustration d'une jointure côté Map
    - Illustration de l'utilisation d'un filtre de Bloom

- **Tester unitairement son code**
  - Présentation de la librairie MRUnit
    - Ecriture de tests unitaires

- **Programmation Hbase**
  - Architecture de HBase
  - Interactions avec HBase
    - Import de données avec HBase
    - Illustration d'un job MapReduce avec HBase

Jour 4

- **Programmation Pig**
  - Types et mots-clés dans Pig
  - Extension de Pig via les classes définies par l'utilisateur (UDF)
    - Implémentation d'une UDF

- **Programmation Hive**
  - Types et mots-clés dans Hive
  - Extension de Hive via les classes définies par l'utilisateur (UDF)
    - Implémentation d'une UDF

- **Créer et utiliser un workflow Oozie**
  - Workflow et coordinateur Oozie
  - Actions possibles avec Oozie

Contact
☎ 01 75 43 05 58
✉ training@emgsgroup.com

Développer des applications pour Hadoop 2.X Hortonworks sous Windows

Code formation : EMHHW
Durée : 4 jours – 28h de cours
Format : Inter-entreprise*

4 jours
28 heures de cours

2550€
Tarif Inter-entreprise*/Hors taxes

Cette formation est également disponible en « Intra-entreprise », nous contacter pour plus d'infos.

| Description

Au travers de ce stage, les formateurs OCTO vous formeront à l'analyse de données en contexte Big Data sous Windows avec la plateforme HDP d'Hortonworks. Chaque participant rentrera dans les détails d'Hadoop 2.X, de YARN et d'HDFS. Il acquerra une vue d'ensemble de MapReduce et une connaissance approfondie de Pig et de Hive. Il apprendra également à utiliser Sqoop pour transférer des données entre Hadoop et SQL Server et à connecter Microsoft Excel à Hadoop en utilisant de driver ODBC de Hive.

| Objectifs pédagogiques

- Identifier les principaux composants de l'écosystème Hadoop
- Expérimenter les outils d'exploration et d'analyse de données
- Connecter Hadoop à Microsoft SQL Server et Excel

| Publics

Analyste, statisticien, développeur

| Pré-requis

- Connaissances de base en programmation.
- Une connaissance du SQL et une familiarité avec Microsoft Windows sont un plus.
- Pas de connaissance préalable d'Hadoop requise.

| Méthode pédagogique

Formation avec apports théoriques, échanges sur les contextes des participants et retours d'expérience pratique du formateur, complétés de travaux pratiques et de mises en situation.

| Programme détaillé

Jour 1

- **Découvrir Hadoop 2.X**
  - L'architecture de Hadoop 2.X
  - The Hortonworks Data Platform (HDP)

- **Le système de fichiers distribué HDFS**
  - Architecture fonctionnelle de HDFS
    - Exercice d'interaction en ligne de commande avec HDFS

- **Alimenter HDFS en données**
  - Prise en main de l'outil Flume
  - Prise en main de l'outil Sqoop
    - Utilisation de Sqoop pour transférer des données entre Hadoop et Microsoft SQL Server

- **Le Framework MapReduce**
  - Architecture et fonctionnement général de MapReduce
    - Exemples d'utilisation d'un job MapReduce

Jour 2

- **Introduction à Pig**
  - Types et mots-clés dans Pig
    - Exploration de données avec Pig

- **Programmation Pig avancée**
  - Mots-clés et fonctionnalités avancées dans Pig
  - Jointures dans Pig
  - Astuces d'optimisation de scripts Pig
    - Analyse de cas d'usages métier divers avec Pig

Jour 3

- Programmation Hive
  - Types et mots-clés dans Hive
  - Concept de table et base de données dans Hive
  - Présentation et explication des types de jointures
    - Démonstration de jointures
    - Analyse de cas d'usages métier

| Programme détaillé

- **Utiliser HCatalog**
  - Fonctionnement et utilisation de HCatalog
    - Démonstration du fonctionnement de HCatalog

Jour 4

- **Programmation Hive avancée**
  - Les vues dans Hive
  - Les différents formats de stockage des tables Hive
  - Optimisation de scripts Hive
    - Illustration des fonctions avancées

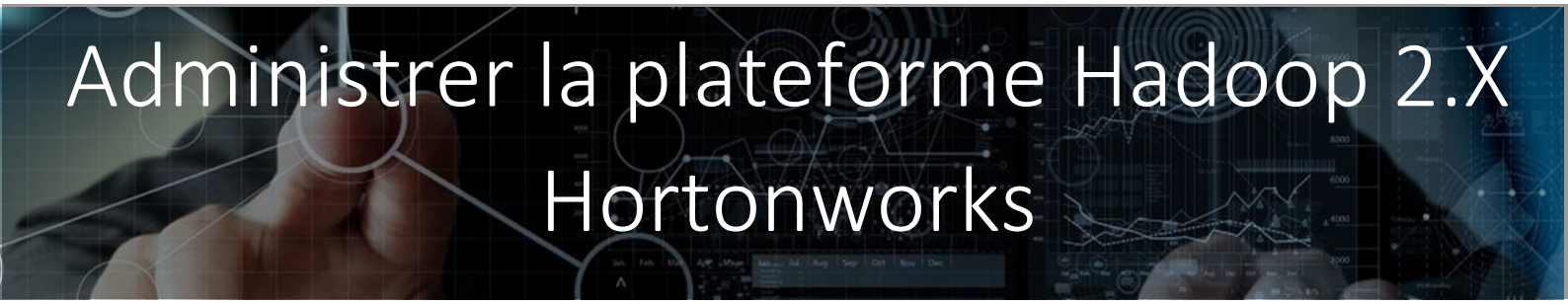
- **Le driver ODBC de Hive**
  - Connexion de Microsoft Excel à Hadoop

- **Hadoop 2.X et Yarn**
  - Architecture de Yarn
    - Démonstration d'une application Yarn

- **Créer et utiliser un Workflow Oozie**
  - Workflow et coordinateur Oozie
  - Actions possibles avec Oozie

Contact

☎ 01 75 43 05 58
✉ training@emgsgroup.com



Administrer la plateforme Hadoop 2.X Hortonworks

Code formation : EMPHH
Durée : 4 jours – 28h de cours
Format : Inter-entreprise*

4 jours
28 heures de cours

2550€
Tarif Inter-entreprise*/Hors taxes

**Cette formation est également disponible en « Intra-entreprise », nous contacter pour plus d'infos.*

| Description

Cette session prépare au rôle d'administrateur au sein d'un contexte technologique innovant et en particulier au cours d'un projet Big Data. A travers des exercices concrets, vous apprendrez à installer, configurer et maintenir un cluster Hadoop.

A la fin de cette formation, vous aurez une compréhension solide de comment Hadoop fonctionne avec le Big Data et, à travers nos mises en pratique, vous saurez déployer tout le cycle de vie pour des clusters multi-nœuds.

| Objectifs pédagogiques

- Dimensionner un cluster Hadoop
- Installer un cluster Hadoop
- Configurer un cluster Hadoop
- Sécuriser un cluster Hadoop
- Maintenir un cluster Hadoop

| Publics

Architecte, administrateur

| Pré-requis

Connaissances de l'environnement Linux.

| Méthode pédagogique

Formation avec apports théoriques, échanges sur les contextes des participants et retours d'expérience pratique du formateur, complétés de travaux pratiques et de mises en situation. Cette formation prépare à la certification éditeur Hortonworks.

| Programme détaillé

Jour 1

- **Big Data, Hadoop et la plateforme Hortonworks : les bases du Big Data**
  - Les produits de la HDP
  - Qu'est-ce que Hadoop ?
  - Une architecture de cluster type
  - Introduction à Ambari

- **Gestion des utilisateurs avec Ambari**
  - Gérer les users et les groupes
  - Gérer les permissions
  - Mise en pratique : « Gestion des utilisateurs avec Ambari »

- **Gestion des services Hadoop via Ambari**
  - Configuration des services
  - Surveillance des services
  - Maintenance des services
  - Mise en pratique : « Gestion des services Hadoop »

- **Utiliser le stockage HDFS**
  - Accéder aux données
  - Gestion des fichiers
  - Mise en pratique : « Utiliser le stockage HDFS »

Jour 2

- **Utiliser le stockage HDFS (suite)**
  - Les web services d'HDFS
  - Mise en pratique : « Utiliser WebHDFS »
  - Protéger les accès
  - Mise en pratique : « Utiliser les ACLs HDFS »

- **Gestion du stockage HDFS**
  - Architecture HDFS
  - Assurer l'intégrité de la donnée
  - Mise en pratique : « Gestion du stockage sur HDFS »
  - Les quotas HDFS
  - Mise en pratique : « Gestion des quotas sur HDFS »

| Programme détaillé

- **Gestion des ressources avec Yarn**
  - Architecture de Yarn
  - Utilisation de Yarn
  - Les différentes façons de gérer Yarn
  - Mise en pratique : « Configurer et gérer Yarn »
  - Mise en pratique : « Gestion de Yarn sans Ambari »

Jour 3

- **Découverte des applications Yarn**
  - Les bases d'une application Yarn
  - Mise en pratique : « Démarrer une application Yarn »

- **Gestion des nœuds dans un cluster**
  - Ajouter, enlever un nœud du cluster
  - Déplacer des composants
  - Mise en pratique : « Ajouter, décommissionner et recommissionner un nœud »

- **Le capacity scheduler de Yarn**
  - Contrôler la répartition des ressources grâce aux queues Yarn
  - Contrôler les accès sur les queues Yarn
  - Mise en pratique : « Configuration des utilisateurs et des groupes pour Yarn »
  - Mise en pratique : « Configurer les ressources avec les queues »
  - Mise en pratique : « Tuning de la gestion des ressources »

- **Gestion des racks sur Hadoop**
  - Les bénéfices de la « rack awareness »
  - Configurer la « rack awareness »
  - Mise en pratique : « Configurer la rack awareness »

Jour 4

- **Activer la haute disponibilité avec HDFS et Yarn**
  - Les principes de la haute disponibilité
  - Haute disponibilité du Namenode
  - Haute disponibilité du Resource manager
  - Mise en pratique : « Configurer la haute disponibilité du namenode »
  - Mise en pratique : « Configurer la haute disponibilité du resource manager »

- **Surveillance de cluster**
  - Surveillance avec Ambari
  - Lever des alertes avec Ambari
  - Mise en pratique : « Configurer les alertes avec Ambari »

| Programme détaillé

- **Protéger ses données**
  - De l'importance des backups
  - Les snapshots HDFS
  - Utiliser DistCP
  - Mise en pratique : « Gestion des snapshots HDFS »
  - Mise en pratique : « Utiliser DistCP »

- **Installer la HDP**
  - Identifier les options de déploiement de cluster
  - Planifier un déploiement de cluster
  - Faire une installation avec Ambari
  - Mise en pratique : « Installer la HDP »

Contact
☎ 01 75 43 05 58
✉ training@emgsgroup.com

Analyse de données pour Hadoop 2.X Hortonworks avec Pig & Hive

Code formation : EMHPH
Durée : 4 jours – 28h de cours
Format : Inter-entreprise*

4 jours
28 heures de cours

2550€
Tarif Inter-entreprise*/Hors taxes

Cette formation est également disponible en « Intra-entreprise », nous contacter pour plus d'infos.

| Description

Cette formation présente les grands outils de l'écosystème Hadoop en se focalisant plus spécifiquement sur Pig et Hive. Le principal objectif est le développement de compétences de data analyst orientées accès et traitement des données sans nécessairement avoir un fort background technique.

| Objectifs pédagogiques

- Identifier et définir les différents composants de l'écosystème Hadoop
- Appréhender l'architecture de Hadoop 2.X
- Expérimenter les outils d'exploration et d'analyse avancée de données

| Publics

Analyste, statisticien, développeur

| Pré-requis

Connaissances de base en scripting (SQL, Python, R) ou en programmation

| Méthode pédagogique

Formation mêlant des apports théoriques à de nombreux travaux pratiques sous forme d'exercices d'application et d'analyse de uses cases métier complétés des retours d'expérience du formateur.

| Programme détaillé

Jour 1

- **Comprendre Hadoop 2.X**
  - L'architecture de Hadoop 2.X
  - The Hortonworks Data Platform (HDP)

- **Le système de fichiers distribué HDFS**
  - Architecture fonctionnelle de HDFS
  - Exercice d'interaction en ligne de commande avec HDFS

- **Alimenter HDFS en données**
  - Prise en main de l'outil Flume
  - Prise en main de l'outil Sqoop
  - Application de ces deux outils d'import et d'export des données

- **Le framework MapReduce**
  - Architecture et fonctionnement général de MapReduce
  - Exemples d'utilisation d'un job MapReduce

Jour 2

- **Introduction à Pig**
  - Types et mots-clés dans Pig
  - Exploration de données avec Pig

- **Programmation Pig avancée**
  - Mots-clés et fonctionnalités avancées dans Pig
  - Jointures dans Pig
  - Astuces d'optimisation de scripts Pig
    - Analyse de cas d'usages métier divers avec Pig

Jour 3

- **Programmation Hive**
  - Types et mots-clés dans Hive
  - Concept de table et base de données dans Hive
  - Présentation et explication des types de jointures
  - Démonstration de jointures
    - Analyse de cas d'usages métier

- **Utiliser HCatalog**
  - Fonctionnement et utilisation de HCatalog
    - Démonstration du fonctionnement de HCatalog

| Programme détaillé

Jour 4

- **Programmation Hive avancée**
  - Les vues dans Hive
  - Les différents formats de stockage des tables Hive
  - Optimisation de scripts Hive
    - Illustration des fonctions avancées

- **Hadoop 2.X et Yarn**
  - Architecture de Yarn
    - Démonstration d'une application Yarn

- **Créer et utiliser un workflow Oozie**
  - Workflow et coordinateur Oozie
  - Actions possibles avec Oozie

Contact

☎ 01 75 43 05 58
✉ training@emgsgroup.com

Développer des applications pour YARN avec Hadoop 2.X Hortonworks

Code formation : EMYHH
Durée : 2 jours – 14h de cours
Format : Inter-entreprise*

2 jours
14 heures de cours

Nous contacter pour le tarif
Tarif Inter-entreprise*/Hors taxes

**Cette formation est également disponible en « Intra-entreprise », nous contacter pour plus d'infos.*

| Description

Cet atelier vous présente le fonctionnement détaillé de YARN et la méthodologie pour développer ses propres applications avec le framework YARN. Durant ces deux journées, nous aborderons les différents patterns d'architecture logicielle avec YARN et les possibilités d'interactions avec Hadoop. Cet atelier permettra aux participants d'avoir une parfaite compréhension du fonctionnement de YARN et la maitrise de son API.

| Objectifs pédagogiques

- Identifier et définir les différents composants de YARN.
- Appréhender le fonctionnement détaillé de YARN.
- Utiliser l'API YARN pour développer des applications Java.
- Configurer le Job Scheduler.
- Maîtriser le contexte d'exécution des conteneurs.

| Publics

Architecte, développeur, expert technique

| Pré-requis

- Bonne connaissance du langage Java.
- La connaissance de l'environnement Linux est un plus.

| Méthode pédagogique

Formation avec apports théoriques, échanges sur les contextes des participants et retours d'expérience pratique du formateur, complétés d'exercices pratiques et de mises en situation.

| Programme détaillé

Jour 1

- **Présentation de l'architecture de Yarn**
  - L'architecture Yarn
  - Les différences entre Hadoop 1 et Hadoop 2
  - Management des logs
  - Administration basique
  - Exercice pratique : exécuter un shell distribué

- **Définition d'une application Yarn**
  - Le cycle de vie d'une application
  - L'API YARN
  - La gestion des dépendances : Local Resource
  - Exercice pratique : installer l'environnement

- **Développer une application Yarn**
  - Interagir avec le Resource Manager
  - Prérequis d'une application de type Yarn client
  - Récupération des métriques et monitoring de son application
  - Exercice pratique : développer un client Yarn

Jour 2

- **Développer son propre application master**
  - Prérequis et fonction d'un Application Master
  - Pattern synchrone ou asynchrone
  - Allocation des ressources
  - Monitoring des conteneurs
  - Exercice pratique : développer un Application Master

- **Traiter avec les conteneurs**
  - Démarrer un conteneur
  - Communiquer avec l'Application Master
  - Ecrire ses propres conteneurs personnalisés
  - Co-localisation des données : communiquer avec HDFS
  - Exercice pratique : développer une application Java s'exécutant dans un conteneur

| Programme détaillé

- **Ordonnancer un Job Yarn**
  - Présentation du Capactity Scheduler
  - Présentation du Fair Scheduler
  - Configuration du scheduler dans Yarn

EMGS GROUP
CONSULTING | RECRUITMENT | TRAINING

Contact
☎ 01 75 43 05 58
✉ training@emgsgroup.com

Analyse de données pour Hadoop 2.X Hortonworks avec HBase

Code formation : EMHHB
Durée : 2 jours – 14h de cours
Format : Inter-entreprise*

2 jours
14 heures de cours

1480€
Tarif Inter-entreprise*/Hors taxes

**Cette formation est également disponible en « Intra-entreprise », nous contacter pour plus d'infos.*

| Description

Inspirée des publications de Google sur BigTable, HBase est un SGBD non relationnel capable de gérer d'énormes quantités de données. Intégré à l'écosystème Hadoop, il permet de distribuer les données en utilisant le système de fichiers distribué HDFS (Hadoop Distributed File System) du framework.

Son fonctionnement, qui repose donc sur le stockage distribué des données sur un cluster de machines physiques, garantit à la fois la haute disponibilité et les hautes performances des bases. Deux arguments de poids qui suffisent à comprendre le succès croissant de la solution. A l'issue de cette formation, les participants disposeront des connaissances et compétences nécessaires à la mise en œuvre de HBase.

| Objectifs pédagogiques

- Découvrir le fonctionnement de HBase
- Savoir configurer et utiliser HBase
- Modéliser une table HBase
- Prendre en main et utiliser les différents outils à dispositionInteragir avec le framework Hadoop.

| Publics

Architecte, administrateur, développeur de logiciel, analyste

| Pré-requis

- Connaissance de l'environnement Linux.
- Connaissance des bases de données relationnelles.

| Méthode pédagogique

Une pédagogie basée sur l'alternance de phases théoriques et de mises en pratique qui permet aux participants d'acquérir une première expérience concrète du stockage de données en environnement HBase. Des échanges avec l'intervenant sur les meilleures pratiques pour garantir la disponibilité des bases et leurs performances.

| Programme détaillé

Jour 1

- **Généralités**
  - Rappels rapides sur l'écosystème Hadoop
  - Hortonworks
  - HDFS
  - Yarn
  - Mise en pratique : lancement d'une tâche MapReduce

- **Introduction a HBase**
  - Fonctionnement général
  - Mise en pratique :
  - Utilisation du client HBase
  - Importation d'une table MySQL avec Sqoop

- **Architecture**
  - Fonctionnement et cycle de vie des régions HBase
  - Orchestration du cluster avec Zookeeper
  - Mise en pratique : manipulation des nœuds ZooKeeper

- **Le framework MapReduce**
  - HMaster et RegionServer
  - Opérations : get, put, scans
  - Mise en pratique : prise en main des fichiers de configuration

Jour 2

- **Commandes**
  - Manipulation des données
  - Manipulation des tables
  - Réparations
  - Réplication de clusters
  - Mise en pratique : utilisation des commandes

| Programme détaillé

- **Configuration et distributions**
  - Configuration HBase et Zookeeper
  - Distributions HBase
  - Backups
  - Mises en pratique :
    - Backup et snapshots
    - Exports avec Pig, imports avec Importtsv

- **Modèle de données HBase**
  - Designer les clés de lignes : patterns et techniques
  - Modèle de table pour une application de messagerie
  - Familles de colonnes
  - Mise en pratique :
    - Familles de colonnes
    - Etude de cas : application de suivi de colis

- **Optimisation**
  - Blocs, caches, filtres de bloom, memstore, logs
  - Filtres de colonnes
  - Mise en pratique : manipulation des filtres de bloom

- **Démonstration : réalisation d'un client Java HBase**

EMGS GROUP
CONSULTING | RECRUITMENT | TRAINING

Contact
☎ 01 75 43 05 58
✉ training@emgsgroup.com

Administrer la base de données HBase avec Hadoop 2.X Hortonworks

Code formation : EMHHB
Durée : 4 jours – 28h de cours
Format : Inter-entreprise*

4 jours
28 heures de cours

2550€
Tarif Inter-entreprise*/Hors taxes

Cette formation est également disponible en « Intra-entreprise », nous contacter pour plus d'infos.

| Description

Inspirée des publications de Google sur BigTable, HBase est un SGBD non relationnel capable de gérer d'énormes quantités de données. Intégré à l'écosystème Hadoop, il permet de distribuer les données en utilisant le système de fichiers distribué HDFS (Hadoop Distributed File System) du framework.

Son fonctionnement, qui repose donc sur le stockage distribué des données sur un cluster de machines physiques, garantit à la fois la haute disponibilité et les hautes performances des bases. Deux arguments de poids qui suffisent à comprendre le succès croissant de la solution.

A l'issue de cette formation, les participants disposeront des connaissances et compétences nécessaires à la mise en œuvre de HBase.

| Objectifs pédagogiques

- Savoir installer HBase
- Sécuriser les accès cluster
- Assurer la maintenance des données
- Monitorer HBase pour faciliter la résolution de problème
- Optimiser les performances

| Publics

Architecte, administrateur, analyste, développeur, tech lead, chef de projet, gestionnaire de bases de données

| Pré-requis

Connaissance de l'environnement Linux et des SGBD relationnels

| Méthode pédagogique

Formation avec apports théoriques, échanges sur les contextes des participants et retours d'expérience pratique du formateur, complétés de travaux pratiques et de mises en situation.

| Programme détaillé

Jour 1

- **Introduction à Hadoop**
  - Présentation de cas d'usage big data
  - Différents types de données : structurée, non structurée
  - Les caractéristiques d'un projet big data
  - Hadoop et Hortonworks

- **HDFS et Yarn**
  - Démonstration pour la découverte et l'utilisation de HDFS (Hadoop Distributed File System)
  - Architecture de Yarn

- **Introduction à HBase**
  - Modèle clé-valeur, lignes, colonnes
  - Architecture
  - Phoenix, une solution pour requêter HBase en SQL

Jour 2

- **Installation**
  - Installations de Standalone : semi-distribuées ou pleinement distribuées
  - Prérequis demandés : Java, Zookeeper, Master Server, Region Server
  - Installation manuelle ou automatisée avec Ambari
  - Mise en pratique : « Installation automatisée avec Ambari »
  - Mise en pratique : « Installation manuelle »
  - Vérification de l'installation

- **Utilisation du client**
  - Commandes générales
  - Manipulation des tables
  - Mise en pratique : « Exploration des commandes »
  - Mise en pratique : « Administration du cluster (régions, balance, etc.) »

| Programme détaillé

- **Ingestion de données**
  - Composants impliqués dans le stockage (re, Write Ahead log, Memstore, HFile, etc.)
  - Modèle de stockage : paires clés-valeur, clés de lignes, familles de colonnes, etc.
  - Appréhender les mécanismes de lecture et d'écriture de données
  - Flush process (memstore, etc.)
  - Compactage des régions
  - Ingestion de données en masse (import tsv, coompletebulkload)
  - Mise en pratique : « Utiliser ImportTSV pour ingérer des données »
  - CopyTable (use cases, exemples)

Jour 3

- **Gestion des opérations**
  - Utilisation d'Ambari pour gérer HBase
  - Haute disponibilité (sauvegarde des maires, lectures HA)
  - Mise en pratique : « Haute disponibilité »
  - Log files (log4j, Linux, GUI Master Server)
  - Mise en pratique : « Log files »
  - Coprocesseur : le pendant des triggers SQL
  - Filtres (filtres de scan, filtres customs)

- **Sauvegarde et restauration des données**
  - Protection des données : réplication HDFS, réplication de clusters, backup et snapshots
  - Mise en pratique : « Snapshots »
  - Réplication de cluster : topologies, configuration
  - Mise en pratique : « Réplication »
  - Snapshots hbase : processus, création et gestion, travailler avec

- **Sécurité**
  - Authentification
  - Autorisations et Access Control Lists
  - Mise en pratique : « Autorisations et Access Control Lists »
  - Commandes Hbase Shell relatives à la sécurité
  - Ranger : un outil pour configurer les autorisations sur l'ensemble du cluster
  - Knox : un point d'accès sécurisé au cluster
  - Authentifications simples
  - Bulk load secure

| Programme détaillé

Jour 4

- Monitoring HBase et diagnostic des problèmes
  - Métriques importantes (Master Server, Region Server)
  - Les outils de monitoring HBase : Nagios, Ganglia, OpenTSDB
  - Identifier les HotSpots
  - Mise en pratique : « Identifier les hotspots »
  - Eviter les hotspots par le design des clés de ligne
  - Utiliser le pré-split

- **Maintenance**
  - Split de régions
  - Mise en pratique : « Split de régions »
  - Load balancer
  - Monitoring de la taille des régions
  - Split et merge manuel de régions
  - Problèmes d'intégrité (utilisation de HBCK)

- **Résolution de problèmes**
  - Vérification des statuts Zookeeper
  - Monitoring des garbage collection de la JVM
  - Mise en pratique : « Monitoring des garbage collection de la JVM
  - Résolution des erreurs au démarrage des serveurs HBase
  - Régler HBase pour obtenir de meilleures performances
  - Régler HDFS pour obtenir de meilleures performances

- **Projet récapitulatif**

FORMATIONS

Hadoop

EMGS GROUP
CONSULTING | RECRUITMENT | TRAINING

Contact
☎ 01 75 43 05 58
✉ training@emgsgroup.com

Déployer & gérer un cluster Couchbase

Code formation : EMCLU
Durée : 4 jours – 28h de cours
Format : Inter-entreprise*

4 jours
28 heures de cours

2295€
Tarif Inter-entreprise*/Hors taxes

Cette formation est également disponible en « Intra-entreprise », nous contacter pour plus d'infos.

| Description

Cette formation Couchbase pour administrateurs apportera aux participants les concepts architecturaux et l'expertise nécessaire à la définition, au déploiement, et à l'opération de clusters Couchbase. Ils découvriront et expérimenteront les procédures et outils dont un administrateur a besoin pour opérer des plateformes critiques et temps-réel utilisant Couchbase.

| Objectifs pédagogiques

- Installer et configurer un cluster Couchbase
- Lancer des tests de charge sur un cluster, et le monitorer
- Savoir partitionner et rééquilibrer un cluster, ajouter et supprimer des nœuds, réaliser un backup et une restauration
- Résoudre les problèmes courants (trouble shooting)

| Publics

Architecte NoSQL, futur administrateur, couchbase, administrateur système, développeur

| Pré-requis

- Expérience en administration système (*nix), ou DBA.
- Connaissance des bases de données relationnelles.
- Aisance avec l'anglais écrit.

| Méthode pédagogique

Formation avec apports théoriques, échanges sur les contextes des participants et retours d'expérience pratique du formateur, complétés de travaux pratiques et de mises en situation.

| Programme détaillé

Jour 1

- Aperçu de l'architecture et des possibilités de Couchbase server 4.0
- Principaux cas d'utilisation de couchbase
- Guide de dimensionnement en fonction de la charge
- Guide d'installation et bonnes pratiques de configuration
- Montée de version de couchbase en production

Jour 2

- Les vbuckets
- Moteur de stockage : cache, persistence sur le disque et réplication réseau
- Préchargement du cache
- Réplication entre les nœuds : gestion de la cohérence par couchbase
- Configuration d'une application cliente pour utiliser couchbase
- Détermination par le client du nœud à contacter pour l'accès aux données
- Création d'un bucket et ajout de données

Jour 3

- Gestion asynchrone des suppressions : pierres tombales et compactage
- Éjection, éviction et gestion du jeu de données actif
- Rééquilibrage d'une grappe après l'ajout ou la suppression de nœuds
- Vues et index
- Gestion des zones de réplication

Jour 4

- Tâches d'administration : augmentation/réduction de la taille d'une grappe,
- Gestion des indisponibilités
- Performances et supervision d'une grappe
- Réplication inter-datacenter et gestion des conflits

| Programme détaillé

- Sauvegardes et restaurations d'une grappe
- Trucs et astuces de résolution des problèmes
- Utilisation de la console web pour administrer et superviser couchbase
- Utilisation des outils en ligne de commande pour administrer couchbase

EMGS GROUP
CONSULTING | RECRUITMENT | TRAINING

Contact
☎ 01 75 43 05 58
✉ training@emgsgroup.com

Développer des applications avec Couchbase

Code formation : EMCOU
Durée : 3 jours – 21h de cours
Format : Inter-entreprise*

3 jours
21 heures de cours

1895€
Tarif Inter-entreprise*/Hors taxes

Cette formation est également disponible en « Intra-entreprise », nous contacter pour plus d'infos.

| Description

Cette formation Couchbase pour développeur apportera aux participants les connaissances nécessaires au développement d'applications Web / NoSQL. Ils découvriront et expérimenteront les procédures et méthodes nécessaires aux cas d'utilisation typiques, tant au niveau du langage que de la modélisation, rencontrés lors du développement d'une application critique et temps-réel utilisant Couchbase.

| Objectifs pédagogiques

- Installer et configurer un cluster Couchbase
- Savoir développer un use-case typique de CRUD
- Modéliser des données dans une base orientée document Utiliser vue (Map / Reduce) pour requêter les données

| Publics

Développeur, architecte NoSQL

| Pré-requis

- Expérience dans un langage de programmation moderne (Java, C/C++, C#/.Net, Python, Ruby, PHP, etc.).
- Connaissance des bases de données relationnelles.
- Aisance avec l'anglais écrit. Remarque : il est recommandé de se documenter, avant le cours, sur l'extension ReactiveX de Java (rxjava).

| Méthode pédagogique

Formation rythmée par des apports théoriques et des ateliers de mise en pratique. Chaque participant crée son propre cluster et s'y connecte via le client Java pour y effectuer des opérations. Les ateliers porteront sur le développement d'une application de gestion de playlist musicale, qui sera élaborée au fur et à mesure des différents modules.

| Programme détaillé

Jour 1

- **Introduction à Couchbase server**
  - Écosystème Couchbase Server
  - Principes fondamentaux de Couchbase Server
  - Architecture de Couchbase Server 4.0
  - Anatomie d'une application Couchbase
  - Où se trouvent les données ?
  - Hiérarchie des données
  - Deux types de bucket
  - Opérations dans Couchbase
  - Nœud unique : opérations d'écriture
  - Nœud unique : opérations de mise à jour
  - Nœud unique : opérations de lecture
  - Nœud unique : éjection du cache
  - Nœud unique : Données manquantes en cache
  - Utilisation de la console d'administration Web
  - Aperçu du SDK Couchbase Java 2.0
  - Introduction à l'application CouchMusic
  - Chargement massif de documents JSON avec Cbdocloader

- **Utilisation du SDK Java**
  - Gestion des connections
  - L'interface Cluster
  - L'interface Bucket
  - Travailler avec des documents
  - L'interface Document
  - L'interface Transcoder
  - Les méthodes insert de l'interface Bucket
  - Les méthodes get de l'interface Bucket
  - Les méthodes replace de l'interface Bucket
  - Les méthodes upsert de l'interface Bucket
  - Les méthodes remove de l'interface Bucket
  - Aperçu de la programmation asynchrone
  - Introduction à RxJava
  - La méthode async de l'interface Bucket
  - La classe Observable

| Programme détaillé

Jour 2

- **Utilisation des vues**
  - Tirer profit de la puissance des vues
  - Moteur de vues de Couchbase
  - Introduction à MapReduce
  - Vues de développement vs de production
  - Code source des vues
  - Introduction à l'API de requêtes sur les vues
  - Tri des résultats d'une requête
  - Indexation et requêtage
  - Requêtes sur des plages de valeurs

- **Modélisation des données**
  - Schémas implicites vs explicites
  - Dénormalisation
  - Clés naturelles vs artificielles
  - Définition d'unmotif de clé
  - Motif de clé basé sur un compteur
  - Motif de recherche
  - Motif de recherche inversée

- **Ingestion de données**
  - Connections client trop nombreuses
  - Cache de configuration inutilisé
  - Utilisation du ConfigCache
  - Non utilisation des vues avec de gros documents
  - Huit raisonnements discutables
  - Bien écrire une vue
  - Quand créer plusieurs buckets
  - Mélanger la liste de nœuds avant la connexion
  - Pourquoi réutiliser un objet
  - Ai-je besoin d'utiliser la lecture depuis les répliqua pour améliorer les performances ?

| Programme détaillé

Jour 3

- **Intégration avec Elastic search**
  - Réplication inter-datacenter (XDCR)
  - Configuration de XDCR
  - Intégration avec Elastic Search
  - Recherche à l'intérieur de documents JSON
  - Recherche plain texte
  - Terminologie Elastic Search

- **Recherche plain texte Couchbase**
  - Fonctionnement
  - Comment bien commencer ?
  - Installation du greffon Couchbase
  - Indexation des documents
  - Score des résultats
  - Requête simple par HTTP
  - Type d'utilisation recommandé

- **Possibilités avancées**
  - Recherche à facettes
  - Requêtes à facettes
  - Résultats de requêtes à facettes
  - Support de recherches géographiques
  - Possibilités impliquant Elastic Search
  - Ressources Elastic Search
  - Chiffrement des communications
  - Éjection des méta-données du cache

Contact

☎ 01 75 43 05 58
✉ training@emgsgroup.com

Savoir utiliser & configurer Elasticsearch

Code formation : EMELA
Durée : 3 jours – 21h de cours
Format : Inter-entreprise*

3 jours
21 heures de cours

1490€
Tarif Inter-entreprise*/Hors taxes

Cette formation est également disponible en « Intra-entreprise », nous contacter pour plus d'infos.

| Description

Elasticsearch est un moteur de recherche conçu dès le départ pour être distribué et gérer des volumes de données massifs. Il se base sur la librairie Apache Lucene et lui ajoute des fonctionnalités supplémentaires pour la mise en cluster, la haute disponibilité ainsi qu'une API puissante.

Cette formation a pour objet de présenter Elasticsearch et toutes les notions importantes pour développer de façon efficace avec Elasticsearch. Elle est aussi l'occasion de jeter un œil sous le capot pour comprendre plus en profondeur le fonctionnement d'Elasticsearch et en tirer le meilleur.

| Objectifs pédagogiques

- Appréhender Elasticsearch et son API
- Découvrir les concepts essentiels (mapping, analyse)
- Assimiler quel type de recherche est adapté à chaque use-cases et comment modéliser
- Maîtriser le fonctionnement d'Elasticsearch pour l'utiliser efficacement

| Publics

Développeur, architecte

| Pré-requis

Disposer de notions sur http

| Méthode pédagogique

Formation rythmée par des apports théoriques et des ateliers de mise en pratique.

| Programme détaillé

Jour 1

- **Introduction**
  - Pourquoi un moteur de recherche ?
  - Pourquoi Elasticsearch ?

- **Notions de base**
  - Node
  - Cluster
  - Index
  - Type
  - Shard
    - Primary
    - Replica
  - Document
  - Mapping
  - Score

- **Prise en main**
  - Installation
  - Configuration
  - Mise en cluster
  - Structure de l'api REST

- Indexation de documents
  - Création d'un index et d'un type
  - Indexation d'un document
  - Suppression d'un document
  - Mise à jour de documents
  - Version

- Analyse de document
  - Mapping et types de champs
  - Propriétés des champs
  - Customisation du mapping
  - Définition d'analyseurs
  - Cas d'usage
    - Langues humaines
    - Index multilingue
    - Typos et problèmes d'orthographe

| Programme détaillé

Jour 2

- **Requêtes**
  - Structure d'un index
  - Logique
    - Physique
  - Queries
    - Types de requêtes
    - Simples
    - Texte
    - Géographique
    - Recherche approximative et tolérance aux fautes
    - Pertinence et score
    - Comprendre le calcul du score avec explain
    - Fonctions pour le score
  - Filtres
    - Nested
    - Parent-child
    - Cycle de vie d'une requête

- **Agrégations**
  - Fonctionnement
  - Notion de scope
  - Types d'agrégations

Jour 3

- **Percolation**

- **Benchmark**

- **Gestion des index**
  - API indices
  - Templates

- **Clustering**
  - Communication entre nœuds
    - Rôles des nœuds
    - Notion de master
  - La vie d'une requête distribuée

| Programme détaillé

- **Elasticsearch en production**
  - Performance
  - Configuration
  - Indexation en masse
  - Monitoring
  - Répartition des index
  - Backups
  - API cat

- **Plugins**
  - Types de plugins
  - Rivers
  - Langages de script
  - Fonctionnalités
  - Installation