Big Data : architecture et infrastructure

Lien : https://innov-systems.com/formation/big-data-architecture-et-infrastructure

DURÉE
5 jours (35h)

RÉFÉRENCE
BSI13

CATÉGORIE
Big Data : fondamentaux et architecture

OBJECTIFS DE LA FORMATION

À l'issue de cette formation, vous serez capable de :

- Maîtriser les aspects fondamentaux de Big Data
- Maîtriser des éléments clés liés aux architectures des SI
- Pouvoir proposer des architectures répondant aux besoins du client
- Comprendre les différents cas d’usage et les outils correspondants
- Avoir une vision globale des différents outils et Framework
- Mettre en pratique les connaissances sur des problématiques réelles

POUR QUI ?

- Administrateurs de clusters Hadoop
- DBA
- Ingénieurs systèmes et réseaux
- Développeurs

Innov Systems

Programme détaillé

1 / Comprendre le Big Data

- Introduction au Big Data
- Les 5V du Big Data
- Technologies de référence du Big Data
- Les technologies concernées
- Les outils
- Les langages : Hortonworks, MapR, Cloudera, IBM Watson
- Démystification du Big Data
- Les acteurs principaux
- Les différents métiers du Big Data
- Différences entre BI, Big Data et Data Science
- Exemples pratiques

2 / Les questions clés sur l'architecte Big Data

- Limites des architectures classiques
- La nouvelles architecture Big Data et son impact sur le SI
- Cycle de vie de la donnée
- Par quelles transformations passe-t-elle ? Comment la gouverner ?
- Les principaux modèles d'architecture d'un SI Big Data
- Définir le "Data Lake" ?
- Le data lake : une nouvelle philosophie pour le stockage et le traitement de la donnée
- Les avantages du data lake et centralisation des données
- Dimensionner l'infrastructure
- Mesurer la scalabilité du système ?
- Cycle de vie d'un projet Big Data
- Cas d'usage, mise en situation

Innov Systems

3 / Solutions Big Data

- Projets, applications, plateformes
- Solutions de stockage des données
- Considérations Hardware
- Stockage
- CPU
- Mémoire
- Réseau
- Systèmes distribués
- La science et l'art de l'analyse prédictive

4 / Architectures distribuées

- Problématiques et enjeux
- Des données cohérentes, disponibles et tolérantes aux pannes ?
- Les architectures massivement parallèles
- L’ouverture aux traitements complexes (datamining, machine learning, etc.)
- Paradigmes de calculs distribués
- Les bases NoSQL et le calcul distribué (exemple avec MongoDB)

5 / Qualité des données (dataquality)

- Liens entre infrastructure et qualité des données
- Les 4 V
- Base à chaud, base à froid
- Les apports d’un outil de Dataquality
- L'utilité d'un ETL
- Illustration via Talend Data Integration
- Analyser les données en les fusionnant avec les données internes
- Le Master Data Management (MDM)

Inov Systems

6 / Installation d'une plateforme Hadoop

- Type de déploiement
- Installation d'Hadoop
- Installation d'autres composants (Hive, Pig, HBase, Flume...)
- Différences entre les distributions Cloudera, Hortonworks et MapR

7 / Stockage et traitement de la donnée

- Le système de fichiers distribué de Hadoop (HDFS)
- MapReduce
- YARN
- Spark

8 / Les bases de données et la gestion des données

- NoSQL (Cassandra, MongoDB)
- NoSQL Vs SGBD Relationnels
- Classification des bases de données NoSQL selon leurs types et leurs cas d'usages

9 / L’analyse et la visualisation des données

- Spark MLIB
- Mahout
- Hadoop Streaming
- Pig et MapReduce
- Hadoop Image Processing Interface (HIPI)

10 / Monitoring du cluster Hadoop

- Suivi de charges
- Journaux (jConsole)
- Gestion des nœuds

Innov Systems

- Accès JMX
- Mise en œuvre d'un client JMX
- Administration HDFS
- Stockage des fichiers : fsck, dfsadmin
- Gestion centralisée de caches avec Cacheadmin

11 / Supervision des plateformes

- Supervision des éléments par le NodeManager
- Monitoring graphique avec Ambari, Kibana, Cloudera Manager
- Visualisation des alertes en cas d'indisponibilité d'un nœud
- Configuration des logs avec log4j

12 / Traiter les données en temps réel avec spark streaming

- Notion de Streaming
- Objectifs, principe de fonctionnement : stream processing
- Présentation des Discretized Streams (DStreams)
- Source de données : HDFS, Flume, Kafka, ...
- Manipulation de l'API
- Comparaison avec Apache Storm

13 / La sécurité et la gouvernance de données

- Signes d’alerte pour mieux gérer la gouvernance et la sécurité des données
- Que faire si vous identifiez ces signes d’avertissement
- La sécurité et le contrôle d’accès Kerberos & Knox
- Atlas et Ranger pour la sécurité et la gouvernance

Inov Systems

Approche pédagogique

- Support Ecrit et Projection
- Exposés Interactifs, Podcasts et Vidéos
- Brainstorming et Jeux de Rôle
- Cas Pratiques et Labs inclus pour leur impact opérationnel
- Test de Validation des Acquis des Connaissances

Prochaines dates programmées

04 au 08 Jan. 2027	Présentiel - Casablanca
01 au 05 Mars 2027	Présentiel - Casablanca
26 au 30 Avr. 2027	Présentiel - Casablanca
21 au 25 Juin 2027	Présentiel - Casablanca
16 au 20 Août 2027	Présentiel - Casablanca

Autres dates possibles sur demande. Contactez-nous pour organiser une session intra-entreprise.

Réservation & Renseignements

Téléphone : +212 522 247 210
Email : contact@innov-systems.com
Web : https//www.innov-systems.com

Innov Systems — Tous droits réservés