



## Cycle certifiant Administrateur BigData

CB095

Durée: 18 jours

9 570 €

### Public :

Chefs de projet, exploitants, administrateurs

### Objectifs :

Comprendre les caractéristiques d'une infrastructure BigData, les contraintes de production et de supervision. Savoir définir les points à surveiller, et connaître les outils et solutions disponibles pour l'administration BigData.

### Connaissances préalables nécessaires :

Connaissances générales des systèmes d'information et des outils et techniques d'exploitation et d'administrations

### Programme :

#### Comprendre les principaux concepts du Big Data ainsi que l'écosystème technologique d'un projet Big Data

L'essentiel du BigData : calcul distribué, données non structurées. Besoins fonctionnels et caractéristiques techniques des projets. La valorisation des données. Le positionnement respectif des technologies de cloud, BigData et noSQL, et les liens, implications.

Concepts clés : ETL, Extract Transform Load, CAP, 3V, 4V, données non structurées, prédictif, Machine Learning.

L'écosystème du BigData : les acteurs, les produits, état de l'art. Cycle de vie des projets BigData.

#### Savoir analyser les difficultés propres à un projet Big Data

Rôle de la DSI dans la démarche BigData. Gouvernance des données: importance de la qualité des données, fiabilité, durée de validité, sécurité des données

Emergence de nouveaux métiers : Data-scientists, Data labs, Hadoop scientists, CDO, ...

Intégration avec les outils statistiques présents et les outils BigData futurs.

#### Déterminer la nature des données manipulées

Les différents modes et formats de stockage.

Les types de bases de données : clé/valeur, document, colonne, graphe. Besoin de distribution. Définition de la notion d'élasticité. Principe du stockage réparti.

Données structurées et non structurées, documents, images, fichiers XML, JSON, CSV, ...

Atelier : installation d'un lac de données Hadoop/HDFS

Atelier : démonstrations avec une base MongoDB et une base Cassandra sur des données de différents types.



# Phirio

---

## Appréhender les éléments de sécurité, d'éthique et les enjeux juridiques

Les risques et points à sécuriser dans un système distribué.

Aspects législatifs et éthiques: sur le stockage, la conservation de données, ..., sur les traitements, la commercialisation des données, des résultats

Atelier : mise en évidence des problèmes liés à la réplication inter-régions et concernant les aspects juridiques des données : droits d'exploitation, propriété intellectuelle, ...

Etude des failles de sécurité sur une infrastructure Hadoop.

---

## Exploiter les architectures Big Data

Les objectifs de la supervision, les techniques disponibles. La supervision d'une ferme BigData.

Objets supervisés. Les services et ressources. Protocoles d'accès. Exporteurs distribués de données.

Définition des ressources à surveiller. Journaux et métriques.

Application aux fermes BigData : Hadoop, Cassandra, HBase, MongoDB

Besoin de base de données avec agents distribués, de stockage temporel (timeseriesDB)

Produits : Prometheus, Graphite, ElasticSearch.

Présentation, architectures.

Les sur-couches : Kibana, Grafana.

Atelier : mise en oeuvre de prometheus pour la supervision d'une ferme Cassandra sur une infrastructure distribuée multi-noeuds.

---

## Mettre en place des socles techniques complets pour des projets Big Data.

Etude des différents composants d'une infrastructure BigData :

Datalake : collecte des différents types de données

Stockage distribué : réplication, sharding, gossip, hachage,

Principe du schemaless, schéma de stockage, clé de distribution, clé de hachage

Systèmes de fichiers distribués : GFS, HDFS, Ceph. Les bases de données : Redis, Cassandra, DynamoDB,

Accumulo, HBase, MongoDB, BigTable, Neo4j, ...

Calcul et restitution : Apport des outils de calculs statistiques

Langages adaptés aux statistiques, liens avec les outils BigData.

Outils de calcul et visualisation :R, SAS, Spark, Tableau, QlikView, ...

Caractéristiques et points forts des différentes solutions.

Atelier : mise en oeuvre du sharding avec une base de données MongoDB sur une infrastructure distribuée

---

## Cassandra

Cassandra

---

## Cassandra

Introduction

Historique, fonctionnalités de Cassandra, licence

Format des données,"key-value", traitement de volumes importants,

haute disponibilité, système réparti de base de données, ...



# Phirio

---

## Installer et configurer le SGBD NoSQL Apache Cassandra

---

Installation et configuration

Prérequis. Plateformes supportées. Etude du fichier de configuration : conf/cassandra.yaml  
Répertoire de travail, de stockage des données, gestion de la mémoire.

Atelier : démarrage d'un noeud et test de l'interface cliente cqlsh.

---

## Appréhender le CQL (Cassandra Query Language)

---

Commandes de base : connexion au système de base de données, création de colonnes, insertion, modification recherche,  
Le CQL : Cassandra Query Language.  
Limitations du CQL.

---

## Créer une base de données et manipuler ses objets

---

Utilisation de Cassandra

Création de bases et interrogation avec cql

Définition de la notion de consistance. Eléments en jeu : Commit.log, Memtable, Quorum

Comment écrire des requêtes ? Approches.

Atelier : premiers pas avec une base de données Cassandra pré-chargée mise à disposition sur l'infrastructure de travaux pratiques

---

## Connaitre la notion de grappe au sein de la base de données

---

Gestion de la grappe.

Principe. Configuration des noeuds.

Notion de bootstrapping et de token.

Paramètres de démarrage des noeuds.

Réplication: topologie du réseau et EndpointSnitch.

Stratégie de réplication.

Méthode d'ajout de noeuds et suppression.

Architecture de stockage mémoire et disque dur, gestion des tombstones, bloom-filter

Atelier : mise en place d'une configuration de production (multi-datacenters, multi-racks)

---

## Administrer et sécuriser un cluster Cassandra

---

Exploitation.

Gestion des noeuds Cassandra.

Sauvegardes, snapshots et export au format JSON.

Principe de cohérence, hinted\_handoff, digest request et read repair.

Sécurité

Atelier : paramétrage, authentification et sécurisation de la base system\_auth.

Gestion des rôles et des autorisations sur une application standard.



# Phirio

---

## Support Hadoop et Spark

---

Principe de map/reduce. Implémentation Hadoop et intégration Hadoop/Cassandra.  
Support Spark :  
Description rapide de l'architecture spark.

Atelier : Mise en oeuvre depuis Cassandra. Execution d'application Spark s'appuyant sur une grappe Cassandra.

---

## Supervision et performances

---

Prometheus: apports et particularité de prometheus pour la supervision cassandra  
Supervision avec nodetool.  
Principe des accès JMX , exports JMX vers des outils de supervision.

Atelier : démonstration avec Prométheus et Grafana.

Performance :  
Présentation de l'outil de test de performance Cassandra-stress

Atelier : mise en place d'un plan de stress et paramétrage.

---

## Elastic Stack

---

Elastic Stack

---

## ElasticStack

---

Présentation, fonctionnalités, licence  
Positionnement Elasticsearch et les produits complémentaires : Kibana,X-Pack,  
Logstash, Beats  
Principe : base technique Lucene et apports d'ElasticSearch  
Définitions et techniques d'indexation

---

## Installation de base

---

Prérequis techniques.  
Installation avec les RPM

---

## Outils d'interrogation

---

Communication en RESTful avec le cluster  
Interface http DevTools, travaux pratiques, démonstration



# Phirio

---

## Traitement des données

Structure des données. stockage, indexation  
Format des données.  
Conversion au format JSON des données à traiter.  
Interrogations avec Search Lite et avec Query DSL (domain-specific language)  
Notion de 'filtre' pour affiner des requêtes.

---

## Autres composants

Démonstrations de Logstash, Kibana et Beats  
Intégration

---

## ElasticStack

Présentation de la pile elastic.  
Positionnement d'Elasticsearch et des produits complémentaires : Kibana, Logstash, Beats, X-Pack  
Principe : base technique Lucene et apports d'ElasticSearch. Fonctionnement distribué

---

## Installation et configuration

Prérequis techniques.  
Installation depuis les RPM.  
Premiers pas dans la console Devtools.  
Etude du fichier : elasticsearch.yml et kibana.yml  
Mise en place de la surveillance d'un cluster ES

---

## Clustering

Définitions : cluster, noeud, sharding  
Nature distribuée d'elasticsearch  
Présentation des fonctionnalités : stockage distribué, calculs distribués avec Elasticsearch, tolérance aux pannes.

---

## Fonctionnement

Notion de noeud maître,  
stockage des documents, shard primaire et réplicat,  
routage interne des requêtes.

---

## Gestion du cluster

Outils d'interrogation : `/_cluster/health`  
Création d'un index : définition des espaces de stockage (shard), allocation à un noeud  
Configuration de nouveaux noeuds : tolérance aux pannes matérielles et répartition du stockage



# Phirio

---

## Cas d'une panne

Fonctionnement en cas de perte d'un noeud :  
élection d'un nouveau noeud maître si nécessaire, déclaration de nouveaux shards primaires

---

## Exploitation

Gestion des logs : ES\_HOME/logs  
Paramétrage de différents niveaux de logs : INFO, DEBUG, TRACE  
Suivi des performances.  
Sauvegardes avec l'API snapshot.

---

## HBase

HBase

---

## Hadoop

Rappels rapides sur l'écosystème Hadoop. Fonctionnalités.  
Le projet et les modules : Hadoop Common, HDFS, YARN, Spark, MapReduce  
Présentation HBase. Historique. Lien avec HDFS.

---

## Comprendre l'architecture et le fonctionnement de HBase

Définitions : table, région, ligne, famille de colonnes, cellules, espace de nommage, ...  
Fonctionnalités : failover automatique, sharding, requêtage  
HBase master node, Region Master, liens avec les clients HBase. Haute disponibilité. Consistance des données.  
Présentation du rôle de Zookeeper.

Atelier : définition d'une architecture HBase en fonction de contraintes d'utilisation

---

## Identifier les apports d'HBase en termes de stockage distribué des données

Format des données dans HBase. Comparaison avec d'autres bases clés/valeurs.  
Présentation des différentes interfaces disponibles.  
Outils HBase : hbase pe et hbase ltt pour les performances, hbase shell pour l'exploitation

Atelier : gestion de base avec hbase shell.



# Phirio

---

## Mener à bien l'installation

Choix des paquets. Vérification des pré-requis.  
Installation et configuration en mode distribué. Mise en oeuvre avec HDFS dans un environnement distribué.  
Test de connexion avec hbase shell.

Atelier : installation d'une grappe de serveurs HBase en mode distribué

Atelier : interrogations depuis le serveur http intégré.

---

## Savoir mettre en place une configuration distribuée

Fonctionnement en mode distribué  
Fonctionnement indépendant des démons (HMaster, HRegionServer, Zookeeper). Gestion de la consistance. Mise en évidence.

Atelier : utilisation des outils d'exploitation : hbck, hfile, ...

Atelier : mise en oeuvre des splits sur un exemple de tables réparties.  
regionsplitter.

---

## Hadoop Cloudera

Hadoop Cloudera

---

## Hadoop Cloudera

Les fonctionnalités du framework Hadoop. Les différentes versions.  
Distributions : Apache, Cloudera, Hortonworks, EMR, MapR, DSE.  
Spécificités de chaque distribution.  
Architecture et principe de fonctionnement.  
Terminologie : NameNode, DataNode, ResourceManager, NodeManager. Rôle des différents composants. Le projet et les modules : Hadoop Common, HDFS, YARN, Spark, MapReduce, Hue, Oozie, Hive, HBase, Zeppelin, ...

---

## Les outils Hadoop

Infrastructure/mise en oeuvre : Avro, Ambari, Zookeeper, Tez, Oozie. Vue d'ensemble. Gestion des données. Exemple de sqoop.  
Restitution : webhdfs, hive, Hawq, Mahout, Elasticsearch, ...  
Outils complémentaires de traitement : Spark, SparkQL, Spark/ML, Storm, BigTop, Zebra; de développement : Cascading, Scalding, Flink; d'analyse : RHadoop, Hama, Chukwa, kafka



# Phirio

---

## Installation et configuration

Présentation de Cloudera Manager.  
Installation en mode distribué.  
Configuration de l'environnement, étude des fichiers de configuration : core-site.xml, hdfs-site.xml, mapred-site.xml, yarn-site.xml et capacity-scheduler.xml  
Création des utilisateurs pour les daemons hdfs et yarn, droits d'accès sur les exécutables et répertoires.  
Lancement des services. Démarrage des composants : hdfs, hadoop-daemon, yarn-daemon, ...  
Gestion de la grappe, différentes méthodes : ligne de commandes, API Rest, serveur http intégré, APIs natives  
Exemples en ligne de commandes avec hdfs, yarn, mapred. Présentation des fonctions offertes par le serveur http

Atelier : organisation et configuration d'une grappe hadoop avec Cloudera Manager

Traitement de données. Requêtage SQL avec Hive et Impala.

---

## Administration Hadoop

Outils complémentaires à yarn et hdfs : jConsole, jconsole yarn. Exemples sur le suivi de charges, l'analyse des journaux.  
Principe de gestion des noeuds.  
Principe des accès JMX. Démonstration avec Prométheus.  
Administration HDFS : présentation des outils de stockage des fichiers, fsck, dfsadmin  
Mise en oeuvre sur des exemples simples de récupération de fichiers. Gestion centralisée de caches avec Cacheadmin.  
Gestion de la file d'attente, paramétrage, capacity-scheduler.

---

## Haute disponibilité

Mise en place de la haute disponibilité sur une distribution Cloudera.

Atelier : passage d'un système HDFS en mode HA

Explication d'une fédération de cluster Hadoop. Intérêts.

---

## Sécurité

Mécanismes de sécurité et mise en oeuvre pratique de la sécurité avec Kerberos.

Atelier : mise en place de la sécurité Kerberos sur une distribution Cloudera. Création des utilisateurs. Travaux sur les droits d'accès et les droits d'exécution. Impact au niveau des files Yarn.

---

## Exploitation

Installation d'une grappe Hadoop. Lancement des services. Principe de la supervision des éléments par le NodeManager.  
Monitoring graphique avec Cloudera Manager.

Atelier : Visualisation des alertes en cas d'indisponibilité d'un noeud.

Configuration des logs avec log4j.



# Phirio

---

## Supervision : définitions

---

Les objectifs de la supervision, les techniques disponibles. La supervision d'une ferme BigData.  
Objets supervisés. Les services et ressources. Protocoles d'accès. Exporteurs distribués de données.  
Définition des ressources à surveiller. Journaux et métriques.  
Application aux fermes BigData : Hadoop, Cassandra, HBase, MongoDB

---

## Mise en oeuvre

---

Besoin de base de données avec agents distribués, de stockage temporel (timeseriesDB)  
Produits : Prometheus, Graphite, Influxdb, ElasticSearch.  
Présentation, architectures.  
Les sur-couches : Kibana, Grafana.

---

## Graphite

---

Composants, architecture  
Modèle de données et mesures  
Format des données stockées, notion de timestamp  
Calculs de l'espace disque nécessaire  
Architecture de production.

---

## InfluxDB

---

Présentation, structure, évolution, installation  
Bucket, token, organisation  
Plugin Telegraph, architecture  
Interface graphique, alertes, langage flux  
Démonstration avec Jolokia2 et Cassandra.

---

## JMX

---

Principe des accès JMX. MBeans. Visualisation avec jconsole et jmxterm.  
Suivi des performances cassandra : débit d'entrées/sorties, charges, volumes de données, tables, ...

---

## Prometheus

---

Installation et configuration de base  
Définition des ressources supervisées, des intervalles de collecte  
Types de mesures : compteurs, jauges, histogrammes, résumés.  
Notions d'instances, de jobs.  
Démarrage du serveur Prometheus  
Premiers pas dans la console web, et l'interface graphique.  
Le langage PromQL  
Node Exporter. JMX Exporteur. MongoDB Exporteur.  
Démonstration avec Cassandra  
Configuration des agents sur les noeuds de calculs. Agrégation des données JMX. Expressions régulières.  
Requêtage. Visualisation des données  
Comparaison avec Graphite et InfluxDB.



# Phirio

---

## Exploration et visualisation des données

---

Mise en oeuvre de Grafana. Installation, configuration.  
Pose de filtres sur Prometheus et remontée des données.  
Etude des différents types de graphiques disponibles,  
Agrégation de données. Appairage des données entre Prométheus et Grafana.  
Visualisation et sauvegarde de graphiques,  
création de tableaux de bord à partir des graphiques.

---

## Kibana, installation et configuration

---

Architectures, paramétrages  
Installation, configuration du mapping avec Elasticsearch.  
Mapping automatique ou manuel  
Démonstration avec Cassandra  
Injection des données avec Logstash, Filebeat et Metricbeat.  
Configuration des indexes  
Exploration des données, création de graphiques, de tableaux de bord

---

## Le scénario

---

La société DataIA rencontre de sérieuses difficultés avec sa toute nouvelle infrastructure BigData : le traitement des données est très lent, ils n'arrivent à aucun résultat dans un temps correct.  
Ils décident de se faire aider par une équipe d'administrateurs pour analyser et surveiller les différents composants de leur architecture afin d'en trouver les failles et de proposer des correctifs.  
Cette équipe saura-t-elle relever le défi ?

---

## La méthode

---

Simulation d'un cas d'étude, avec un travail collaboratif sur des données réelles, accessibles en opendata, et des labs techniques (Prometheus, Graphite, Influxdb, ElasticSearch, ..)  
Épreuves personnelles et épreuves en commun vont permettre de contrôler les connaissances et d'échanger entre stagiaires, tout en bénéficiant du soutien et des explications complémentaires du formateur sur les thèmes proposés

---

## Les jeux

---

Battle d'architecture, la techno mystère, l'intrus, les points de faiblesse, etc...

---

## Le debrief

---

Retour des travaux, bilan des points individuels et classement des joueurs.  
Retour d'expérience des participants