Cycle certifiant Architecte BigData

CB090

Durée: 18 jours

9 110 €

Public :

Chefs de projet,architectes

Objectifs :

Comprendre les concepts et les apports des technologies BigData. Connaître les caractéristiques techniques des bases de données NoSQL, les différentes solutions disponibles. Identifier les critères de choix. Connaître les apports des outils comme Hadoop, Spark, le rôle des différents composants, la mise en oeuvre pour du stream processing, des traitements de Machine Learning, des calculs statistiques avec les graphes.

Connaissances préalables nécessaires :

Connaissance des bases des systèmes d'information, et notions de calculs statistiques.

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.

Atelier : Démonstration d'un prédiction Machine Learning avec Dataiku DSS

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

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, ...

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.

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.

HBase

HBase

HBase

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.

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 : utilisationn des outils d'exploitation : hbck, hfile, ...

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

Neo4J

Neo4J

Neo4J

Présentation Neo4j, les différentes éditions, license
Fonctionnalités, stockage des données sous forme de graphes
CQL : Cypher Query Language
Positionnement par rapport aux autres bases de données, apports de Neo4j
L'analyse de données.
Cas d'usage

Installation et configuration

Les différentes méthodes d'installation

Atelier : installation de Neo4J Enterprise Edition en cluster.

Premiers pas avec l'interface web.
Création de données, requêtage
Import de données

Cypher Query Language

Syntaxe, description des relations avec CQL,les patterns
Les clauses d'écriture : set, delete, remove, foreach,
de lecture : match, optional match, where, count, case, ...
Les fonctions : count, type, relationship, ...
Principe de profondeur et de direction de relation dans une recherche
Les listes et les projections maps
Les algorithmes de Graphe

Atelier : création d'un graphe,

Requêtes de recherche, navigation dans le graphe

Exploitation

Sauvegardes et restaurations
Optimisation des transactions
Indexation
Client jmx
Points de surveillance

Développement

Description des APIs disponibles: .Net, Java, Javascript, Python
Connexions, sessions et transactions
Principe de causalité entre transactions
La bibliothèque Apoc

Atelier : connexion et récupération de données provenant de Cassandra

Sécurité

Principe et activation
Paramétrage

Atelier : création d'un compte sécurisé

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

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

Spark

Spark

Spark

Présentation Spark, origine du projet,apports, principe de fonctionnement. Langages supportés.
Modes de fonctionnement : batch/Streaming.
Bibliothèques : Machine Learning, IA
Mise en oeuvre sur une architecture distribuée. Architecture : clusterManager, driver, worker, ...
Architecture : SparkContext, SparkSession, Cluster Manager, Executor sur chaque noeud. Définitions : Driver program, Cluster manager, deploy mode, Executor, Task, Job

Savoir intégrer Spark dans un environnement Hadoop

Intégration de Spark avec HDFS, HBase,
Création et exploitation d'un cluster Spark/YARN. Intégration de données sqoop, kafka, flume vers une architecture Hadoop et traitements par Spark.
Intégration de données AWS S3.
Différents cluster managers : Spark interne, avec Mesos, avec Yarn, avec Amazon EC2

Atelier : Mise en oeuvre avec Spark sur Hadoop HDFS et Yarn. Soumission de jobs, supervision depuis l'interface web

Développer des applications d’analyse en temps réel avec Spark Structured Streaming

Objectifs , principe de fonctionnement: stream processing. Source de données : HDFS, Flume, Kafka, ...
Notion de StreamingContext, DStreams, démonstrations.

Atelier : traitement de flux DStreams en Scala. Watermarking. Gestion des micro-batches.

Intégration de Spark Structured Streaming avec Kafka

Atelier : mise en oeuvre d'une chaîne de gestion de données en flux tendu : IoT, Kafka, Spark Structured Streaming, Spark. Analyse des données au fil de l'eau.

Faire de la programmation parallèle avec Spark sur un cluster

Utilisation du shell Spark avec Scala ou Python. Modes de fonctionnement. Interprété, compilé.
Utilisation des outils de construction. Gestion des versions de bibliothèques.

Atelier : Mise en pratique en Java, Scala et Python. Notion de contexte Spark. Extension aux sessions Spark.

Manipuler des données avec Spark SQL

Spark et SQL
Traitement de données structurées. L'API Dataset et DataFrames
Jointures. Filtrage de données, enrichissement. Calculs distribués de base. Introduction aux traitements de données avec map/reduce.
Lecture/écriture de données : Texte, JSon, Parquet, HDFS, fichiers séquentiels.
Optimisation des requêtes. Mise en oeuvre des Dataframes et DataSet. Compatibilité Hive

Atelier : écriture d'un ETL entre HDFS et HBase

Atelier : extraction, modification de données dans une base distribuée.
Collections de données distribuées. Exemples.

Support Cassandra

Description rapide de l'architecture Cassandra. Mise en oeuvre depuis Spark. Exécution de travaux Spark s'appuyant sur une grappe Cassandra.

Spark GraphX

Fourniture d'algorithmes, d'opérateurs simples pour des calculs statistiques sur les graphes

Atelier : exemples d'opérations sur les graphes.

Avoir une première approche du Machine Learning

Machine Learning avec Spark, algorithmes standards supervisés et non-supervisés (RandomForest, LogisticRegression, KMeans, ...)
Gestion de la persistance, statistiques.
Mise en oeuvre avec les DataFrames.

Atelier : mise en oeuvre d'une régression logistique sur Spark

Kafka

Kafka

Kafka

Le projet Kafka : historique, fonctionnalités, principe de fonctionnement.
Présentation de l'architecture et du rôle de chaque composant : broker, producer, consumer
Liaison éventuelle avec Zookeeper. Impacts.

Acquérir les bonnes pratiques de distribution de messages

Etude de la configuration du broker

Atelier : création d'une configuration multi-broker, démarrage de plusieurs noeuds

Atelier : création d'un topic simple et mise en oeuvre d'une chaîne de base. Visualisation des messages avec kafka-console-consumer

Savoir configurer Kafka pour intégrer les données de différents formats et de sources différentes

Kafka Connect : présentation des fonctionnalités : intégration de données d'origines multiples, modes de fonctionnement (standalone ou distribué)
Types de connecteurs

Atelier : configuration de connecteurs, ingestion de données, création d'une chaîne de transformation

Appréhender les différentes APIs de Kafka.

Conception d'applications avec Kafka. Principe de fonctionnement.

Atelier : développement de prototypes en Python, Java, Scala

Couplage avec SparkStreaming en mode batch, en mode continu
Principe et architecture de Kafka Streams

Mettre en oeuvre KSQL

Conception d'application avec KSQL. La sur-couche KSQL.
Présentation de ksqlDB. Création de flux en KSQL. Les ajouts SQL pour permettre le traitement en flux.
Micro-batchs. Introduction au water-marking.

Atelier : mise en oeuvre d'une chaîne de traitement avec KSQL

Travailler en sécurité avec Kafka

Intégrité et confidentialité :
Chiffrement SSL et authentification SASL.
Sécurisation de l'infrastructure : Zookeeper, proxy Rest
Disponibilité : La réplication. Facteur de réplication. Partitions

Atelier : tests de haute disponibilité dans une configuration multi-noeuds

Exploitation

Mise en oeuvre de kafka-tool
Gestion des logs
Détection de pannes : méthodes et outils
Benchmarks

Intégration SQL

Intégration SQL

Intégration SQL

Besoin. Adéquation entre les objectifs et les outils.
Faciliter la manipulation de gros volumes de données en conservant une approche utilisateurs.
Rappels sur le stockage : HDFS, Cassandra, HBase
et les formats de données : parquet, orc, raw, clés/valeurs
Les outils : Hive, Impala, Tez, Presto, Drill, Phoenix, Spark-sql, Spark Dataframe

Hive

Présentation Hive. Mode de fonctionnement. Rappel sur map/reduce.
Hive : le langage HiveQL. La surcouche Tez.

Atelier : création de tables, requêtage, connexion avec Hbase.

Impala et Phoenix

Présentation Impala. Cadre d'utilisation. Contraintes. Liaison avec le métastore Hive.

Atelier : mise en évidence des performances.

Présentation Phoenix. Cadre d'utilisation. Contraintes.

Atelier : connexion et requêtage sur une table Hbase.

Presto

Cadre d'utilisation. Sources de données utilisables.

Atelier : mise en oeuvre d'une requête s'appuyant sur Cassandra et PostgreSQL.

Spark-sql et Spark DataFrame

Les différentes approches. Syntaxe Spark-sql, Spark/SQl. APIs QL.
Utilisation du métastore Hive.

Atelier : mise en oeuvre d'une requête s'appusant sur une table HBase et sur HDFS. Requêtage en spark-sql sur un fichier csv.

Drill

Utilisation d'APIs JDBC, ODBC. Indépendance Hadoop. Contraintes d'utilisation. Performances.

Atelier : lecture de fichiers Parquets dans du HDFS, jointures, connexion et requêtage sur une table Hbase.

Comparatifs

Compatibilité ANSI/SQL. Approches des différents produits.
Critères de choix.

Le scénario

Afin de déterminer quelles sont les espèces végétales adaptées à chaque zone géographique, des ingénieurs BigData d'un laboratoire de recherche doivent mettre en place un process de collecte et stockage de photos satellites et mettre à disposition des data-scientists du laboratoire une architecture adéquate pour les analyses.

La méthode

Simulation d'un cas d'étude, avec un travail collaboratif sur des données réelles, accessibles en opendata, et des labs techniques (cluster cassandra, hadoop, spark, elasticstack, etc ..)
Épreuves personnelles et épreuves en commun vont permettre de valider les connaissances et d'échanger entre participants, 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