Pythagore F.D.
Apprendre à apprendre

CB090

Cycle Certifiant architecte BigData

Durée: 10 jours

Prix et dates: nous consulter

Public:
Chefs de projet,architectes

Objectifs:
Comprendre les concepts et les apports des technologies BigData. 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:
Introduction : 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.
Quelques éléments d'architecture.Concepts clés : ETL, Extract Transform Load, CAP, 3V, 4V, données non structurées, prédictif, Machine Learning.
Quelques applications : Watson (IBM), Amazon Rekognition
L'écosystème du BigData : les acteurs, les produits, état de l'art.
Cycle de vie des projets BigData. Emergence de nouveaux métiers : Datascientists, Data labs, Hadoop scientists, CDO, ...
Rôle de la DSI dans la démarche BigData. Gouvernance des données :importance de la qualité des données, fiabilité, durée de validité, sécurité des données
Aspects législatifs : sur le stockage, la conservation de données, etc ...sur les traitements, la commercialisation des données, des résultats

Pythagore F.D.

Apprendre à apprendre

CB090

Cycle Certifiant architecte BigData

Stockage distribué : Caractéristiques NoSQL. 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 : Définitions : réplication, sharding, gossip protocol, hachage,
Systèmes de fichiers distribués : GFS, HDFS, Ceph
Les bases de données : Cassandra, DynamoDB, Accumulo, HBase, MongoDB, CouchBase, Riak, BigTable, ..
Caractéristiques NoSQL :structure de données proches des utilisateurs, développeurs
Les types de bases de données : clé/valeur, document, colonne, graphe.
Données structurées et non structurées, documents, images, fichiers XML, JSON, CSV, ...
Les différents modes et formats de stockage.Stockage réparti : réplication, sharding, gossip protocol, hachage,
Systèmes de fichiers distribués : GFS, HDFS,Quelques exemples de produits et leurs caractéristiques : Cassandra, MongoDB, CouchDB, DynamoDB, Riak, Hadoop, HBase, BigTable, ...
Qualité des données, gouvernance de données.

Indexation et recherche : Moteurs de recherche.Principe de fonctionnement.
Méthodes d'indexation. Mise en oeuvre avec elasticsearch.
Exemple de Lucene/solr.
Recherche dans les bases de volumes importants.
Exemples de produits et comparaison : Dremel, Drill, ElasticSearch, MapReduce,

Pythagore F.D.

Apprendre à apprendre

CB090

Cycle Certifiant architecte BigData

Calcul et restitution, intégration : Différentes solutions : calculs en mode batch, ou en temps réel,
sur des flux de données ou des données statiques.
Les produits : langage de calculs statistiques, R Statistics Language, sas, RStudio;
outils de visualisation : Tableau, QlikView
Ponts entre les outils statistiques et les bases BigData
Outils de calcul sur des volumes importants : storm en temps réel, hadoop en mode batch.
Zoom sur Hadoop : complémentarité de HDFS et MapReduce.
Restitution et analyse : logstash, kibana, elk, pentaho
Présentation de pig pour la conception de tâches MapReduce sur une grappe Hadoop.

Hadoop : 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
Oozie, Pig, Hive, HBase, ...

Les outils Hadoop : Infrastructure/Mise en oeuvre :
Avro, Ambari, Zookeeper, Pig, Tez, Oozie, Falcon, Pentaho
Vue d'ensemble
Gestion des données.
Exemple de sqoop.
Restitution : webhdfs, hive, Hawq, Mahout,ElasticSearch ..
Outils complémentaires:
Spark, SparkQL, SparkMLib, Storm, BigTop, Zebra
de développement : Cascading, Scalding, Flink/Pachyderm
d'analyse : RHadoop, Hama, Chukwa, kafka

Pythagore F.D.
Apprendre à apprendre

CB090

Cycle Certifiant architecte BigData

Installation et configuration : Trois modes d'installation : local, pseudo-distribué, distribué
Première installation.
Mise en oeuvre avec un seul noeud Hadoop.
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 users 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, etc ..
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
Travaux pratiques :
Organisation et configuration d'une grappe hadoop

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, accès JMX.
Travaux pratiques :
mise en oeuvre d'un client JMX
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
Déplacement d'un NameNode. Mise en mode maintenance.

Haute disponibilité : Mise en place de la haute disponibilité sur une distribution Ambari.
Travaux pratiques :
Passage d'un système HDFS en mode HA

Pythagore F.D.
Apprendre à apprendre

CB090

Cycle Certifiant architecte BigData

Sécurité : Mécanismes de sécurité et mise en oeuvre pratique :
Activation de la sécurité avec Kerberos dans core-site.xml, et dans hdfs-site.xml pour les NameNode et DataNode. Sécurisation de yarn avec la mise en oeuvre d'un proxy et d'un Linux Container Executor.
Travaux pratiques :
Mise en place de la sécurité Kerberos sur une distribution Ambari. Création des utilisateurs. Travaux sur les droits d'accès et les droits d'exécution. Impact au niveau des files Yarn, Oozie et Tez.

Exploitation : Installation d'une grappe Hadoop avec Ambari. Tableau de bord. Lancement des services.
Principe de la supervision des éléments par le NodeManager.
Monitoring graphique avec Ambari.
Présentation de Ganglia,Kibana
Travaux pratiques :
Visualisation des alertes en cas d'indisponibilité d'un noeud.
Configuration des logs avec log4j.

NoSQL avec Cassandra : Historique, fonctionnalités de Cassandra, licence
Format des données,"key-value", traitement de volumes importants,
haute disponibilité, système réparti de base de données, ...

Installation et configuration : Prérequis. Plate-formes supportées. Etude du fichier de configuration : conf/cassandra.yaml
Répertoire de travail, de stockage des données, gestion de la mémoire.
Démarrage d'un noeud et test de l'interface cliente cqlsh.

CQL : Commandes de base : connexion au système de base de données,
création de colonnes,insertion, modification recherche,
Le CQL : Cassandra Query Language. Exécution de scripts.
Comment écrire des requêtes? Approches.

Pythagore F.D.
Apprendre à apprendre

CB090

Cycle Certifiant architecte BigData

Gestion de la grappe	: Principe.Préparation du premier noeud : adresse d'écoute. Configuration de nouveaux noeuds.Notion de bootstrapping et de token. Paramètres listen_address et rpc_address. Réplication : topologie du réseau et EndpointSnitch.Stratégie de réplication. Ajout de noeuds, suppression. Cassandra dans un cloud. Mise en oeuvre avec OpenStack.
Supervision	: OpsCenter : installation, lancement. Utilisation de base. Supervision avec nodetool cfstats, ou export JMX vers des outils de supervision comme Nagios.
Exploitation	: Sauvegardes. Import/export au format JSON.
Support Hadoop	: Principe de MapReduce. Implémentation Hadoop. Mise en oeuvre depuis Cassandra.
Support Spark	: Description rapide de l'architecture spark. Mise en oeuvre depuis Cassandra. Execution de travaux Spark s'appuyant sur une grappe Cassandra.
Flux de données avec Storm	: Présentation de Storm:fonctionnalités, architecture, langages supportés Définitions:spout, bolt, topology
Architecture	: Etude des composants d'un cluster Storm : master node 'nimbus' et worker nodes Positionnement par rapport à un cluster Hadoop.Le modèle de données. Différents types de flux.
Premiers pas	: Configuration d'un environnement de développement. Installation d'un cluster Storm. Travaux pratiques sur le projet storm-starter
Flux de données	: Définition du nombre de flux dans un noeud, création de topologies regroupants des flux entre différents noeuds, communication entre flux en JSON, lecture de flux d'origines diverses (JMS, Kafka, ...)

Pythagore F.D.

Apprendre à apprendre

CB090

Cycle Certifiant architecte BigData

Haute disponibilité	: Tolérance aux pannes: principe de fiabilisation des master node, workers node, nimbus Garantie de traitement des flux: principe,paramètres TOPOLOGY_MESSAGE_TIMEOUT_SECS, TOPOLOGY_ACKERS Traitements temps réel avec Trident. Scalabilité: parallèlisme dans un cluster storm, ajouts de noeuds, commande 'storm rebalance'