Phirio

Hadoop : administration cloudera

CB032

Durée: 3 jours

2 400 €

4 au 6 février
15 au 17 juin

2 au 4 septembre
30 novembre au 2 décembre

Public :

Chefs de projet, administrateurs et toute personne souhaitant mettre en oeuvre un système distribué avec Hadoop. Les travaux pratiques sont réalisés sur une distribution Hadoop Cloudera.

Objectifs :

Connaître les principes du framework Hadoop et savoir l'installer et le configurer. Maitriser la configuration et la gestion des services avec Cloudera Manager.

Connaissances préalables nécessaires :

Connaissance des commandes des systèmes unix/linux.

Programme :

Introduction

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

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.