



Hadoop Hortonworks : administration avec Ambari

CB031

Durée: 3 jours

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 Hortonworks.

Objectifs :

Connaître les principes du framework Hadoop et savoir l'installer, le configurer et l'administrer avec Ambari (tableaux de bord, supervision, gestion des services, etc ...)

Connaissances préalables nécessaires :

Connaissance des commandes des systèmes unix/linux et des bases TCP/IP

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, Oozie, Tez, Pig, Hive, HBase, ...

Les outils Hadoop

Infrastructure/Mise en oeuvre : Avro, Ambari, Zookeeper, Pig, 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

Installation en mode distribué. Configuration du cluster et 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, ...
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

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. Déplacement d'un NameNode. Mise en mode maintenance.
Gestion de la file d'attente, paramétrage, capacity-scheduler.

Haute disponibilité

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

Atelier : passage d'un système HDFS en mode HA

Explication/démonstration d'une fédération de cluster Hadoop.

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 Ambari. Création des utilisateurs. Travaux sur les droits d'accès et les droits d'exécution. Impact au niveau des files de traitement Yarn.

Sécurisation de yarn avec les Linux Container Executor.

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.

Atelier : visualisation des alertes en cas d'indisponibilité d'un noeud.
Configuration des logs avec log4j.