

# BigData : intégration SQL, hive, SparkDataframes

Référence : PYCB045

Durée : 2 jours

Certification : **Aucune**

## CONNAISSANCES PREALABLES

- Connaissance générale des systèmes d'informations et des bases de données.

## PROFIL DES STAGIAIRES

- Experts en bases de données, chefs de projet. • Toute personne souhaitant comprendre le fonctionnement et les apports des bases NoSQL.

## OBJECTIFS

- Connaître les caractéristiques techniques des bases de données NoSQL, les différentes solutions disponibles. • Identifier les critères de choix.

## CERTIFICATION PREPAREE

Aucune

## METHODES PEDAGOGIQUES

- Mise à disposition d'un poste de travail par stagiaire
- Remise d'une documentation pédagogique papier ou numérique pendant le stage
- La formation est constituée d'apports théoriques, d'exercices pratiques, de réflexions et de retours d'expérience
- Le suivi de cette formation donne lieu à la signature d'une feuille d'émargement

## FORMATEUR

Consultant-Formateur expert Bigdata

## METHODE D'EVALUATION DES ACQUIS

- Auto-évaluation des acquis par le stagiaire via un questionnaire
- Attestation de fin de stage adressée avec la facture

## CONTENU DU COURS

### Présentation

- Besoin
- Comment concilier le quoi et le comment ?
- 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, Pig, Spark/QL

### Hive et Pig

- Présentation
- Mode de fonctionnement
- Rappel sur map/reduce
- Hive : le langage HiveQL. Exemples
- Pig : le langage pig/latin. Exemples

### Impala

- Présentation

- Cadre d'utilisation
- Contraintes
- Liaison avec le métastore Hive
- Travail Pratique : Mise en évidence des performances

### Presto

- Cadre d'utilisation
- Sources de données utilisables
- Travail Pratique : Mise en oeuvre d'une requête s'appuyant sur Cassandra, HDFS et PostgreSQL

### Spark DataFrame

- Les différentes approches
- Syntaxe Spark/SQL
- APIs QL
- Compilation catalyst
- Syntaxe, opérateurs
- Comparaison avec Presto

- Travail Pratique : Mise en oeuvre d'une requête s'appuyant sur HBase et HDFS

### **Drill/sqoop**

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

### **Comparatifs**

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