



# Phirio

## BigData : intégration SQL, Hive, SparkDataFrames

CB045

Durée: 2 jours

1 570 €

6 au 7 février

22 au 23 mai

11 au 12 septembre

27 au 28 novembre

### Public :

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

### Objectifs :

Comprendre les connexions existantes entre les mondes relationnels et NoSQL en environnement Big Data. Savoir mettre en oeuvre Hive, Impala, Phoenix, les Spark Dataframes.

### Connaissances préalables nécessaires :

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

### Programme :

#### Présentation

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.

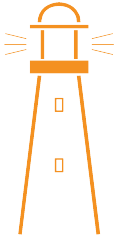
#### Impala et Phoenix

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.



# Phirio

---

## 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/scala, pyspark. APIs QL. Utilisation du métastore Hive.

Atelier : mise en oeuvre d'une requête s'appuyant sur une table HBase et sur HDFS. Requête 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 requête sur une table Hbase.

---

## Comparatifs

---

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