

Durée : 2 jour(s)

Objectifs

Comprendre les concepts du BigData et connaître les technologies associées.

Pré-requis

PUBLIC : Chefs de projets, architectes, et toute personne souhaitant connaître les outils et solutions pour mettre en place une architecture BigData.

PREREQUIS : bonne culture générale sur les systèmes d'information.

Plan de cours

1. IntroductionLe besoin : volumes importants de données, traitements optimisés de flux de données au fil de l'eau (nouvelles technologies et nouveaux usages). Domaines concernés : recherche scientifique, médical, e-commerce, sécurité, ... Développement des techniques sur différents aspects : stockage, indexation/recherche, calcul Définition de ETL : Extract Transform Load Les acteurs aujourd'hui

2. Stockage

Caractéristiques NoSQL : Structure de données proches des utilisateurs, développeurs 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, BigTable, ...

Les bases de données

Quelques exemples de produits et leurs caractéristiques : cassandra, MongoDB, CouchDB, DynamoDB...

3. Indexation et recherche

Moteurs de recherche

Principe de fonctionnement

Méthodes d'indexation

Exemple de Lucene, et mise en oeuvre avec solr

Recherche dans les bases de volumes importants

Exemples de produits et comparaison : dremel, drill, elasticsearch, MapReduce...

4. 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

Outils de calcul sur des volumes importants : storm en temps réel, hadoop en mode batch.

Zoom sur Hadoop : complémentarité de HDFS et MapReduce.

5. Evolutions

Les offres Saas BigData, type Google BigQuery.

Les limites actuelles. Les avancées annoncées.