ESMT
École Supérieure Multinationale des Télécommunications

Formation continue

Formation Traitement distribué pour les Big Data

APERÇU / PLAN DU COURS

DESCRIPTION DU COURS

Brève description : Les entreprises produisent d'énormes quantités de données chaque jour. Ces données sont stockées puis traitées et analysées pour en tirer de la valeur. Grâce aux plateformes de stockage et de traitements distribués de type Hadoop, il est devenu plus facile pour les ingénieurs de répondre aux problématiques du Big Data avec une grande efficacité et à un coût réduit.

Ce séminaire vous offre l'occasion de vous essayer au traitement distribué de données massives via la plateforme Hadoop et ses outils comme Hive et Spark.

Objectifs	Les participants seront capables de : 1. Stocker des données sur Hadoop 2. Implémenter en Python des algorithmes MapReduce et les exécuter sur une plateforme Hadoop 3. Charger des données sur Hive et les traiter à l'aide du langage HiveQL 4. Charger des données dans Spark et les traiter à l'aide du langage Spark SQL
Dates	**08 au 10 Juin 2021**
Durée	**3 jours**

PUBLIC CIBLE

- Développeurs
- Ingénieurs recherche et développement
- Experts en business intelligence
- Analystes de données

PRE-REQUIS

- Notions en administration système sous Linux, en programmation orientée objet et en bases de données. Connaissances de base du langage SQL.

AGENDA ET CONTENU DU COURS

Contenu Pédagogique	**Lôécosystème du Big Data** ▪ Les origines et caractéristiques du Big Data ▪ Les avancées technologiques ▪ Lôinformatique parallèle ▪ Les acteurs du Big Data **Le paradigme MapReduce** ▪ Historique ▪ Le pattern MapReduce ▪ **Travaux pratiques :** programmation MapReduce **Hadoop** ▪ Historique ▪ Socle technique ▪ HDFS ▪ Hadoop ï YARN ▪ Ecosystème Hadoop ▪ Les distributions Hadoop ▪ **Travaux pratiques** : programmation Hadoop MapReduce **Apache Hive** ▪ Présentation ▪ Création dôune base de données sous Hive ▪ Commandes HQL ▪ **Travaux pratiques** : traitement de données avec HiveQL **Spark SQL** ▪ Apache Spark ▪ Les RDD ▪ **Travaux pratiques** : calcul distribué avec PySpark ▪ Spark SQL et DataFrames ▪ **Travaux pratiques** : traitement de données avec Spark SQL

MODES DôANIMATION PEDAGOGIQUE

- Exposés théoriques ;
- Questions-réponses ;
- Exercices dôapplication
- Travaux pratiques

SUPPORTS PEDAGOGIQUES

- Support de cours ;
- Les Travaux Pratiques se feront en local et dans le Cloud de AWS via le service Amazon Elastic Map Reduce (EMR) ;
- Le langage de programmation utilisé sera Python.