BIG DATA FACILE

Carrière et Développement Professionnel
Formation en ligne

Plan de cours

Titre de cours	**Gestion et Analyse des Données Massives**
Session	Hiver 2022 / Mars 12 – Avril 17
Durée	30h / 12 séances de 2.5h – dispensées via Zoom les Samedis et Dimanches De **7h00 à 9h30** AM heure de l'EST *(GMT -4 Toronto/Montréal)*
Instructeur	Khaled Tannir **Email:** khaled.tannir@bigdatafacile.com
Description du cours	Ce cours familiarise les participants avec les différents aspects des données massives et leur gestion à la fois sur site et dans le Cloud. L'accent est mis sur des expériences pratiques allant de l'ingestion de données à l'analyse de celles-ci, à la fois au repos et en mouvement (données en continu). Ceci tout en incluant la définition du Big Data et de ses 5 V : *Volume, Velocity, Variety, Veracity et Value*. Les architectures des écosystèmes de traitement et de stockage distribués tels que Hadoop et Spark sont couvertes, suivies d'une introduction au langage Scala et à PySpark.
Résultats d'apprentissage	À la fin de ce cours, vous devriez être en mesure de: ○ Distinguer les principales caractéristiques du Big Data (5 V : Volume, Vélocité, Variété, Véracité et Valeur). ○ Installez et exécutez la machine virtuelle incluant Hadoop et Spark ○ En local ou sur le cloud Amazon. ○ Utiliser les frameworks les plus populaires (Apache Hadoop, Apache Spark) pour ingérer, stocker, traiter, transformer et analyser de grands ensembles de données. ○ Copier / déplacer des données vers et depuis HDFS, exécuter des tâches MapReduce (principaux composants du framework Hadoop) ○ Nettoyer, filtrer, normaliser les données non structurées à l'aide de scripts Pig Latin. ○ Écrire des scripts HiveQL (Hive Query Language) pour interroger des énormes ensembles de données stockés dans HDFS.

- Utilisez HCatalog pour partager des métadonnées entre Hive et Pig et d'autres outils de l'écosystème Hadoop.
- Ingérer des données (importer / exporter des ensembles de données en mode batch) dans HDFS à l'aide de Sqoop
- Ingérer des données en temps réel à l'aide de Flume / Nifi et Spark Streaming.
- Construire des dataflows d'ingestion de données avec Nifi.
- Utiliser Apache Zeppelin pour exécuter des applications Spark de manière interactive (codée en Scala et PySpark)
- Expérimenter SparkSQL avec Scala et PySpark (à l'aide de Apache Zeppelin).
- Collecter des données en temps réel à partir de source réelle telles que Twitter / Meetup et les stockez-les dans HDFS afin de les analyser à l'aide des outils tels que Hive/Impala.

Matériel de cours

Lectures, exercices, tutoriels, ateliers et fichiers de données sont fournis via la plateforme *mescours*.

Méthodes d'Instruction

L'approche d'enseignement et d'apprentissage est basée sur l'expérience, la collaboration et la résolution de problèmes.

Avant de suivre le cours

Avant de suivre le cours, vous devez :

- Avoir une connaissance de base des commandes Shell Linux.
- Etre familier avec un langage de programmation tel que Java et Python.
- Être familier avec le langage SQL et capable d'écrire des requêtes SQL de base
- Connaître les techniques et logiciels de virtualisation (ex : VirtualBox, VMWare, …).

Dans le cas contraire, des documents vous seront fournis.

Environnement technique pour les ateliers

Vous avez besoin d'un ordinateur portable avec :

- 16 Go (recommandé) ou au moins 8 Go de mémoire
- 50+ Go d'espace libre sur le disque dur

Si votre ordinateur portable ne dispose pas de suffisamment de ressources pour exécuter des machines virtuelles, vous devriez envisager la solution cloud AWS. L'environnement virtuel du cours été mis à jour et chargé sur AWS.

- L'exécution de l'environnement technique sur Amazon EC2 nécessite un compte AWS. (non inclus dans la formation)
- Pour créer un compte sur Amazon AWS, vous devez fournir vos informations de carte de crédit. Pour plus d'informations sur la création d'un compte AWS @ https://aws.amazon.com/

- Le coût associé dépend de votre utilisation, mais la moyenne est de 0,80 $ par heure (pour le calcul et le stockage)

t2.large	2	Variable	8 GiB	EBS Only	$0.0928 per Hour
t2.xlarge	4	Variable	16 GiB	EBS Only	$0.1856 per Hour
t2.2xlarge	8	Variable	32 GiB	EBS Only	$0.3712 per Hour

Pour plus d'informations sur les tarifs AWS :
https://aws.amazon.com/ec2/pricing/on-demand

EVALUATION

Elément	%	Explication
Présence et participation active	**10%**	• Ce cours est constitué d'une communauté d'apprenants dont vous êtes membre à part entière ; votre participation active est donc essentielle à sa réussite. Cela signifie : assister aux cours ; visiter mescours, suivre les lectures et faire les exercices assignés; et participer à des discussions/activités en classe. • Vous devez être présent pour au moins 75% du temps de la formation. (2 absences autorisées).
Project de fin de formation	**90 %**	Résoudre un problème technique ou un cas d'utilisation basé sur l'apprentissage en classe. L'examen des études de cas et des problèmes du monde réel est l'un des moyens les plus efficaces d'acquérir de nouvelles compétences et d'assimiler le matériel de cours. **Vous devez obtenir 65% des points pour réussir le projet de fin de formation.**
Total	**100%**	

Un Certificat d'accomplissement de la formation vous sera remis à la fin du cours et après avoir obtenu au moins 65% des points au projet final

CONTENU DU COURS

Classe	Sujets
	1ère Partie
Classe 1	Introduction au cours, sa logistique et son déroulement ***Introduction aux Big Data*** • Qu'est-ce que le Big Data et d'où viennent les données • Comment gérer de très grandes quantités de données • Concepts de base pour stocker, traiter, extraire de la valeur et des connaissances à partir des données *L'écosystème Hadoop* • *Introduction à Hadoop* • *Historique, performances, composants de base* • *Présentation de l'écosystème Hadoop et des composants* • *Distributions Hadoop - Comparaison des meilleurs fournisseurs Hadoop* Pratique: • Installation de VirtualBox • Installation de la machine virtuelle du cours (Hadoop et Spark) • Transfert des fichiers des cours vers la VM
Classe 2	**Stockage et traitement des données à grande échelle** Présentation des composants principaux de Hadoop - HDFS / MapReduce • Qu'est-ce que HDFS et comment ça marche • Qu'est-ce que MapReduce et comment ça marche • Exécution de tâches MapReduce à l'aide de l'API Java • Exécution de tâches MapReduce à l'aide de Hadoop Streaming API Pratique: • Interagir avec HDFS à l'aide de la ligne de commande • Exécution de tâches MapReduce à l'aide de Java • Surveillance des tâches MapReduce à l'aide de YARN
Classe 3	**Traitement des données à grande échelle à l'aide de Spark** • Qu'est-ce que Spark et comment ça marche • Présentation des RDD et des DataFrames Spark • Introduction au langage Scala • Présentation d'Apache Zeppelin Pratique:

	• Lancer d'Apache Zeppelin et création de notes • Écrire des applications Spark avec Scala et PySpark
2ème Partie	
Classe 4	**Analyser des données structurées à grande échelle -1** Présentation de Hive/Impala et du langage de requête Hive • Concepts et composants principaux de Hive • Amélioration des performances avec le partitionnement et le compartimentage des données • Interrogation des données à l'aide de HiveQL (Hive & Impala) • Interrogation des données interactivement avec Impala Pratique: • Interrogation d'ensembles de données structurées à l'aide de HiveQL (Hive & Impala)
Classe 5	**Analyser des données structurées à grande échelle -2** Présentation de Spark SQL • Utilisation des Dataframes de Spark SQL pour charger et interroger des données • Création de tables virtuelles en mémoire avec l'API Spark SQL
Classe 6	**Stockage des données à grande échelle** **Format de fichier pour les grands ensembles de données** Avro / Parquet • Introduction aux formats de fichiers • Présentation du format des fichiers Hadoop : Avro et Parquet • Apprendre comment encoder les fichiers Avro / Parquet • Performances et utilisation de chacun des formats Avro / Parquet Pratique: • Spark SQL : interrogation d'ensembles de données structurés à l'aide de l'API Spark SQL et de l'API du langage SQL • Manipulation des outils Apache Avro Tools et Parquet Tools ainsi que Apache kite-sdk pour la création de fichiers Avro et Parquet

Classe 7	**Analyser des données Non Structurées à grande échelle** Introduction à PIG et le langage Pig Latin Script • Concepts de base • Présentation du langage Pig Latin et comment ça fonctionne • Nettoyer/filtrer et préparer les données à l'aide de Pig Latin • Qu'est-ce que HCatalog et comment ça fonctionne • Utilisation de HCatalog pour partager des métadonnées dans Hadoop Pratique: • Les bases de Pig Latin • Partage de metadonnées avec HCatalog
Etudes de Cas :	• Analyser des données COVID-19 • Analyser des données météo • Analyser les données des crimes de la ville de Chicago
3ème Partie	
Classe 8	**Ingérer des données à grande échelle -1** Ingestion des données à la demande Apache Sqoop • Concepts et principes de fonctionnement • Anatomie des operations Import/Export de données • Configuration d'une operation Sqoop Pratique: • Importer de données à partir d'un serveur base de données MySQL • Importer / Exporter des données vers HDFS, Hive et HBase

Classe 9	**Ingérer des données à grande échelle -2**
	
	Ingestion des données en temps réel
	Introduction à Flume et Nifi
	• Ingérer des données en temps réel
	• Configuration d'un agent Flume (définir Source / Channel / Sink)
	  ○ Ingérer des données dans HDFS / Hive
	  ○ Utilisation de la réplication et du multiplexage des canaux
	• Comprendre un FileFlow Nifi
	  ○ Présentation des composants de Nifi
	  ○ Anatomie d'un flux d'ingestion (dataflow)
	  ○ Présentation des FlowFiles et comment les utiliser
	  ○ Création d'un dataflow d'ingestion des données

	Pratique: • Ingestion de données en temps réel à l'aide de Flume • Création de dataflow Nifi pour ingérer et transformer les données
Classe 10	**Ingérer des données à grande échelle –3** Ingestion des données grande vitesse en temps réel -1 Apache HBase • Introduction aux bases de données NoSql (quatre familles) • Introduction à HBase - La base de données Hadoop NoSQL • Caractéristiques et fonctionnement de HBase • Stockage et récupération des données de HBase Pratique: • Manipulation des données dans HBase *(create, read, update, delete)* • Ingérer des données dans HBase à partir de Sqoop ; Flume et Nifi
Classe 11	**Ingérer des données à grande échelle –4** Ingestion des données grande vitesse en temps réel -2 Introduction à Apache Kafka • Qu'est-ce que Kafka et comment ça marche • Comprendre les Borkers et les Topics Kafka • Utilisation des des Producers et Consumers dans Kafka Pratique : • Lancer un serveur Kafka • Création de topics Kafka et insertion de données avec Flume et Nifi
Classe 12	**Ingérer des données à grande échelle –5** Ingestion des données grande vitesse en temps réel -3 Introduction à Spark Streaming • Comprendre le rôle de Spark Streaming • Utiliser Spark Streaming pour ingérer des données en temps réel • Ingestion des données avec Spark Structured Streaming Pratique : • Ingestion de données avec Spark Sreaming • Ingestion de données avec Spark Structured Streaming

Etudes de Cas :	• Ingérer et analyser des données financières en temps réel • Ingestion de données météo en temps réel et création d'un tableau de bord • Ingestion et analyse des données vélo libre de la ville de Montréal (Bixi)
Atelier Twitter	**Création d'un tableau de bord temps réel des données de Twitter** • Création d'une application Twitter • Ingérer des données en temps réel à partir de Twitter (Flume et Nifi) • Stockage et Analyse des données Twitter à l'aide de Hive • Mise en place d'une analyse de sentiments (basic) • Export et Visualisation des données avec Microsoft Excel (Tableau, Microstrategy,…)
Projet de fin Formation - 90%	