Module : Big Data Analytics

Dernière mise à jour : 11/02/2022

Code	HE	HNE	ECTS
BD-24	42h	33h	3

Responsable Module	Sirine ZAABOUTI (sirine.zaabouti@esprit.tn)
Enseignants – Intervenants	InesSlimene, InesChannoufi, Sirine ZAABOUTI
Unité pédagogique	GL-BD
Unité d'enseignement	Informatique Décisionnelle
Prérequis	BD-02 / AP-09
Niveaux et Options	4ème ERP-BI, 4ème DS, 4ème IA, 4ème WIN

Objectif du module :

A la fin de ce module l'apprenant sera capable d'identifier les concepts du big data, d'analyser des données volumineuses et de manipuler des plateformes distribuées.

Mode d'évaluation :

La moyenne de ce module est calculée comme suit :

Moyenne (BDATA) = contrôle continu *40% + Examen *60%.

La note du contrôle continu est la moyenne des travaux pratiques présentiels et les travaux individuels non présentiels (Homework) tout au long de la formation.

Acquis d'apprentissage :

À la validation de ce module l'étudiant sera capable de :

	Acquis d'apprentissage	Niveau d'approfondissement (*)

AA1	Définir le concept Big Data et citer ses termes clefs	1
AA3	Illustrer les spécificités d'une architecture distribuée	2
AA4	Faire de la programmation parallèle avec MapReduce	2
AA5	Comparer les langages de requête Hadoop	4
AA6	Utiliser la suite ELK pour le stockage et la visualisation des données distribuées	3
	Choisir les représentations graphiques appropriées	6
AA7	Définir l'outil de traitement des données distribuées en mémoire : Spark	1
AA8	Développer des applications avec Spark Streaming	5
AA9	Différencier entre l'apprentissage supervisée et l'apprentissage non supervisée	4
	Evaluer les performances d'une méthode	6

* : (1 : Mémoriser, 2 : Comprendre, 3 : Appliquer, 4 : Analyser, 5 : Evaluer, 6 : Créer).

Contenu détaillé

Chapitre1 : Introduction du concept big data

- Définir le concept Big Data
- Identifier les caractéristiques du Big Data.
- Citer les différents cas d'utilisation du Big Data.
- Décrire l'écosystème Hadoop.
- Identifier les distributions Hadoop existantes.
- Manipuler la distribution Cloudera.

Situation(s) d'apprentissage	Cours
Durée	1h30min
Rendu(s)	

Chapitre2 : Système de fichier distribué

- Expliquer le principe du système de fichiers distribué HDFS.
- Identifier et différencier entre les composants de l'architecture HDFS.
- Comparer l'architecture HDFS1 et HDFS2.
- Expérimenter la gestion de fichiers sous HDFS

Situation(s) d'apprentissage	Cours
Durée	1h30min
Rendu(s)	Homework

Chapitre3 : Traitement de données

- Identifier et différencier entre les composants de l'architecture MapReduce.
- Distinguer le fonctionnement map et le fonctionnement reduce.
- Comparer l'architecture MapReduce1 et MapReduce2
- Appliquer des programmes MapReduce.

Situation(s) d'apprentissage	Cours intégré
Durée	1h30,1h30
Rendu(s)	Workshop

Chapitre4 : Requêtage des données structurées

- Décrire le composant Hive.
- Expliquer le langage de requête HiveQL.
- Illustrer les performances de Hive.
- Manipuler le langage HiveQL.

Situation(s) d'apprentissage	Cours intégré
Durée	1h30,1h30
Rendu(s)	TP

Chapitre5 : Transfert des données

- Manipuler flume pour extraire des données à partir de tweeter et les stocker sous HDFS
- Manipuler sqoop pour extraire des données à partir d'une base de données relationnelle

Situation(s) d'apprentissage	Workshop
Durée	6h
Rendu(s)	Workshop

Chapitre6 : Moteur de recherche et d'analyse des données distribuées

- Mettre en place ElasticSearch
- Utiliser ElsticSearch en tant que Base de données
- Voir les données avec kibana

Situation(s) d'apprentissage	Workshop
Durée	3h
Rendu(s)	Workshop

Chapitre7 : Introduction au Framework de traitements big data & analyses complexes

- Reconnaitre l'historique du Framework Apache Spark.
- Identifier les différentes versions de Spark (Scala, Python et Java).
- Comparer avec l'environnement Apache Hadoop.
- Distinguer les différents modules de Spark.
- Manipuler Apache spark

Situation(s) d'apprentissage	Cours

Durée	3h
Rendu(s)	

Chapitre8 : Structure de données Distribuées de Framework Spark

- Retenir les notions des Resilient Distributed DataSets (RDD).
- Créer, manipuler et réutiliser des RDD.
- Identifier les accumulateurs et variables broadcastées.
- Utiliser des partitions.

Situation(s) d'apprentissage	Cours intégré
Durée	1h30min,1h30min
Rendu(s)	TP

Chapitre9 : Requêtage des données structurées et semi-structurées avec le Framework Spark

- Mettre en pratique le SQL, DataFrames et Datasets.
- Identifier les différents types de sources de données.
- Pratiquer avec les RDD.
- Distinguer la performance de Spark SQL.

Situation(s) d'apprentissage	Cours intégré
Durée	1h30,4h30
Rendu(s)	TP

Chapitre10 : Streaming des données en temps réels

- Décrire le principe de fonctionnement Spark Streaming.
- Identifier des Discretized Streams (DStreams).
- Distinguer les différents types de sources.
- Manipuler l'API Spark Streaming

Situation(s) d'apprentissage	Cours intégré
Durée	1h30,4h30
Rendu(s)	TP

Chapitre11 : Bibliothèque d'apprentissage automatique de Framework Spark

- Définir le Machine Learning.
- Distinguer les différentes classes d'algorithmes.
- Reconnaitre le SparkML et MLlib.
- Manipuler les différents algorithmes dans MLlib.

Situation(s) d'apprentissage	Cours intégré
Durée	1h30,4h30
Rendu(s)	TP

Evaluation :

	Oral assessment	Written exam/ MCQ	Report/ Homework	Presentation	TP	Project
Définir le concept Big Data et citer ses termes clefs	X	X				
Illustrer les spécificités d'une architecture distribuée	X			X		
Faire de la programmation parallèle avec MapReduce					X	
Comparer les langages de requête Hadoop	X			X	X	
utiliser la suite ELK pour le stockage et la visualisation des données distribuées					X	X
Choisir les représentations graphiques appropriées			X		X	X
Définir l'outil de traitement des données distribuées en mémoire : Spark	X	X		X		
Développer des applications avec Spark Streaming					X	X
Différencier entre l'apprentissage supervisée et l'apprentissage non supervisée	X	X		X		
Evaluer les performances d'une méthode					X	X

Références :

Références bibliographiques	bigdatauniversity.com
	Support formation IBM BigInsight
	databricks.com