

Plan de cours

Cours :	INF 745 – Mégadonnées (« Big data »)
Trimestre :	Été 2021
Enseignant :	Robert J. Laurin

1. Mise en contexte

Dans la dernière décennie, l'utilisation stratégique de l'information a changé le monde des affaires. Ce changement s'est produit grâce à l'utilisation d'entrepôts de données comme outil d'aide à la décision mais aussi grâce aux méthodes d'analyse de Mégadonnées (« Big data »). Il est devenu essentiel pour le gestionnaire de connaître les enjeux que représentent les outils d'exploitation de la Mégadonnée pour survivre dans le monde des affaires comme dans celui des technologies de l'information.

2. Place du cours dans le programme

Ce cours s'inscrit dans le programme de Diplôme de 2^e cycle de technologies de l'information ainsi que dans le diplôme en génie logiciel. Il a comme préalable le cours INF735 « Entrepôt et forage de données ».

3. Descripteur du cours

Objectifs : Comprendre les concepts sous-jacents à la réalisation d'un environnement de Mégadonnées et les différents outils disponibles et nécessaires dans cet univers. Comprendre les différentes couches de logiciels et les enjeux techniques derrière la réalisation des objectifs d'affaires. Comprendre les différentes approches des bases de données « No SQL » ou non-relationnelles.

Contenu : Définitions et liens avec le contexte d'affaires d'une entreprise. Revue des enjeux et définition des besoins. Caractéristiques et architecture d'un environnement de mégadonnées. Modèles de données. Couches logicielles et techniques. Démarche de construction. Connexion des outils de forage. Conception d'un plan de mise en place d'un environnement de mégadonnées. Mise en pratique des notions et des outils présentés en classe

4. Objectifs spécifiques

Au terme du cours, l'étudiante ou l'étudiant sera capable :

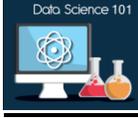
1. d'appliquer les principaux concepts sous-jacents à l'environnement de données éparses et leur organisation en Mégadonnées;
2. d'identifier les différentes approches informationnelles et de choisir l'architecture de données non-relationnelle appropriée;
3. de décrire les étapes cruciales dans la construction et la modélisation de Mégadonnées;
4. d'utiliser les outils d'exploitation et de forage de données.

5. Planification hebdomadaire

CONTENU DU COURS	Séances
Bloc 1 - Introduction: besoin, concepts et définitions <ul style="list-style-type: none">• Positionnement du cours• Définitions de « Mégadonnées » et « Méga-Analyses »• Pourquoi maintenant ?• Évolution des données, les 3 V• Une approche BD différente : NoSQL• Théorème de CAP	1 - 2
Bloc 2 – Les couches logicielles <ul style="list-style-type: none">• Fondements de Hadoop• HDFS• MapReduce• Serveurs, racks et réseaux• Les couches de la pile	2-3
Bloc 3 – Les modèles de données <ul style="list-style-type: none">• Les architectures NoSQL<ul style="list-style-type: none">○ Clé-valeur○ En colonnes○ Documents○ Graphiques○ En mémoire• Choix du schéma de BD• Distribution et réplication	3 – 4

CONTENU DU COURS	Séances
Bloc 4 –Acquisition de données <ul style="list-style-type: none"> • Méthodes et outils selon le type de source <ul style="list-style-type: none"> ○ Données structurées et semi-structurées ○ Données du Web 2.0 ○ Données non-structurées • Métadonnées 	4-5
Bloc 5 – Exploitation des données <ul style="list-style-type: none"> • Techniques de Méga-analyse • Analyse des médias sociaux • Outils de Méga-analyse <ul style="list-style-type: none"> ○ Python ○ R ○ Tableau ○ Etc. 	5-6
Bloc 6 – Gestion du projet et réalisation <ul style="list-style-type: none"> • Gestion d’un projet de Mégadonnées • Architecture hybride • Équipe • Particularités • Éthique • Conclusion 	6+
Examen Final (via Moodle)	

CONTENU DES TUTORIAUX EN LIGNE	Heures
<p data-bbox="342 285 659 310">Tutoriaux en ligne :</p> <p data-bbox="391 342 704 380"><u>1^{er} : Big Data 101</u> (6%)</p> <div data-bbox="436 415 552 531" data-label="Image"> </div> <p data-bbox="436 569 1070 638">https://cognitiveclass.ai/courses/what-is-big-data/ Temps estimé à 3 heures. En anglais seulement.</p> <p data-bbox="391 743 725 781"><u>2^{ième} : Hadoop 101</u> (10%)</p> <div data-bbox="436 821 612 913" data-label="Image"> </div> <p data-bbox="436 951 1055 1058">https://cognitiveclass.ai/courses/course-v1:BigDataUniversity+BD0111EN+2016 Temps estimé à 5 heures. En anglais seulement.</p> <p data-bbox="391 1165 850 1203"><u>3^{ième} : Spark Fundamentals I</u> (9%)</p> <div data-bbox="436 1241 597 1329" data-label="Image"> </div> <p data-bbox="436 1367 1055 1436">https://cognitiveclass.ai/courses/what-is-spark/ Temps estimé à 5 heures. En anglais seulement.</p> <p data-bbox="391 1543 904 1581"><u>4^{ième} : Moving Data Into Hadoop</u> (8%)</p> <div data-bbox="436 1619 597 1711" data-label="Image"> </div> <p data-bbox="436 1749 1055 1856">https://cognitiveclass.ai/courses/flume-sqoop-moving-data-into-hadoop/ Temps estimé à 4 heures. En anglais seulement.</p>	<p data-bbox="1175 285 1208 310">25</p>

CONTENU DES TUTORIAUX EN LIGNE	Heures
<p><u>5^{ème} : Data Science 101</u> (7%)</p>  <p>https://cognitiveclass.ai/courses/data-science-101/ Temps estimé à 3 heures. En anglais seulement.</p> <p><u>6^{ème} : Python For Data Science</u> (10%)</p>  <p>https://cognitiveclass.ai/courses/python-for-data-science/ Temps estimé à 5 heures. En anglais seulement.</p>	

6. Approche pédagogique préconisée

Le cours permet à l'étudiante ou l'étudiant de prendre en charge de ses apprentissages. L'enseignant est là comme guide et soutien dans cette démarche. Les notions théoriques indispensables à l'atteinte des objectifs seront exposées en classe en exploitant l'expérience vécue par chaque personne dans son milieu de travail. L'utilisation de logiciels de manipulation de données apportera une dimension pratique indispensable à l'atteinte des objectifs du cours. Avant chaque prestation, les étudiantes et les étudiants s'engagent à lire les chapitres du livre et les notes de cours mentionnés dans la planification hebdomadaire.

Le chargé de cour sera disponible virtuellement pour assister dans les exercices du laboratoire personnel.

7. Évaluation de l'apprentissage

Les examens comportent des questions :

- à choix multiples,

- à développement court (moins d'une demi-page),
- à développement long.

L'enseignant évaluera seul les examens et les travaux.

Aucun travail en retard ne sera accepté, basé sur la date et l'heure de l'envoi électronique. L'enseignant peut retrancher jusqu'à 5% de chaque évaluation pour la qualité du français.

Description sommaire	Pondération
6 Tutoriaux en ligne avec Quiz -évaluation individuelle	50 %
Examen final -évaluation individuelle	50 %

8. Plagiat

Conformément à l'article 9 du Règlement des études de l'Université de Sherbrooke, le plagiat, soit le fait dans une activité évaluée de faire passer pour sien les idées et le travail d'autrui, est un délit académique qui peut être sanctionné par les autorités disciplinaires compétentes. Peuvent être imposées à titre de sanctions, l'une ou plusieurs des mesures suivantes :

- a) la réprimande simple ou sévère consignée au dossier étudiant pour la période fixée par l'autorité disciplinaire ou, à défaut, définitivement. En cas de réprimande fixée pour une période déterminée, la décision rendue demeure au dossier de la personne aux seuls fins d'attester de l'existence du délit en cas de récidive;
- b) l'obligation de reprendre une production ou une activité pédagogique, dont la note pourra être établie en tenant compte du délit survenu antérieurement;
- c) la diminution de la note ou l'attribution de la note E ou 0;
- d) le renvoi du dossier à la personne responsable de l'évaluation d'une production ou d'une activité pédagogique pour qu'elle attribue une nouvelle note en tenant compte du délit.

Par plagiat, on entend notamment :

- copier intégralement une phrase ou un passage d'un livre, d'un article de journal ou de revue, d'une page Web ou de tout autre document en omettant d'en mentionner la source ou de le mettre entre guillemets;
- reproduire des présentations, des dessins, des photographies, des graphiques, des données sans en préciser la provenance et, dans certains cas, sans en avoir obtenu la permission de reproduire;
- utiliser, en tout ou en partie, du matériel sonore, graphique ou visuel, des pages Internet, du code de programme informatique ou des éléments de logiciel, des données ou résultats d'expérimentation ou toute autre information en provenance d'autrui en le faisant passer pour sien ou sans en citer les sources;
- résumer ou paraphraser l'idée d'un auteur sans en indiquer la source;
- traduire en partie ou en totalité un texte en omettant d'en mentionner la source ou de le mettre entre guillemets;
- utiliser le travail d'un autre et le présenter comme sien (et ce, même si cette personne a donné son accord);
- acheter un travail sur le Web ou ailleurs et le faire passer pour sien;
- utiliser sans autorisation le même travail pour deux activités différentes (autoplégat).

9. Adresse électronique pour la remise des travaux

Robert.J.Laurin@usherbrooke.ca

10. Logiciels (optionnels)

Aucun nécessaire