

Sigle : INF6303 Gr. 01**Titre : Techniques d'analyse des mégadonnées****Session : Automne 2020 Horaire et local****Professeur : Yapi, N'Dah Daniel****1. Description du cours paraissant à l'annuaire :****Objectifs**

Connaître les concepts fondamentaux liés à la production et l'exploitation des mégadonnées. Maîtriser les techniques d'analyse statistique, d'exploration et de visualisation de mégadonnées pour la prise de décision.

Contenu

Description des défis du traitement de données massives (big data) : Échantillonnage et prétraitement des données : structurées, non structurées et flux. Techniques statistiques d'analyse des mégadonnées : modélisation et inférence statistique, estimation de paramètres et de distributions, maximum de vraisemblance et méthodes Bayésiennes, analyse de corrélation, analyse factorielle. Techniques de stockage, de réduction et de visualisation de mégadonnées. Apprentissage par ordinateur sur les mégadonnées : classification et regroupement, régression linéaire et non-linéaire, apprentissage profond. Étude d'applications utilisant les données massives : recherche d'information, systèmes de recommandation, analyse de réseaux sociaux.

Descriptif – Annuaire

2. Objectifs spécifiques du cours :

Au terme de cette activité, l'étudiant(e) aura acquis des connaissances sur :

- les concepts fondamentaux de l'analyse des mégadonnées;
- les techniques de prétraitement et de visualisation des mégadonnées;
- les techniques de classification et de regroupement des mégadonnées;
- les techniques de détection d'anomalies;
- les applications de l'analyse des mégadonnées.

3. Stratégies pédagogiques :

Toutes les activités auront lieu en présentiel.

- Séances de cours magistral en présentiel, de 3h/semaine comprenant :
- Cours magistral (Architecture Big Data et concepts analytiques)
- Laboratoires pratiques (Implémentation sous Docker avec l'écosystème Apache)
- Lectures d'articles
- Disponibilité d'une page **Moodle** contenant le matériel du cours et les résultats des évaluations des travaux.
- Un examen de mi-session, un examen final et un projet finale en **présentiel**.

Communication via Moodle et courriels et forums.

4. Heures de disponibilité ou modalités pour rendez-vous :

Courriel : Sur demande : yapida01@uqo.ca

5. Plan détaillé du cours sur 15 semaines :

Semaine	Thèmes	Dates
1	Cours en ligne Zoom Architecture des Mégadonnées et Calcul Distribué • Limites des systèmes centralisés et passage à l'échelle. • Écosystème Hadoop et système de fichiers distribué (HDFS). • Virtualisation légère et conteneurisation. <i>Lab : Déploiement d'un cluster Big Data avec Docker.</i>	9 sept. 2026
2	Paradigme MapReduce et Traitement par Lots (Batch) • Modèle de programmation MapReduce. • Nettoyage et transformation de données massives. • Ordonnancement de tâches. <i>Lab : Ingestion et traitement par lots d'un corpus massif</i>	16 sept. 2026
3	Cours en ligne Zoom Écosystème NoSQL : Stockage Distribué • Théorème CAP et consistance éventuelle. • Bases de données orientées documents (MongoDB). • Bases de données orientées colonnes (Cassandra). <i>Lab : Modélisation et requêtes d'analyse sur bases NoSQL</i>	23 sept. 2026
4	Traitement en Mémoire : Apache Spark et RDDs • Architecture de Spark et évaluation paresseuse. • Résilient Distributed Datasets (RDDs) • Opérations de transformation et d'action. • Inférence <i>Lab : Programmation fonctionnelle distribuée avec PySpark</i>	30 sept. 2026
5	Analyse de Données Structurées à l'Échelle (Spark SQL) • Spark DataFrames et Datasets. • Optimisation de requêtes (Catalyst Optimizer). • Agrégations complexes sur téraoctets de données.. <i>Lab : Exploration et analyse statistique de datasets géants.</i>	7 oct. 2026
6	Semaine d'études	14 oct. 2026
7	Examen de mi-session	21 oct. 2026
8	Cours en ligne Zoom Apprentissage Automatique Distribué I (Spark MLlib) • Défis de l'IA sur les mégadonnées. • Extraction et transformation de caractéristiques à l'échelle. • Algorithmes de classification et régression distribués <i>Lab : Entraînement de modèles prédictifs sur grappe.</i>	28 oct. 2026

9	Apprentissage Automatique Distribué II (Spark MLlib) • Regroupement (Clustering) avec K-Means distribué. • Moteurs de recommandation (Filtrage collaboratif, ALS). • Évaluation et validation de modèles à grande échelle <i>Lab : Segmentation de données et systèmes de recommandation.</i>	4 nov. 2026
10	Ingestion et Messagerie Distribuée (Apache Kafka) • Architecture Pub/Sub et courtiers de messages. • Découplage des flux d'ingestion analytique. • Gestion des partitions et de la tolérance aux pannes. <i>Lab : Déploiement et routage de messages avec Kafka.</i>	11 nov. 2026
11	Analyse de Flux de Données en Temps Réel (Streaming) • Différence entre Batch et Micro-batch / Continuous streaming. • Apache Spark Streaming et Structured Streaming. • Architectures Lambda et Kappa. <i>Lab : Pipeline d'analyse de données en continu.</i>	18 nov. 2026
12	Architectures Modernes : Data Lakehouse et Cloud • Formats de stockage analytiques optimisés (Parquet, ORC). • Évolution vers Delta Lake (Transactions ACID sur Data Lake). • Découplage calcul/stockage dans les environnements infonuagiques. <i>Lab : Création et requêtage d'un Data Lakehouse.</i>	25 nov. 2026
13	Visualisation Analytique et Recherche Documentaire Massive • Moteurs de recherche textuelle distribués (Elasticsearch). • Création de tableaux de bord en temps réel (Kibana). • Intégration du pipeline analytique de bout en bout <i>Lab : Indexation massive et visualisation de logs analytiques.</i>	2 déc. 2026
14	Présentation des projets	9 déc. 2026
15	Examen final	16 déc. 2026

6. Évaluation du cours :

L'étudiant(e) dans ce cours sera évalué(e) par les laboratoires/devoirs, examens de mi-session et final, ainsi que par des projets de session. La pondération de la note finale se fera comme suit :

- **Devoirs/Laboratoires : 15 %**
- **Projets : 25 %**
- **Examen de mi-session : 30 %**
- **Examen final : 30 %**

Le projet de session consistera à concevoir un pipeline analytique complet (ingestion, stockage, analyse distribuée, et visualisation) sur un jeu de données réel. Ces derniers peuvent être dans l'analyse de données multimédias : texte, son, image ou Web. Un projet typique sera donné par le professeur

L'évaluation se fera sur deux volets :

- Un rapport d'environ 15 pages. Les normes de présentation des travaux (p. ex. page de garde, marge) doivent être **absolument** respectées.
- Une présentation de 15 à 20 minutes. Les projets se feront individuellement ou en équipes de deux si le nombre d'étudiant(e)s inscrit(e)s dépasse 10.

Une moyenne générale inférieure à **64 %** est éliminatoire et conduit automatiquement à l'échec de l'étudiant(e).

Tout retard dans la remise d'un travail entraîne une pénalité de **15 %/jour** sur la note attribuée à ce travail, **jusqu'au maximum d'une semaine**. La qualité du français sera considérée lors de la correction des travaux.

7. Politiques départementales et institutionnelles :

- Politique du département d'informatique et d'ingénierie relative à la tenue des examens
- Note sur le plagiat et sur la fraude
- Politique relative à la qualité de l'expression française écrite chez les étudiants et les étudiantes de premier cycle à l'UQO
- Absence aux examens : cadre de gestion, demande de reprise d'examen (formulaire)

À l'UQO, **les violences à caractère sexuel, c'est tolérance zéro!**

La communauté universitaire s'engage à lutter contre les inconduites, le harcèlement et les violences à caractère sexuel : parce que **le respect, c'est l'affaire de tout le monde!**

N'oubliez pas de faire la formation obligatoire :

uqo.ca/bimi/formation-obligatoire

Pour de plus amples renseignements :

bimi@uqo.ca



8. Principales références :

1. **Note** : Le cours du professeur est la référence bibliographique principale
2. Chambers, B., & Zaharia, M. (2018). Spark: The Definitive Guide. O'Reilly.
3. Dean, J., & Ghemawat, S. (2004). MapReduce: Simplified Data Processing on Large Clusters. OSDI'04: Sixth Symposium on Operating System Design and Implementation
4. White, T. (2015). Hadoop: The Definitive Guide. O'Reilly.
5. Narkhede, N., Shapira, G., & Palino, T. (2017). Kafka: The Definitive Guide. O'Reilly.
6. Leskovec, J., Rajaraman, A., & Ullman, J. D. (2020). Mining of Massive Datasets (3rd ed.). Cambridge University Press.
7. Sadalage, P. J., & Fowler, M. (2012). NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence. Addison-Wesley Professional.
8. Marz, N., & Warren, J. (2015). Big Data: Principles and best practices of scalable realtime data systems. Manning Publications.
9. Zaharia, M., Chowdhury, M., Das, T., Dave, A., Ma, J., McCauley, M., ... & Stoica, I. (2012). Resilient Distributed Datasets: A Fault-Tolerant Abstraction for In-Memory Cluster Computing. NSDI'12: 9th USENIX Symposium on Networked Systems Design and Implementation

9. Page Web du cours :

<https://moodle.uqo.ca>