T3 : Données massives scientifiques (Big data), recherche par les données

En cette époque de ruée vers la donnée, les données massives (big data) résultant de l'explosion des capteurs, de l'open-data, de la complexité et de l'interdisciplinarité des recherches actuelles deviennent centrales.

Dans cette thématique, nous expliciterons les masses de données et leurs enjeux. Nous ferons le point sur les outils et méthodes qui vous sont nécessaires dans le contexte de votre projet de développement logiciel et de recherche par les données. Nous ferons un tour d'horizon pour identifier si selon le contexte recherche, les limites des technologies en matière de traitement et de stockage de haute volumétrie sont atteintes et si alors il faut utiliser des technologies adhoc labélisées “big data” ou si des techniques plus “traditionnelles” suffisent.

Mots-clés : données ouvertes (open data), base de données, masse de données (big data), persistance des données, structuration des données, map-reduce, NoSQL, R, Julia, Python.

Présentations

mercredi 1 Juillet, 9h00-12h30, amphi D

T3.P1 – 9h00-9h40 : *Définition et enjeux du big data (Nouveaux domaines, Nouveaux métiers, Nouveaux champs). Exemple d'applications : collecter, stocker, analyser.*

Intervenant : Philippe Lacomme (ISIMA, Clermont-Ferrand) et Raksmey Phan (ISIMA)

T3.P2 - 9h40-10h15 : *Les enjeux techniques actuels : stocker. Les nouvelles catégories de bases de données (les différentes solutions, API, disponibilités, …). Exemple : MongoDB, Cassandra, Oracle. Principe de modélisation avec les bases NoSQL. Analyse Critique de cette approche.*

Intervenant : Philippe Lacomme (ISIMA, Clermont-Ferrand)

T3.P3 - 10h15-10h30 : *Les bases de données NOSQL orientée graphes : illustration avec Neo4j.*

Intervenant : Cédric Fauvet (Néotechnology)

T3.P4 - 11h00-11h45 : *Donnez du sens à vos données (Elasticsearch, un moteur de recherche open source pour implémenter vos services de recherche big data et d'analyse).*

Intervenant : David Pilato (Elastic)

T3.P5 - 11h45-12h30 : *Hadoop, MapReduce et Spark pour vos développements de service.*

Intervenant : Miguel Liroz Gistau (INRIA, Montpellier)

Ateliers Préparatoires

T3.AP03 : *Initialisation à Python.*

Intervenant : Sékou Diakité (Institut UTINAM, Besançon).

Ateliers

T3.A01a : *Base de données NoSQL – Oracle.*

Intervenant: Raksmey Phan (ISIMA, Clermont-Ferrand)

T3.A01b : *Base de données NoSQL orientée graphe Neo4J.*
Intervenant: Cédric Fauvet (Néotechnology)

T3.A02 : *Base de données NoSQL - Mongo DB.*
Intervenant: Raksmey Phan (ISIMA, Clermont-Ferrand)

T3.A02 bis : *Base de données NoSQL – Cassandra.*
Intervenant: Raksmey Phan (ISIMA, Clermont-Ferrand)

T3.A03 : *Analyse de données massives pour la recherche de patterns liés aux comportements d'utilisateurs : application à l'analyse de log d'une université.*
Intervenant: Jonathan FONTANEL (QUALIAC ERP, Chamalière)

T3.A04 : (Fusionné avec T8.A03) *Atelier Hadoop : Map-Reduce et Spark pour le calcul scientifique.*
Intervenant: Pierre Senellart (Télécom Paris Tech et National University of Singapore)

T3.A06 : *Julia pour vos calculs scientifiques intensifs.*
Intervenant: Mickael Canouil (UMR 8199 GIM3, Lille)

T3.A08 : *Python, apprentissage statistique et analyse de données pour la modélisation prédictive avec scikit-learn.*
Intervenant: Olivier Grisel (INRIA Saclay)

T3.A10 : *Traitement des données en parallèle avec Map-Reduce et Spark.*
Intervenant: Miguel Liroz Gistau (INRIA, Montpellier)

T3.A13 : *Prise en main d'elasticsearch et de Kibana.*
Intervenant: David Pilato (Elastic)

Groupes de travail

T3.GT02 : *Comment faire parler, analyser mes données? Données structurées et semi-structurées.* Jonathan FONTANEL (QUALIAC ERP, Chamalière), Nicolas Larrousse, TGIR Huma-Num (Paris)

T3.GT03 : *Foire aux bibliothèques thématiques scientifiques python et autres par effet de glue. Echange et retour d'expérience. Extension des bibliothèques de références numpy, scipy, matplotlib à sa thématique scientifique. Ouverture à d'autres langages (fortran, C, R, julia, …) , intégration de l'existant par le côté glue du langage.* Mickael Canouil (GIM3, Lille).

T3.GT05 : *Les différentes méthodes d'optimisation d'un code Python (Cython, Numba, programmation parallèle ou utilisation de GPU avec CUDA). Exemple de calcul d'une fractale (Mandelbrot).* Tristan Colombo

T3.GT06 : *Programmation orientée objet interprétée.* Yves Auda (GET/OMP, Toulouse)

T3.GT07 : *Intégration de données (projet apache Camel).* Stéphane Deraco (DSI CNRS, Toulouse)

T3.GT08 : *Comment et pourquoi certifier son centre de données ?* Françoise Genova (Obs. de Strasbourg)

T3.GT09 : *Comment contribuer à RDA? Comment constituer un GT, trouver des partenaires ? Françoise Genova (Observatoire de Strasbourg)*