Nom et Prénom : NIBAREKE THERENCE
Date de soutenance : 07/07/2021
Directeur de Thèse : LAASSIRI JALAL

Sujet de Thèse :

Proposition, implementation and evaluation of Machine Learning models for Big Data analysis and deployment of processing in Cloud Computing platforms

Résumé :

Avec l'avènement de IOT et du web 3.0 plusieurs téraoctets de textes, images, vidéos, audio par seconde. Le Big Data désigne de grandes quantités de données structurées et non structurées produites à une grande vitesse.Les institutions tant gouvernementales que privées analysent des téraoctets de textes, chiffres, audio, vidéos, images pour une prise de décision stratégique (le E-commerce, banques, réseaux sociaux, E-santé, smart cities, capteurs). Les capteurs météorologiques prédisent la météo, Les services de sécurité analysent les comportements des individus et prévenir les menaces.Les SGBDR(SQL,Oracle) ne peuvent plus gérer cette génération de données. Les SGBD NOSQL, les systèmes distribués Hadoop, Spark et Machine Learning permettent une analyse descriptive et prédictive des données. Toutefois la performance de ces outils et algorithmes reste un défi majeur.

Cette thèse a pour objectif d'étudier et comparer les outils Big Data et Machine Learning, Proposer un modèle de déploiement des traitements Big Data dans le Cloud Computing.Nous avons comparé les composants de Hadoop. MapReduce a été plus performant en termes de temps d'exécution. Nous avons développé un modèle qui permet de prédire le risque de diabète à partir des données du corps humain avec une précision de 0.766%. Le déploiement de l'analyse des données dans le Cloud Computing nous a permis d'optimiser les coûts en infrastructure avec une plus grande disponibilité.

MOTS-CLES:
Big Data, Hadoop Ecosystem, Spark, NOSQL databases, Cloud computing, Machine Learning, Performance Analysis, Diabetes prediction, tweets analysis, Data analytics

Abstract :

With the advent of IOT and web 3.0 several terabytes of text, images, videos, audio per second.Big Data refers to large amounts of structured and unstructured data produced at high speed. Both government and private institutions analyze terabytes of text, numbers, audio, videos, images for strategic decision-making (E-commerce, banks, social networks, e-health, smart cities, sensors). Weather sensors predict the weather, Security services analyze individual behavior and prevent attack risk. RDBMS(SQL,Oracle) can no longer manage this generation of data. NOSQL DBMS, Hadoop, Spark and Machine Learning distributed systems allow descriptive and predictive data analysis. However, the performance of these tools and algorithms remains a major challenge.

This thesis aims to study and compare Big Data and Machine Learning tools, Propose a model for deploying Big Data processing in Cloud Computing. We compared the components of Hadoop . MapReduce performed better in terms of execution time. We have developed a model that predicts the risk of diabetes from human body data with an accuracy of 0.766%. Deploying data analytics in cloud computing has allowed us to optimize infrastructure costs with greater availability.

KEY WORDS:
Big Data, Hadoop Ecosystem, Spark, NOSQL databases, Cloud computing, Machine Learning, Performance Analysis, Diabetes prediction, tweets analysis, Data analytics