

Entrepôt de Données dans l'ère Data Science : De la Donnée au Modèle

Cyrille Ponchateau, Ladjel Bellatreche, Mickael Baron

LIAS/ISAE-ENSMA, Université de Poitiers
Téléport 2 - 1 avenue Clément Ader - BP 40109,
86961 Futuroscope Chasseneuil Cedex
France
firstname.lastname@ensma.fr

Résumé. Dans l'ère Data Science, un nombre important de domaines scientifiques souhaitent analyser leurs données. Souvent dans ces domaines, les données des tests sont représentées par des séries chronologiques. Ces dernières sont une classe de données temporelles, comprenant un enregistrement chronologique de valeurs, considérées comme un tout et non comme une liste de données individuelles et indépendantes. De plus, les séries chronologiques sont généralement composées d'un grand nombre de valeurs et peuvent être stockées dans des bases de données classiques, parfois en très grande quantité. Dans cet article, nous proposons un moyen de stockage des séries chronologiques, par abstraction de la série par son modèle (équation différentielle), conduisant ainsi à la notion nouvelle d'*entrepôt de modèles* ayant pour but de proposer une autre représentation des séries chronologiques et une solution alternative de stockage. Cependant, cette méthode induit un coût non-négligeable en temps de calcul. Notre proposition est implémentée et validée en utilisant des données réelles issues des expérimentations du domaine de l'automatique.

1 Introduction

Les dernières « avancées » en matière des technologies de l'information, matériels, de plateformes de déploiement de grande puissance et de la maturité de la technologie d'entreposage de données ont motivé un nombre important d'organismes industriels et de recherche afin de stocker leurs expérimentations à des fins d'analyse. Souvent, les données générées par ses expérimentations sont représentées par des séries chronologiques.

Les séries chronologiques sont une classe de données temporelles, comprenant un enregistrement chronologique de valeurs. Une série chronologique est considérée comme un tout et non comme une liste de données individuelles et indépendantes (Fu (2011); Esling et Agón (2012)). En effet, il existe une corrélation entre la valeur d'un point et les valeurs des points adjacents, aucune valeur ne peut donc être considérée comme totalement indépendante des autres Shumway et Stoffer (2015). De plus, ces dernières sont généralement composées d'un grand nombre de valeurs, parfois continuellement mises à jour. Les séries chronologiques peuvent