Brochure Solution

Que permettent les offres SAS hautes-performances ?

Les offres hautes-performances de SAS permettent de développer des modèles sur des données exhaustives et non plus seulement sur un sous-ensemble. Il est désormais possible de développer et d'exécuter des modèles utilisant des milliers de variables et des millions de documents, et d'obtenir des réponses plus précises et plus rapides. Les traitements, qu'ils relèvent des statistiques, du data mining, du text mining, de l'économétrie, des prévisions ou de l'optimisation, sont exécutés en mémoire sur une architecture distribuée très évolutive.

En quoi les offres analytiques hautes-performances sont-elles importantes ?

Avec les offres hautes-performances de SAS, les entreprises ont la possibilité d'analyser les big data, d'en tirer une connaissance plus précise, plus rapidement, et donc de prendre des décisions plus efficaces au moment opportun. Pouvoir résoudre des problématiques complexes, des problèmes jugés jusqu'ici insolubles, tester davantage d'idées, simuler différents scenarii permet de saisir plus rapidement les nouvelles opportunités.

A qui les offres hautes-performances de SAS sont-elles adressées ?

Ces offres sont conçues pour des analystes (data miners, statisticiens, data scientists, analystes métier etc.) et leur offre un environnement pour élaborer et développer des modèles rapidement et efficacement. Grâce à cet environnement, l'IT dispose d'une infrastructure stable, conçue pour supporter les montées en charge, pour gérer et exécuter les traitements analytiques.

Offres SAS hautes-performances

Réalisez plus rapidement des analyses plus précises, et résolvez des problèmes complexes sur de gigantesques volumes de données structurées et non-structurées.

Les offres analytiques hautes-performances permettent de résoudre des problématiques métier complexes qui requièrent des outils d'analyse sophistiqués pouvant accéder à des sources de données extrêmement volumineuses (big data), y compris textuelles. Les obstacles inhérents à l'analyse de gros volumes de données avec les outils de modélisation traditionnels, et aux restrictions imposées par les infrastructures informatiques actuelles, sont éliminés.

Les offres hautes-performances de SAS sont conçues pour exécuter des calculs analytiques complexes en mémoire sur un environnement distribué, permettant ainsi de préparer, d'explorer, de modéliser plusieurs scenarii et d'obtenir un résultat précis et rapide sur de gigantesques volumes de données. Les résultats sont obtenus en temps quasi réel, le délai se chiffre en secondes ou en minutes, et non plus en heures ou en jours.

Une telle réduction du temps d'exécution des traitements analytiques rend possible de tester plus d'hypothèses (what-if). Les modèles sont ajustés et ré-exécutés rapidement. Les analyses prédictives sont d'autant plus enrichies et améliorées qu'il est possible de combiner les données structurées et non-structurées, d'utiliser plus de variables et d'exécuter les modèles plus fréquemment et beaucoup plus rapidement.

Les offres hautes-performances de SAS sont disponibles sur les appliances Oracle, Pivotal/Greenplum ou Teradata ainsi que sur des matériels standards utilisant le système de fichiers distribués d'Hadoop (HDFS) et Cloudera.

Principaux atouts

- **Saisissez de nouvelles opportunités rapidement et en toute confiance, détectez les risques inconnus et prenez les décisions opportunes.** Disposer de résultats plus fins et plus précis procure d'énormes avantages aux entreprises, générant de nouvelles sources de revenus et accroissant leur rentabilité.
- **Utilisez des techniques de modélisation avancées sur l'intégralité de vos données (yc. non-structurées) et exécutez vos modèles plus fréquemment pour répondre à vos problématiques les plus complexes.** L'utilisation d'analyses statistiques avancées sur l'intégralité de vos données permet d'améliorer la précision des résultats et de prendre ainsi des décisions mieux ciblées. La combinaison de données structurées à des données textuelles permet de mettre à jour des relations indécelables jusqu'alors et donne plus de puissance aux modèles
- **Obtenez en un temps record les informations qui vous permettront de valoriser et d'accélérer vos décisions stratégiques.** Diminuez le temps d'exécution des modèles analytiques et obtenez des informations rapides permettant d'améliorer la prise de décision au sein de l'entreprise. Les performances ultra-rapides des offres hautes-performances de SAS permettent de tester de nombreux scénarii alternatifs, de détecter les changements et de proposer très rapidement des recommandations optimales sur les marchés très volatiles.

sas
THE POWER TO KNOW.

- **Tirez parti d'une infrastructure analytique ultra-évolutive et fiable, optimisée pour tester plus d'hypothèses et de scenarii sur l'intégralité de vos données.** L'infrastructure in-memory permet aux analystes de résoudre les problématiques les plus complexes sans contrainte liées à l'architecture et l'IT peut gérer efficacement les requêtes de puissance de traitement supplémentaires immédiatement ou dans le futur.

Présentation de l'offre

L'offre SAS hautes-performances permet aux entreprises d'analyser les big data pouvant provenir aussi bien de référentiels structurés que de collections de documents textes, et de produire au moment opportun des résultats plus précis, en quelques minutes. Chacun peut ainsi prendre de meilleures décisions sur la base de meilleures informations et ce plus rapidement. L'environnement distribué en mémoire permet d'exécuter les modèles plus fréquemment, d'ajouter ou de retirer rapidement des variables et d'utiliser des méthodes analytiques sophistiquées.

Ces offres sont spécifiques pour chaque domaine - statistiques, data mining, text mining, prévision, optimisation et économétrie - et s'exécutent sur des architectures de traitement en mémoire ultra-évolutives. D'énormes volumes de données sont chargés en mémoire très rapidement et les algorithmes analytiques parallélisés s'exécutent directement sur ces données centralisées en mémoire vive.

SAS® High-Performance Statistics

Régression linéaire haute-performance

La régression des moindres carrés ordinaires dans SAS High-Performance Statistics gère la régression pour que les utilisateurs puissent analyser la relation entre une variable expliquée et un ensemble de variables explicatives. Elle comporte de nombreuses méthodes de sélection de modèles et propose différents indicateurs de diagnostic. Parmi ses fonctionnalités exclusives figurent le biais de sélection et l'instruction CLASS qui permet de sélectionner les effets pour les variables catégorielles.

Régression logistique haute-performance

La régression logistique est une méthode de référence pour les prédictions binaires, binomiales et multinomiales. La régression logistique hautes performances inclut des modifications visant à optimiser les algorithmes tout en distribuant le calcul sur la grille. Elle ajuste aussi les modèles de régression logistique aux données binaires, binomiales et multinomiales. Dans un contexte haute-performance, la sélection des modèles est réalisée en quelques secondes ou minutes, ce qui permet aux analystes de tester davantage de variables et de créer des modèles de meilleure qualité.

Modèles linéaires mixtes haute-performance

Ces fonctions permettent d'ajuster différents types de modèles linéaires mixtes aux données ; les modèles ajustés permettent d'obtenir des interférences statistiques. Un modèle linéaire mixte est une généralisation du modèle linéaire standard utilisé dans la procédure GLM – la généralisation signifiant que les variables peuvent présenter une corrélation et une variabilité non contrainte. Par sa souplesse, le modèle linéaire mixte permet de modéliser non seulement les moyennes de vos données (comme dans le modèle linéaire standard), mais aussi leurs variances et covariances

SAS® High-Performance Data Mining

Réduction de variables haute-performance

Cette fonction procède à une sélection non supervisée de variables en identifiant un sous-ensemble expliquant conjointement une variance maximale. La procédure HPREDUCE effectue une analyse de la variance et réduit le nombre de dimensions en sélectionnant le sous-ensemble parmi les variables d'origine contribuant le plus à la variance globale.

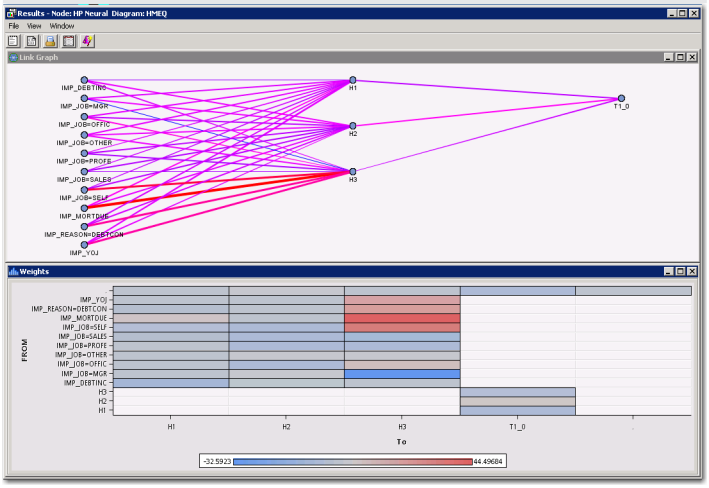
Réseaux neuronaux et forêts aléatoires (fonction expérimentale) haute-performance

Les réseaux neuronaux haute-performance ont été conçus pour que la phase d'apprentissage soit extrêmement rapide et que leur utilisation soit aussi simple que possible. Ce tout en assurant une performance optimale ainsi qu'une bonne généralisation. Ainsi, le paramétrage des réseaux est fait principalement de manière automatique. Les réseaux neuronaux haute-performance tirent parti de l'environnement de calcul parallélisé pour renforcer la puissance prédictive de l'algorithme. La procédure SAS HPFOREST crée un modèle prédictif comportant plusieurs arbres de décision. L'apprentissage et l'exécution de plusieurs centaines d'arbres de décision peuvent se faire en parallèle, indépendamment, sur différents nœuds de la grille. Une nouvelle fonctionnalité permet d'exécuter un score à partir d'un modèle de forêt aléatoire obtenu auparavant après la phase d'apprentissage de la procédure HPFOREST.

SAS® High-Performance Text Mining

SAS High-Performance Text Mining permet d'obtenir très rapidement des informations à partir de larges collections de données non structurées. Ces données peuvent contenir des millions ou des dizaines de millions de documents, d'emails, de notes, de bribes de rapports, de sources provenant des media sociaux etc. Cette offre prend en charge l'analyse des textes, l'extraction des entités, la réduction des termes à leur radical, la détection des synonymes, la découverte de thèmes et décomposition de la valeur singulière (SVD). Les résultats provenant de l'analyse textuelle peuvent ensuite être utilisés dans les analyses de data mining afin d'améliorer la performance prédictive des modèles.

SAS High-Performance Data Mining offre aux analystes la possibilité de réaliser un plus grand nombre d'itérations du réseau neuronal, améliorant ainsi la performance du modèle développé.



SAS® High-Performance Econometrics

Trois procédures de SAS/ETS® ont été transformées pour la haute-performance. La procédure COUNTREG modélise des variables dépendantes constituées de nombres entiers. La modélisation de sévérité ajuste la répartition des probabilités pour la sévérité (amplitude) des événements aléatoires. La procédure haute-performance sur les modèles à variable dépendante limitée permet de construire de modèles avec des variables ayant une plage de valeurs réduite. Les traitements à optimiser sont distribués sur les nœuds qui peuvent eux-mêmes distribuer l'exécution d'une partie de leurs traitements grâce au multi-threading, effectuant ainsi l'optimisation sur le sous-ensemble de données qui lui est affecté et permettant d'énormes gains de performance

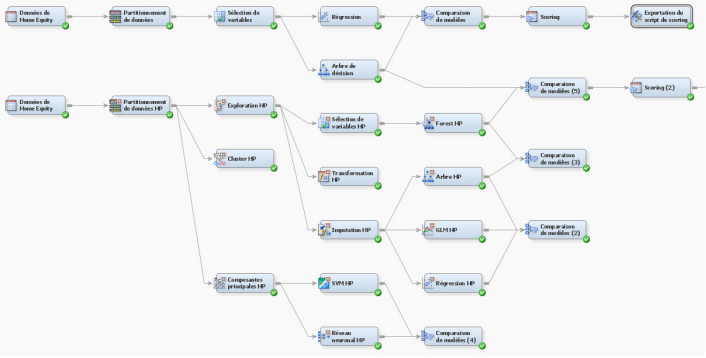
SAS® High-Performance Forecasting

SAS High-Performance Forecasting a été conçue pour traiter efficacement de très larges volumétries de données datées, sous forme hiérarchique. Elle génère automatiquement des séries temporelles et leurs prévisions en une étape en traitant chaque niveau de la hiérarchie en une seule passe.

SAS® High-Performance Optimization

L'optimisation haute-performance s'avère très utile dans la résolution de certains types de problèmes linéaires et mixtes en nombres entiers car elle réduit considérablement le temps nécessaire à l'ajustement des paramètres. Elle permet aussi de réaliser des optimisations locales sur des fonctions non-linéaires et distribue les calculs pour accélérer le traitement des problèmes d'optimisation non-linéaire ayant plusieurs optima locaux.

En utilisant le code L4G de SAS ou bien avec SAS Enterprise Miner, challengez vos méthodes de développement des modèles statistiques, sur l'exhaustivité de vos données et quel que soit le degré de complexité envisagé.



Principales caractéristiques

SAS® High-Performance Statistics

Régression linéaire haute-performance

- Modèles linéaires généralisés et paramétrage de références pour les effets de classe.
- Instruction FREQ pour l'analyse groupée et instruction WEIGHT pour l'analyse pondérée.
- Méthodes de sélection d'effets multiples.

Régression non linéaire haute-performance

- Evaluation des dérivées d'expressions analytiques fournies par les utilisateurs pour obtenir des estimations de paramètres robustes.
- Evaluation des expressions fournies par les utilisateurs et de leur intervalle de confiance.
- Estimation de paramètres avec les méthodes des moindres carrés et du maximum de vraisemblance.

Régression logistique haute-performance

- Prédictions binaires, binomiales et multinomiales.
- Syntaxe de création de modèles avec instructions CLASS et MODEL incluant les effets.
- Modèles de liaison cumulatifs pour les données ordinales et modèles logistiques généralisés pour les données multinomiales non ordonnées ; création de modèles (sélection de variables).
- Instructions WEIGHT et FREQ respectivement pour les analyses pondérées et groupées.

Arbres de décision haute-performance

- Création d'arbres de décisions pouvant être utilisés dans SAS® Enterprise Miner™
- Prise en compte de variables explicatives nominales et continues, et de variables à expliquer nominales.
- Critères de découpe : Gini, entropie et FastCHAID ; algorithme d'élagage type C4.5.

Modèles linéaires généralisés haute-performance

- Estimation des paramètres selon le maximum de vraisemblance.
- Spécification des modèles avec les instructions CLASS et MODEL incluant les effets.
- Multiples fonctions de distribution et de liaison – y.c. les familles de distibution Tweedie.
- Modèles de liaison pour les modélisations logistique ordinale et généralisée.

Modèles linéaires mixtes haute-performance

- Prise en charge de structures de covariance multiples - composants de la variance, symétrie composée, non structurée, AR(1), Toeplitz – et analyse des facteurs.
- Méthodes d'estimation REML et ML et nombreux algorithmes d'optimisation.
- Algorithmes de densité et de dispersion spécifiques mettant à profit les environnements de calcul distribué.

SAS® High-Performance Data Mining

Réduction de variables haute-performance

- Réduction du nombre de dimensions sur les données structurées en entrée et sélection d'un sous-ensemble des variables d'origine.
- Sélection non supervisée de variables par l'identification d'un jeu de variables présentant conjointement une variance de données maximale (analyse de covariance).

- Calcul et affichage distribués de la matrice CORR, COV ou SSCP.
- Utilisation de l'instruction CLASS pour gérer les données catégorielles.
- Prise en compte des effets principaux et des interactions avec l'instruction VAR.
- Génération de statistiques et d'informations matricielles exploitables directement par les procédures statistiques.

Réseaux neuronaux haute-performance

- Standardisation automatique des variables à expliquer et explicatives.
- Paramétrage intelligent par défaut des réseaux neuronaux (ex : fonction d'activation et d'erreur).
- Sélection et utilisation automatiques d'un sous-ensemble de données de validation.
- Interruption automatique de la phase d'entraînement lorsque l'erreur de validation cesse d'augmenter.
- Pondération des observations individuelles.

Forêts aléatoires haute-performance

- Création d'un ensemble de plusieurs centaines d'arbres de décision pour prévoir une variable à expliquer unique.
- Entraînement de plusieurs centaines d'arbres de décision en parallèle, indépendamment, sur différents nœuds.
- Sélection aléatoire des variables explicatives à utiliser pour le découpage d'un nœud parmi toutes celles disponibles.
- Evaluation de la seule variable la plus étroitement associée à celle à expliquer pour le découpage.
- 4SCORE haute-performance : scoring d'un modèle de forêts aléatoires entraîné résultant de la procédure HPFOREST.

Nœuds hautes-performances dans SAS Enterprise Miner

- Les capacités hautes-performances des fonctions présentées sont aussi présentes dans SAS Enterprise Miner, dans les nœuds HP Explore, HP Data Partition, HP Transform, HP Variable Selection, HP Regression, HP Neural, HP Forest, HP Impute et HP Tree.

SAS® High-Performance Text Mining

- Analyse textuelle : détection des différents termes, réduction au radical, détection des synonymes, fréquences et pondération des termes.
- Réduction de la matrice des termes par document à une représentation numérique et structurée grâce à la méthode de décomposition de la valeur singulière (SVD).
- Résultats directement exploitables dans les procédures haute-performance de data mining.
- Informations détaillées et représentations graphiques et tabulaires des termes et de leurs distributions.
- Pondération de la variable expliquée.

SAS® High-Performance Econometrics

Régression de quantile haute-performance

- Ajustement des modèles de régression dont la variable expliquée représente des comptages.
- Modèles : loi de Poisson, loi binomiale négative et loi de Poisson zéro-modifiée – ajustement de régresseurs indépendants pour la loi de Poisson zéro-modifiée.

Modèles de sévérité haute-performance

- Ajustement de la répartition des probabilités pour la sévérité (amplitude) des événements aléatoires.
- Ajustement des modèles de régression à l'échelle des distributions de sévérité.
- Sélection automatique de la fonction de distribution la plus appropriée parmi neuf distributions de probabilités.
- Possibilité d'ajouter des distributions de probabilités supplémentaires.
- Modélisation de la troncature et de la censure de données.

Modélisation qualitative à variables indépendantes limitées haute-performance

- Modèles de régression linéaire, de censure, de régression tronquée avec hétéroscédasticité et modèles de frontière stochastique de production et de coûts.

SAS® High-Performance Forecasting (disponible sur Teradata ou Pivotal)

- Génération automatique, en une étape, de séries temporelles et de leurs prévisions à partir de données datées. Traite les différents niveaux d'une hiérarchie en une seule passe sur les données.
- Pour les séries temporelles standards, la procédure HPFORECAST sélectionne automatiquement la meilleure méthode de lissage parmi les méthodes de lissage simple, linéaire, amorti, saisonnier (additif et multiplicatif), de Winters (additif et multiplicatif).
- Transformations disponibles : logarithmique, quadratique, logistique et Box-Cox.

SAS® High-Performance Optimization

Optimisation locale haute-performance

- Optimisation d'un objectif - défini par l'utilisateur - sous contraintes non-linéaires.
- Intègre le solveur basé sur les algorithmes génétiques de SAS/OR® ainsi que les autres méthodes d'optimisation locale.

Optimisation haute-performance

- Décomposition d'un problème global en sous-problèmes pouvant être résolus plus rapidement.
- Vraisemblance de l'identification d'une solution globalement optimale accrue grâce à la fonctionnalité multi-start.
- Tuning parallélisé permettant un bon réglage des options et de trouver la meilleure combinaison de leurs valeurs.

Eléments communs aux procédures High-Performance

- Les procédures suivantes sont disponibles dans chacune des offres présentées ci-dessus :
  High-performance data summarization (HPSUMMARY), high-performance correlation (HPCORR), high-performance sampling (HPSAMPLE), high-performance binning (HPBIN), high-performance imputation (HPIMPUTE), high-performance DS2 (HPDS2) et high-performance data mining database (HPDMDB).

SAS Institute s. a. s. – DOMAINE DE GREGY - GREGY-SUR-YERRES - 77257 BRIE COMTE ROBERT - FRANCE
TEL. : +33(0) 1 60 62 11 11 FAX : +33(0) 1 60 62 11 99 **www.sas.com/france**