	RTE direction R&D Innovation **& Ecole Centrale de Lille option DAD** **(Décision & Analyse de Données)** **Proposition de projet IMPACT 2014-2015** **« ACP rapide pour des processus d'analyse Bigdata »**	

Sujet : *Tester les performances d'un algorithme stochastique innovant, à convergence exponentielle, de calcul de composantes principales sur des jeux de données importants. La méthode en question est décrite dans une publication récente fournit en annexe de ce sujet. Il s'agit de tester cette méthode et de la comparer, dans le logiciel d'analyse R, à l'algorithme standard disponible et ceci sur un jeu de données réelles fourni par RTE.*

Contexte et enjeux :

L'analyse en composantes principales (ACP) ou « Principal Component Analysis (PCA) » en anglais, est une technique devenue relativement standard dans les processus d'analyse de données. Que ce soit à des fin pratiques de réduction de dimension où pour se concentrer sur un jeu restreint de données à « information maximale » au sens de la corrélation, l'ACP est une brique ayant vocation à s'insérer dans des processus d'analyse et de traitement de volume de données pouvant être très importants comme ceux que génèrent le système électrique et le réseaux de transport géré par RTE.

L'ACP se fonde sur l'analyse des éléments propres, et donc leurs calculs, de la matrice de corrélation des données à analyser. L'enjeu est la maitrise de ces phases de calculs au « meilleur » coût d'autant plus lorsque l'on est contraint en temps d'exécution (dans le cadre de processus opérationnels en exploitation dans les métiers de RTE) et que la taille des données est critique (jeux de données pluriannuels, sur plusieurs milliers à dizaines de milliers de variables, au pas minute par exemple).

On trouvera dans la littérature des références à des techniques de « fast PCA », « Iterative PCA », « Dynamic PCA ». Pour ce projet, on se fondera sur la publication récente (septembre 2014) fournie en annexe de ce sujet : « A Stochastic PCA Algorithm with an exponential convergence rate » d'Ohad Shamir du Weizmann Institute of Science.

On utilisera préférentiellement le logiciel open source d'analyse de données R et RTE fournira comme cas d'application un jeu de données de consommations électriques élémentaires et/ou de productions sur le système français. Des essais sur données simulées pourront être menés préalablement éventuellement.

Encadrant RTE :

Samir ISSAD
Ingénieur Responsable d'études / Engineer
RTE - R&D Innovation
Immeuble LE COLBERT
9, Rue de la Porte de Buc , BP 561
78005 VERSAILLES CEDEX
samir.issad@rte-france.com
Tel : +33 (0)1.39.24.40.16
Fax : +33 (0)1.39.24.41.75