Syllabus projet Spark batch

Sujet:

Vous travaillez pour l'entreprise Data's Madness, spécialisée dans le data engineering. Un nouveau client vient d'arriver et on vous demande de vous occuper de lui. Il s'agit de l'entreprise Untel, spécialisée dans la vente de matériel informatique sur internet, qui possède des données mais qui ne sait pas comment les traiter. Votre rôle est de développer un job spark permettant de répondre à leurs problématiques.

Ils possèdent trois datasets. Un dataset de clients, un dataset de ventes et un dataset de produits.

Ils souhaiteraient avoir un reporting sur la vente de leurs produits ainsi que sur leurs clients.

Quels sont leurs plus gros clients ?
Quels sont les produits qu'ils vendent le plus ? Qui rapportent le plus ?
Tout autre proposition de reporting est intéressante ! Proposez d'autres utilisations de leurs données. Le format de restitution est libre : Fichiers CSV, visualisation graphique...

Organisation

Le projet suivant sera à réaliser par groupe de trois étudiants (choix des groupes libre).

Consignes

Vous devez restituer votre travail au travers d'une présentation de 15 minutes maximum détaillant les étapes de votre démarche et du traitement que vous proposerez. Le code du traitement sera à envoyer avant le vendredi 7 février 2020 23H59 à l'adresse nicolas.martinr.p@gmail.com, avec comme objet [Projet Spark] + vos noms. Le mail contiendra aussi une description des données de sortie de votre projet.
Vous serez évalués à la fois sur la méthode de traitement des données mais aussi sur l'application des bonnes pratiques de Spark.