kadaoui_alexandre_Hadood_Spark

Alexandre

23/12/2020

R Markdown

Dans le cadre de notre partiel, nous devons réaliser un total de 12 travaux retracant notre parcours et notre travail durant les 30 heures de cours.
Le travail à faire est le suivant :
- Une entête comportant un titre, un lien Github avec le ou les noms des auteurs.
- Une synthese de ce travail
- Un extrait commenté avec des parties de codes clé avec explication et commentaire.
- Une évalutation du travail avec nos 5 criteres.
- Une conclusion du travail

Definition des 5 critères de notations :

1) Effort de présentation :
2) Le knit est réalisable et bien présenté.
3) Explications simples et efficaces.
4) Le Code reproductible à d'autres DataFrame avec facilité.
5) Description des fonctions utilsés et du raisonnement.

Hadoop et Spark

Travail réalisé par "Florine Comlan / Ramya Hountondji" le 18/11/2020.

https://github.com/fcom-stack/PSBX (https://github.com/fcom-stack/PSBX)

Synthese :

Le PDF nous présente ici deux frameworks utilisé dans la gestion, le stockage et la manipulation de volumes de données extrêmement importants que l'on peut qualifier de "big data"
Ces deux frameworks sont complémentaires
Là où Hadoop aura comme bénéfice principal un stockage de donnée efficace à moindre coût via son utilisation et sa gestion de clusters de serveurs, Spark sera lui plus à même de traiter et manipuler rapidement ces quantités monumentales de données distribuées.
Le PDF présente ensuite une liste complète des diverses commandes d'installation et de gestion des outils Hadoop et Spark.

Evaluation du travail :

1) Effort de présentation : Le PDF est très bien mis en page, les divers schémas et images apportent une excellente visibilité sur les concepts.
2) Le knit est réalisable et bien présenté : Le knit ne pose aucun soucis à réaliser ici.

3) Explications simples et efficaces : Bien que techniques, les explications concernant les fonctionnements d'Haddop et Spark sont claires, simplifiées et structurées. Tout est expliqué de manière à ce qu'un novice puisse comprendre sans trop de soucis l'utilité des frameworks, leurs différences et leur complémentarité pour diverse tâches.
4) Le Code reproductible à d'autres DataFrame avec facilité : Le code mis en avant dans la dernière partie du PDF est clair et bien présenté afin de faciliter sa réutilisation en tant que documentation.
5) Description des fonctions utilsés et du raisonnement : Il n'est ici pas réellement question des fonctions ou de raisonnement mais plus d'une présentation de ces deux frameworks, leurs points forts, leur complémentarité et leurs contextes d'utilisation.

Conclusion :

En conclusion ce PDF très bien réalisé permet de mettre en lumière l'infrastrucutre nécessaire au stockage et à la gestion de données de l'ordre du Big Data. En présentant non seulement ces deux frameworks mais également le contexte qui les entoure (description du fonctionnement des clusters par exemple), les auteurs parviennent à expliquer simplement et de manière complète la nécessité de ces framework dans le traitement et stockage de big data ainsi que leurs forces et l'importance de leur complémentarité pour des tâches différentes, hadoop pour le stockage de ces données et Spark pour la gestion et la manipulation de celles-ci.