Spark Cheat Sheet

by [deleted] via cheatography.com/30326/cs/9053/

SparkSQL - Transformations

`df = sqlCon tex t.c rea teD ata Fra me( data)`	Créer un DataFrame à partir d'une collection Python (liste)
`df = sqlCon tex t.c rea teD ata Fra me( data, ['name', 'age'])`	Créer un DataFrame à partir d'une collection Python (liste)
`df = sqlCon tex t.r ead.te xt( 'fi les.txt')`	Crér un DataFrame à partir d'un fichier
`ageCol = people.age`	Créer un DataFrame à partie d'une colonne
`df.sel ect ('*')`	Sélect-ionner toutes les colonnes
`df.sel ect ('n ame', 'age')`	Sélect-ionne 1 ou plusieurs colonnes
`df.sel ect (df.name, (df.age + 10).al ias ('a ge'))`	Sélect-ionner 2 colonnes, changer la valeur d'une colonne et renommer la colonne
`df.dro p(d f.age)`	Suppre-ssion d'une colonne. Retourne un nouveau DataFrame

SparkSQL - Transformations (cont)

`df.sel ect (ex plo de( df4.in tli st).al ias - ( a nine )`	Chaqu élémen est dans une nouvel ligne
`df.gro upB y(d f.n ame ).a gg( {'*': 'count '}).co ll ect()`	
`df.gro upB y(d f.n ame ).count`	
`df.gro upB y().av g().co llect()`	
`df.gro upB y(' nam e').av g(' age', 'grade ').c ol le ct()`	

SparkSQL - Actions

`df.show(n, truncate)`	Affiche les n premières lignes du DataFrame
`df.take(n)`	Affiche les n premières lignes sous forme de une liste
`df.col lect()`	Retourne les enregistrements sous forme de liste
`df.count()`	Retourne le nombre de lignes dans le DataFrame
`df.des cri be( *cols )`	Calcule les statistiques descriptives des colonnes numériques
`linesD F.c ache()`	Enregistre le DataFrame dans le cache donc pas besoin de réexécuter toutes les transformations et actions. A utiliser si on réutilise souvent le DataFrame.
`aDF.un ion All (bDF)`	Concaténation de deux DataFrame

/!\ ATTENTION /!\

- S'assurer qu'on a assez d'espace dans le programme "driver".pour utiliser `collect()`.
- Ne jamais utiliser `collect()` en production. Préférer `take(n)`.

`lambda a, b : a + b`	Fontion anonyme. 1 expression
`slen = udf(lambda s: len(s), Interg erT ype())`	Fonction lambda ou nommée et type du retour
`inleds DF.f il ter (is Comment`	Retourne un DataFrame dont les lignes respectent la/les condit- ion(s)
`where( fun ction)`	Retourne un DataFrame dont les lignes respectent la/les condit- ion(s)
`df.dis tinct()`	Retourne un DataFrame avec les lignes uniques
`orderBy(cols, *kw)`	Retourne un DataFrame dans l'ordre croissant ou décroi- ssant
`df.sor t("a ge", ascending = False)`	Retourne un DataFrame dans l'ordre croissant ou décroi- ssant

Sponsored by **CrosswordCheats.com**
Learn to solve cryptic crosswords!
http://crosswordcheats.com