**SPARK SQL**

Table des matières

[Chapitre 1 Introduction & Définition 2](#_Toc104278151)

[Chapitre 2 Data-Frame 2](#_Toc104278152)

[Chapitre 3 Data-set 2](#_Toc104278153)

[Chapitre 4 API Data-Frame 3](#_Toc104278154)

[a-Df à partir de fichiers : 3](#_Toc104278155)

[b-Df à partir fu .Class : 3](#_Toc104278156)

[c-Df à partir de structype : 3](#_Toc104278157)

[Chapitre 5 API Data-Set 3](#_Toc104278158)

[a-Ds à partir de fichiers : 4](#_Toc104278159)

[b-Ds à partir de fichiers : 4](#_Toc104278160)

[Chapitre 6 Caractéristiques de Spark SQL 4](#_Toc104278161)

[Chapitre 7 Comparaison entre RDD, Dataframe et DataSet 4](#_Toc104278162)

[Chapitre 8 Architecture 4](#_Toc104278163)

# 

# Introduction & Définition

Le spark sql est une couche supérieure à spark core. C’est utilisé essentiellement pour traiter des données structurées.

Données structurées

RDD

Données non-structurées

Les data frame et data set sont des cursors (comme pl sql).

Les data frame et data set sont des RDD structurées. Les données sont organisée en colonnes.

Les data frame et data set sont immutable (assimilables à une table dans une base de données relationnelles mais distribuées et en mémoire.

Pour faire du spark sql il faut utiliser les **API data set et data frame** qui sont fournies par apach spark sql.

Le principe de fonctionnement de ces deux API repose sur les algorithmes d’optimisation **Tungstene et Catalist.**

**Tungstene** : responsable de la sérialisation et désérialisation (donc gestion de la mémoire) intervient pour compresser les données en mémoire.

**Catalist :** Optimisation des requêtes, génération du DAG.

**NB :** La sémantique formelle est une science en informatique pour prouver que deux algorithmes sont identiques. (Ou 1 algo est faux ou est juste)

**Un data set** : c’est quand le type de données est connu à l’avance (le cas des fichiers csv), fortement typé)

**Un data frame**: C’est data set avec des lignes dont le schéma n’est pas connu, c’est ce qu’on appelle row (un data frame non-typé)

# Data-Frame

Data frame est très typé, peut être créé à partir d’une rdd, d’une table SGBD(ou table hive), d’une ou plusieurs dataframes avec des transformations, disponible en java scala, python, R

Toutes les données sont de type générique **ROW**

# Data-set

Un data set est un data frame fortement typé disponible uniquement en java et scala. (pas de python car il est un langage pas typé alors que les data sets sont très typés)

# API Data-Frame

Il existe trois façon pour obtenir un data frame :

1. Soit à partir de directement de fichiers
2. Soit à partir du .Class java (.Class étant variable fournie par java pour tous les objets implémentés)
3. Soit à partir de structype

## a-Df à partir de fichiers :

1. Pour les fichier csv

**Dataset< ROW> dfFromCsv =SparkSession.read().format(‘’csv’’).option(“header,true”).load(‘’csvfile.csv’’)**

1. Pour les fichiers parquet:

**Dataset <row> dfFromParquet = sparcession.read().parquet(‘’parquetfilepath’’)**

1. Pour les fichiers orc :

**Dataset<row> dfFromORC = sparkcession.read().orc(‘’filepath’’)**

## b-Df à partir fu .Class :

Dataset<row> people = sparkcession.createDataFrame(peopleRDD,person.Class)

People.show() ;

## c-Df à partir de structype :

Dataset <Row> peopleDataFrame = spark.createDataFrame rowRDD , s

Code à revoir sur le cour

***Manipulations de dataframe:***

Dataset<row> selectedColumnsRows = dataFromCsv.select(col(‘’nom\_colonne1’’),col(‘’nom\_colonne2”).plus(1))); // plus 1 ajouter 1 à la colonne 2 (colonne\_2 sont des integer)

Geq (greater or equal) cad >=

Dataset<Row> filtredRow = dffromcsv.filter(col(‘’colonne1’’).geq(13));

On peut convertir une data frame en rdd avec la fonction suivante :

Java rdd<row> rowJavaRDD = dfFromCsv.toJavaRDD() ;

# API Data-Set

C’est exactement la même chose que le data frame sauf qu’on rajoute .asancoder(.Class) à la fin

**Encoder<Person> ancoder = Encoders. bean Person. class**

## a-Ds à partir de fichiers :

**Dataset< Person> dfFromCsv =SparkSession.read().format(‘’csv’’).option(“header,true”).load(‘’csvfile.csv’’).as(ancoder);**

## b-Ds à partir de fichiers :

# Caractéristiques de Spark SQL

Spark SQL est un module de spark.

Il a 3 API : DATAFRAME, DATASET, SQL (SQL LIKE POUR les dataframes)

# Comparaison entre RDD, Dataframe et DataSet

Voir le cours

# Choix d’utilisation

RDD sont utilisées(quand on n’a pas le choix se sont les traitements de bas niveau) pour les données non-structurées (très moinsoptimisées) par rapport au data frames et encore beaucoup moins par rapport aux datasets. (fonctionnel plutôt que DSL : ) exemple :c hercher des mots dans un texte, media, video, photos, etc..

Data Frame et data sets quand on fait des traitements de haut niveau (fonction aggrégation, sum, etc), sur des données semi structurées ou structurées

Utlisation du langage DSL

# Architecture

