Pyspark Implementation

Problem definition: can figure out which ItemCollection will be available for storing books.

Used by Google Cloud Platform

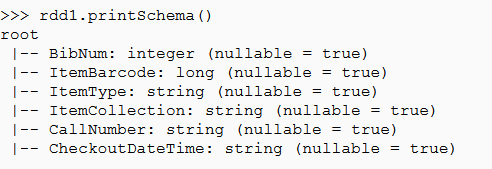
from pyspark.sql import SQLContext, Row, Col

sqlContext = SQLContext(sc)

1. **Read csv files and creates dataframe**

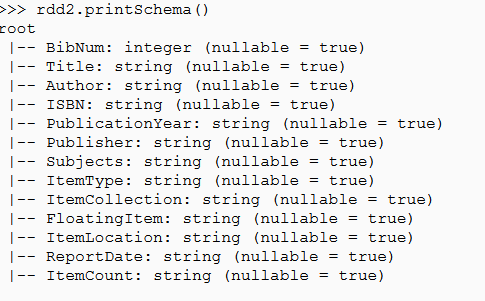
rdd1=sqlContext.read.format(‘csv’).options(header = ‘true’, inferSchema=’true’).load(‘gs://cpsc651-1024474/checkouts/’)

rdd1.printSchema()



rdd2=sqlContext.read.format(‘csv’).options(header = ‘true’, inferSchema=’true’).load(‘gs://cpsc651-1024474/inventory/inventory2.csv’)

rdd2.printSchema()



1. **This allows to check overlapped columns between two dataframe**

common\_cols=[x for x in rdd1.columns if x in rdd2.columns]

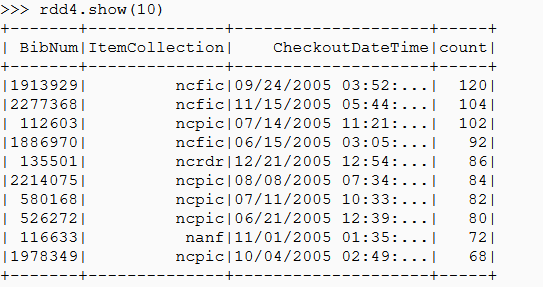
1. **Join each dataframe without overlapped column**

rdd3=rdd1.join(rdd2, on=common\_cols)

1. **Group by BibNum, ItemCollection, CheckoutDateTime to see which location of books are borrowed depending on CheckOutDateTime.**

rdd4=rdd3.groupby([“BibNum”,”ItemCollection”,”CheckoutDateTime”]).count().sort(col(“count”).desc())

rdd4.show(10)



Result : ncfic location has smallest space for storing books.