Apache Spark Book Crossing Analysis

Skills Required: Apache Spark SQL, Apache Hive

Description:

This case study is used to analyze Book Crossing Data set which has details about books, ratings and users.

Input Data

http://www2.informatik.uni-freiburg.de/~cziegler/BX/BX-CSV-Dump.zip

BX-Books.csv
ISBN, Book-Title, Book-Author, Year-Of-Publication, Publisher

BX-Book-Ratings.csv
User-ID ISBN Book-Rating

BX-Users.csv
User-ID Location Age

Requirement

Create tables in hive and store the data related to Books, Book-Ratings and Users in Parquet format.
Use Apache Spark SQL to query the hive tables and perform the following analysis.
Store the output of the analysis into hive tables.

a. Find the most popular Author among each of the following age groups:
- less than 10 years
- 10 to 18 years
- 19 to 35 years
- 36 to 45 years
- 46 years and above

The most popular author is one who got highest number of ratings >= 6

b. Find the Most Popular Author in each Country

c. Find the state in each country which has the highest number of readers