**درس : Big Data استاد : خانم دکتر صفری**

**دانشجو : رقیه یزدانی تمرین : Spark DataFrame**

# Spark DataFrame

# ایجاد بستر لازم برای امکان کار با Spark

مشابه تمرین قبلی spark، cloudera vm را start کرده و سرویس‌های لازم را بالا می‌آوریم. ابتدا یک ترمینال باز کرده و دستور spark-shell را اجرا می‌کنیم، می‌توانیم version مربوط به spark و همینطور scala را ببینیم:

**Spark 1.6.0**

**Scala 2.10.5**

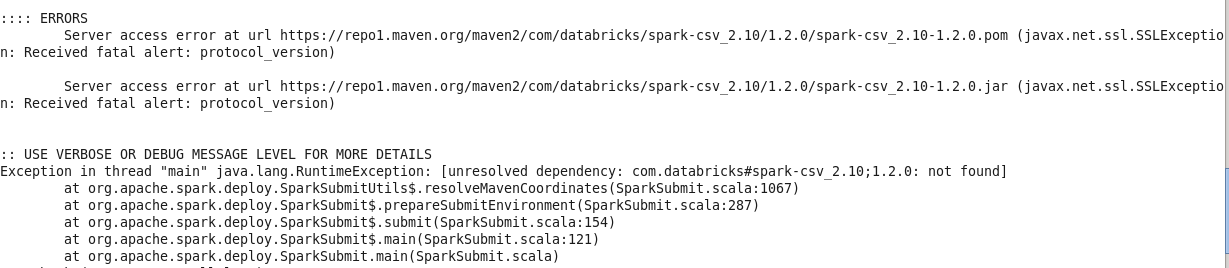


حال در سایت [https://spark-packages.org](https://spark-packages.org/package/databricks/spark-csv) نسخه پکیج spark-csv را که متناسب با نسخه scala بدست آمده در بالا باشد، انتخاب می‌کنیم. برای کار با csv لازم است که هنگام اجرای pyspark، پکیج مربوط به csv را هم import کنیم. برای اینکار دستورات زیر را اجرا می‌کنیم. در دستور دوم، پس از مشخص کردن نام package، عدد اول (2.10) ورژن scala و عدد دوم (1.5.0) ورژن spark-csv می‌باشد.

$ PYSPARK\_DRIVER\_PYTHON=ipython

$ pyspark –packages com.databricks:spark-csv\_2.10:1.5.0

پس از اجرای دستورات بالا خطایی به شکل زیر تولید شد :



برای رفع این مشکل که به نظر می‌رسد مربوط به نسخه TLS در پروتکل SSL می باشد، پس از جستجو در نت، بصورت زیر عمل شد. ابتدا متغیر زیر به فایل .bashrc اضافه شد و پس از لود آن، مجددا دستور بالا اجرا شد که این بار مشکلی نداشت.

export JAVA\_TOOL\_OPTIONS="-Dhttps.protocols=TLSv1.2"

s

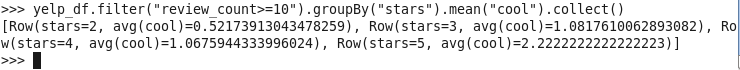
حال می‌توانیم فایل csv را لود کنیم، پس از لود کردن فایل، با استفاده از فراخوانی تابع printSchema ساختار داده اطلاعات لود شده، نمایش داده می شود که شامل نلم ستون؛ نوع داده آن و nullable بودن یا نبودن می‌باشد.



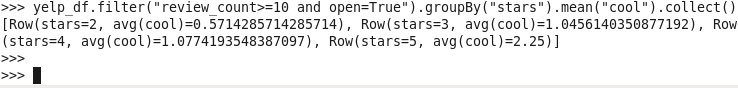
**سوال 1 :** برای یافتن میانگین ستون cool بصورت زیر عمل می‌کنیم. با توجه به نتیجه؛ گزینه 3 صحیح است.



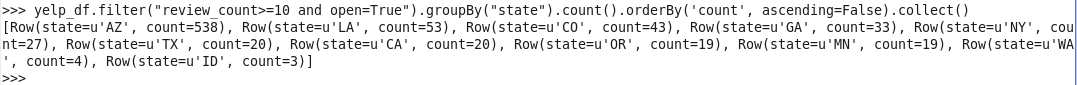
**سوال 2 :** برای این سوال، ابتدا با تابع filter، سطرهای دارای شرط review\_count>=10 را جدا کرده و سپس روی فیلد stars گروه بندی کرده و برای هر گروه میانگین cool را محاسبه می‌کنیم. با توجه به نتیجه؛ گزینه 3 صحیح است.



**سوال 3 :** برای این سوال، ابتدا با تابع filter، سطرهای دارای شرط review\_count>=10 و open=True را جدا کرده و سپس روی فیلد stars گروه بندی کرده و برای گروه cool، میانگین را محاسبه می‌کنیم. با توجه به نتیجه؛ گزینه 4 صحیح است.



**سوال 4 :** برای این سوال، ابتدا با تابع filter، سطرهای دارای شرط review\_count>=10 و open=True را جدا کرده و سپس روی فیلد state گروه بندی کرده و برای هر گروه count، را محاسبه می‌کنیم. سپس نتیجه را بصورت نزولی بر اساس count مرتب می کنیم. با توجه به نتیجه؛ گزینه 2 یعنی CO با 43 تا review جایگاه سوم را دارد و صحیح است.



**سوال 5 :** برای این سوال، ابتدا با روی ستون business\_id گروه‌بندی کرده و به ازای هر گروه؛ تعداد سطرها را بدست ‌آورده و سپس در مرحله آخر، ماکزیمم تعداد سطرها را بدست می‌آوریم. با توجه به نتیجه؛ گزینه 4 صحیح است.

