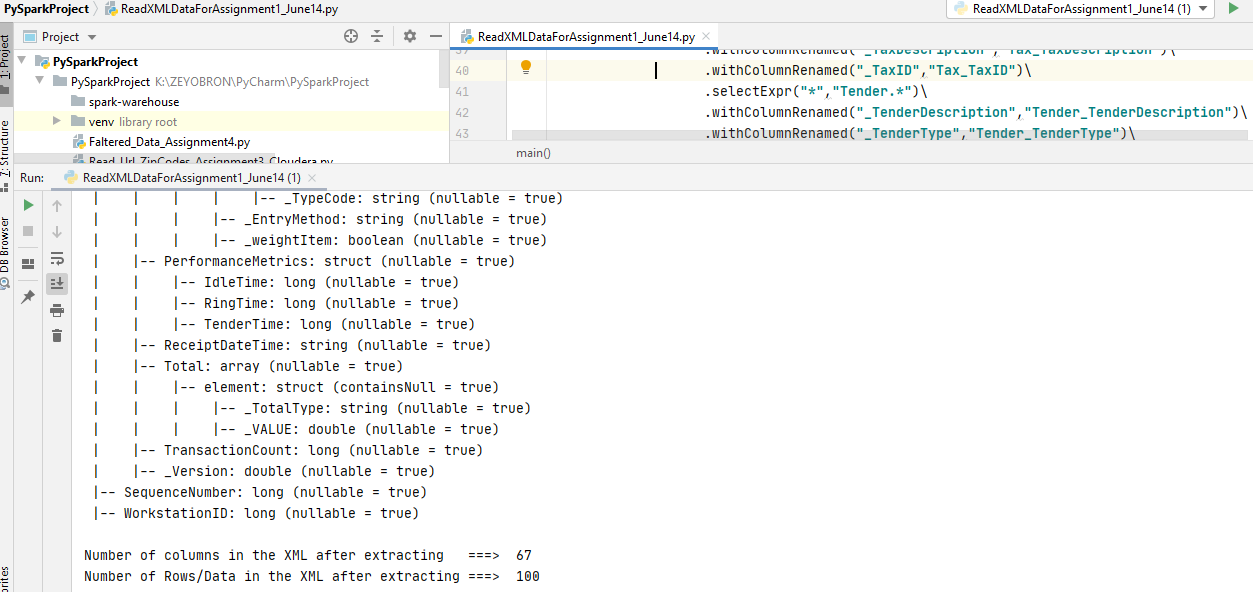
**Task 1**: **--- Pyspark Do all seamless read and writes with different file formats**

****

**Output:**

****

**Task 2 -- Hit results url,**

**Convert that to pyspark rdd**

**Convert that to dataframe**

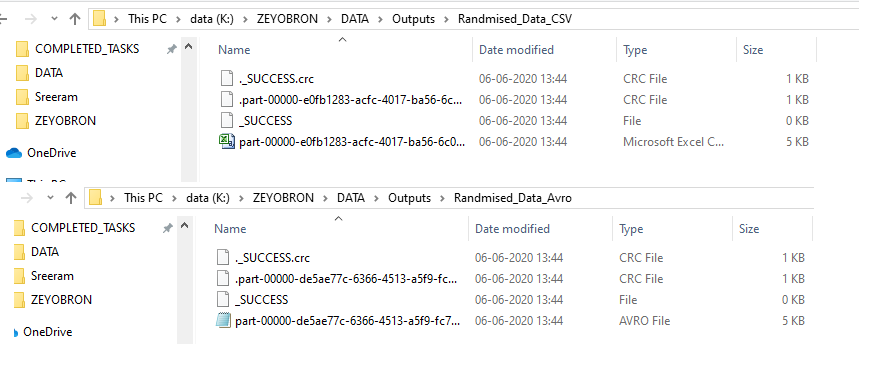
**Usernames --- We removed Numericals**

**Write it to avro in Cloudera**

**Write it to Json in Cloudera**

****

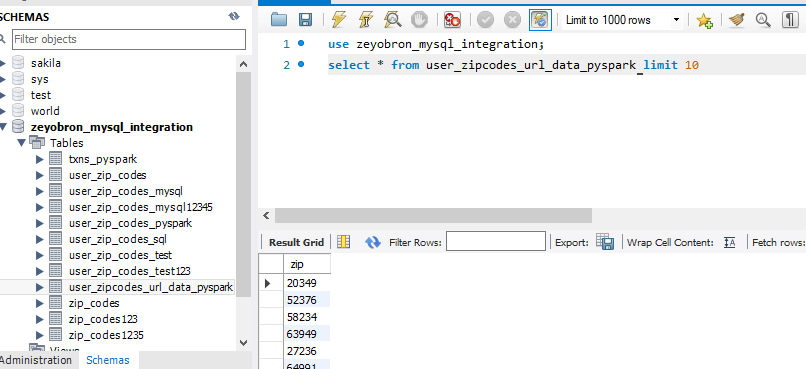
**Output**

****

**Task 3 --- Pyspark Zip codes**

**Every run I need 100 zip codes in RDBMS Table**

****

****

**Task 4 --- Pyspark Row column file split**

**Pyspark Row column file split for PractitionerLatest**

**PROV\_TIN PROV\_PH\_DTL PROV\_ADDR\_DTL**

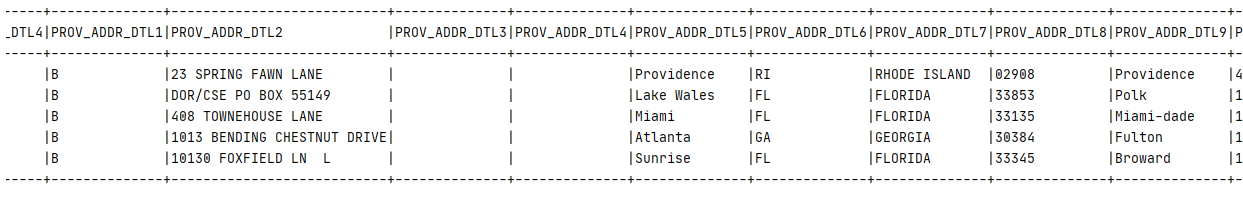
**050460690 B|8888023776||Y SPRING FAWN LANE|||**

**050460690 S|||Y B|23 2203 BECK AVE APT D7|||**

**PROV\_TIN PROV\_PH\_DTL1 PROV\_PH\_DTL1 PROV\_PH\_DTL2**

****

**Output**

****