**CSEE5590-0001/490-0003: Big Data Programming**

**Lesson Plan # 3**

**Lesson Title:** *Data Frame and SQL*

**Lesson Description:** *Distributed Collection of Data*

**Lesson Overview:**

* Data frames
* Construction of Data Frames
* SparkSQL
* Transformation
* Laziness
* Actions
* Basic Commands on Data frames
* Basic commands of SQL on Data frames

**In Class Exercise**

**Dataset:** [**https://umkc.box.com/s/tg08jqi9circsjpycwxoujnux85kbxw6**](https://umkc.box.com/s/tg08jqi9circsjpycwxoujnux85kbxw6)

**DataFrames & SQL in Scala/Pyspark**

Consider the dataset attached above:

References:

<https://stackoverflow.com/questions/51689460/select-specific-columns-from-spark-dataframe>

<https://jaceklaskowski.gitbooks.io/mastering-spark-sql/spark-sql-aggregate-functions.html>

How to read CSV into dataframe:

<https://stackoverflow.com/questions/29704333/spark-load-csv-file-as-dataframe>

**Part – 1:**

1. Import the dataset and create data frames directly on import.
2. Save data to file.
3. Check for Duplicate records in the dataset.
4. Apply Union operation on the dataset and order the output by Country Name alphabetically.
5. Use Groupby Query based on treatment.

**Part – 2:**

1. Apply the basic queries related to Joins and aggregate functions (at least 2)
2. Write a query to fetch 13th Row in the dataset.

Part –3:(bonus)

1.Write a parseLine method to split the comma-delimited row and create a Data frame.

**ICP Guidelines (In Class Students):**

1. ICP Submission is in pairs of two students.
2. Once completed, must be presented to TA or Instructor before the completion of the class
3. Submission after class is considered as late submission. (Check the late submission policy in the syllabus)
4. ICP Code with brief explanation in wiki should be pushed to GitHub (submit your screenshots as well. The screenshot should have both the code and the output)Submit GitHub Link through the Feedback Form( <https://forms.gle/1Z6WUmrktrzp7FGn7> **)**

**Submission Guidelines (for online students):**

1. Submit your source code and documentation to GitHub and represent the work through wiki page properly with detailed explanation (submit your screenshots as well. The screenshot should have both the code and the output)
2. Comment your code appropriately.
3. Submit a brief demo video 2-3 min showing your assignment with a voice over explaining your work through the Submission Link.
4. Use the following Google link to submit your assignment

(ICP Submission Link#):

( <https://forms.gle/1Z6WUmrktrzp7FGn7> **)**

***Cheating, plagiarism, disruptive behavior and other forms of unacceptable conduct are subject to strong sanctions in accordance with university policy. See detailed description of university policy at the following URL:*** [*https://catalog.umkc.edu/special-notices/academic-honesty/*](https://catalog.umkc.edu/special-notices/academic-honesty/)