Virginia Tech – Office of Academic Decision Support

Data Engineer (AP0190104)

Presentation Assignment

Data Processing Task Requirements

1. You are provided with NBA dataset (nba.csv). Identify the issues with the data and develop a script to address those issues. Also, develop a script to aggregate the data for comparing NBA team salary.

1. You are provided with the patient visits data (patients.csv) from the hospital. Identify the issues with the data and develop a script the clean the data.
2. You are provided with the Ebola cases data (ebola.csv). Develop a script to extract useful textual information from the indicator, and transform the data into a useful form. Specifically, you need to extract the indicator class, such as 'Cumulative Cases', 'Cumulative Deaths', 'Number Cases 21', 'Number Cases 07', 'Proportion Cases 07', 'Proportion Cases 21', etc., and indicator sub class, like “All”, “confirmed”, “probable”, “suspected”, etc. The final output table should have the following fields:

* Indicator\_class
* Indicator\_sub\_class
* Country
* Date
* Value

Notes:

1. Please use Python or SQL to complete the data processing tasks.
2. You may use Jupyter or any other notebooks to develop and demo (present) your code detailing each processing steps.
3. If you have AWS personal account, you can develop and demonstrate the data processing pipeline work in AWS environment.
4. Please contact Daniel Hung at [chung@vt.edu](mailto:chung@vt.edu) for any questions related to this assignment.