Assignment:

Download the data from the URL:

https://archive.ics.uci.edu/ml/datasets/Educational+Process+Mining+%28EPM%29%3A+A+Learning+Analytics+Data+Set

Refer to the same above URL for understanding the data and its features.

Perform the following tasks:

1. Save the data from multiple files present in the data source directory 'Processes' into HDFS file/files. Use of any technology (HDFS commands, shell script, Java, Spark etc.) is allowed for preprocessing if required. Justify your choice.
2. Read the above HDFS file/files in a Spark dataframe.
3. Assign a unique key to each record and call the corresponding column as 'record_id'. Store the data with a unique record_id in a hive table named 'processes_source'
4. Check if start_time and end_time are in the appropriate datetime format of dd.MM.yyyy HH:mm:ss. Erroneous record_ids, erroneous column name and value should be logged in a separate hive table named 'error_records_log'. Example is given below

record_id	column name	value
1111	start_time	10.09.22 00.0.123
2122	end_time	10.09.2222 11.12.1111

5. Similarly read the data from files final_grades.xlsx and intermediate_grades.xlsx in Spark.

6. Create a domain in Java to hold the data of each student. Appropriately design the fields of the required classes. The object of the top level Java class should hold all the information about a single student. The details of the information to be stored in this class are as follows:
    a. Student id
    b. List of session wise process details
    c. List of intermediate grades for each session
    d. List of final grades

7. Create an HBase table with a single column family. Insert the data about each student in the HBase table.