Big Data Hadoop & Spark Exam

[Time: 3.5 hrs]

[Total Marks: 100]

Given is the boston.csv dataset with the following variable information:

CRIM - Per Capita crime rate
ZN - Proportion of residential land zoned for lots over 25000 sq. ft
INDUS - Proportion of non-retial business acres
CHAS - Charles River dummy variable (1 - if tracts bounds river, 0 -otherwise)
NOX - Nitrogen Oxide concentration
RM - Average number of rooms per dwelling
AGE - Proportion of owner-occupied unit built prior 1940
DIS - Weighted MEan of distances of five Boston Employement Centres
RAD - Index of accessibilities to Radial highways
TAX - Full-value-property-tax rates per $10,000
PT - Pupil-teacher Ratio
B - the proportion of blacks
LSTAT - Lower Status of the Population (%)
MV - Median Value of homes (Target Variable)

		Marks
Q.1	Read the given CSV file in a Hive table	**[20]**
	Perform the following tasks using *PySpark*	
Q2.	Read the data from Hive table as spark dataframe	**[15]**
Q3.	Get the correlation between dependent and independent variables	**[20]**
Q4.	Build a linear regression model to predict house price	**[25]**
Q5.	Evaluate the Linear Regression model by getting the RMSE and R-squared values	**[20]**