CASE STUDY

Scenario 1

Problem Statement: Danny has joined an IT company and they have going to conduct an assessment on BIGDATA which may have the questions below. He is not sure as to how to answer these.

You are an expert in Big Data Analytics and has come to for help. Help him with his queries.

Cases:

1. load the given textfile in HDFS. [10]
2. Perform WordCount on the text file using mapreduce.[10]
3. Create a HBase table 'Census' using java with Column Family as 'Personal', 'Professional'. [10]
4. Put 2 rows in the Census table each having columns name and gender in personal and occupation in professional and display data using HBase shell.[10]
5. Load the groceries data file using PigStorage in Grunt shell with a schema and describe and display the data.[10]

The files for 1st scenario:


groceries.csv


words.txt

TECHADEMY

Scenario 2

Problem Statement: movierots has a dataset that contains the list of movies and they have approached you to process and analyze the data.

Solution: The best practice would be to load the data in HDFS and using apache spark to analyze and process the data.

Cases:

1. Create a RDD from the movies.csv file. [5]
2. Print the first 5 rows from the RDD. [5]
3. Print the data from the file and also count the number of records.[10]
4. Create a dataframes that takes the csv file as input and 1st row as header and print the schema.[10]
5. Write a SparkSql query to print all the records from the dataframe.[10]
6. Write a sparkSQL query to print the movies that released in the year 2011 and store the result in another dataframe. [10]

The file for the second scenario is attached below.

movies.csv