* **Week 6 Paper 1 - Working with SQL** *Please click on the link above to submit this week's assignment.* Assume that you have the following tables. **Employees table -** Contains a list of employees

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| ID | FirstName | LastName | Age | Dept\_number | City | State | Salary |
| 1 | John | Smith | 45 | 100 | Chicago | IL | 1000 |
| 2 | Jane | Doe | 25 | 100 | Phoenix | AZ | 5000 |
| 3 | Mary | Smith | 40 | 200 | Chicago | IL | 2500 |
| 4 | George | Edwards | 50 | 300 | Phoenix | AZ | 3000 |

* **Departments table** - Contains a list of Departments in the company

|  |  |
| --- | --- |
| Dept\_number | Dept\_Name |
| 100 | HR |
| 200 | IT |
| 300 | Accounting |

* Write a 1-3 page paper providing the SQL statements and the expected/actual results for the following queries.    NOTE:  You are not expected to install any database software for this assignment. However, you can use an existing DB to verify your queries.  NOTE: If you are not familiar with SQL, go to the SQL Basics links provided.
  1. Display the first name, last name, department name and age of all employees. Note you must display the Department Name and not the department number.
     1. Display the same information ordered by Last Name
     2. Display the information in decreasing order of Age
  2. Display all details for employees that are not from Chicago
  3. Display all details for employees with age 40 or over
  4. Calculate and display the total salary of all employees
  5. Add a new employee with the following information:
     1. FirstName: Jane
     2. LastName: Smith
     3. Age: 45
     4. Department: Accounting
     5. City: Chicago
     6. State: IL
     7. Salary: 5000
  6. Increase the salary of John Smith by $2000. Note that this is to add-to and not to replace the existing salary amount.
  7. Compute and display the total salary by City
* **Requirements for the written assignments:**
  1. Assignment file must have a .doc or .docx extension; screen shots should be in .jpg, .gif, or .pdf
  2. Points for this assignment = 15

* [**Week 6 Assignment 1 - Understand the data**](https://lewisuniversity.blackboard.com/webapps/assignment/uploadAssignment?content_id=_3830072_1&course_id=_178844_1&group_id=&mode=view) *Please click on the link above to submit this week's assignment.* In this assignment, you will perform the steps that are required before actually running queries in a cluster. You will end up saving time and money by first understanding the data, the relationships and the information that is required to be generated from the data.
  1. Download the the data used for this assignment, from this URL  <http://www.seanlahman.com/baseball-archive/statistics/>
  2. Unzip the file and then locate the "readme" file inside the "core" folder to understand the data.
  3. Draft and submit a document containing answers to the following.
* (Q1) Use the information from the readme file to answer the following questions:
  + 1. Which CSV file/table would you use to determine the total number of players?
    2. Which CSV file/tablewould you use to determine a player such as Derek Jeter's  salary for the year 2010?
    3. Which CSV file/tablewould you use to determine the player's date of birth and country of birth?
    4. Which CSV file/tablewould you use to determine whether the player was inducted into the Hall of Fame?
    5. Which CSV file/table would you use to determine the name of the team a player played in, in the year 2000?
    6. Which CSV file/table would you use to determine the number of home runs scored by a player such as Derek Jeter in 2010?
    7. Which CSV file/table and which column would you use to check if the player is still alive?
* (Q2) Provide the Hive query language (HQL) commands for the following.  You will run these HQL commands in the Azure cluster in Week 6 Assignment 2.  Provide the names of the data file(s) that you will need to include in the queries and the results that you expect to get from the queries.  NOTE :  The HQL commands look similar to  SQL queries.
  + 1. What is the total number of baseball players?
    2. How many players were born in the year 1960 and earlier?
    3. How many players were born in the USA?
    4. How many players were born outside the USA?
    5. Display the number of players born in each year starting from 1960 thru 2000.  For example, the output should show:   1980  4    ( where 4 is the number of players born in 1980)
       1. Use the Group By and Order By and count(\*)
    6. How many players and managers were inducted into the Hall of Fame?
    7. Provide a list of all players for any team and from any year. For example, print the list of players who played for Chicago Cubs in 2000.
* **Requirements for the assignments:**
  1. The assignment and write up are due by end of Week 6 .
  2. Assignment file must have a .doc or .docx extension; screen shots should be in .jpg, .gif, or .pdf format
  3. Points for this assignment = 30

* [**Week 6 Assignment 2 - Analyze the data**](https://lewisuniversity.blackboard.com/webapps/assignment/uploadAssignment?content_id=_3830073_1&course_id=_178844_1&group_id=&mode=view) *Please click on the link above to submit this week's assignment.* In this assignment you will be running the HQL queries with the baseball data.  Your options are:
  1. VirtualBox (you can use any one of the following)
     1. Bitnami Hadoop (Difficulty level : High because of command-line use only )
     2. Cloudera CDH    (Difficulty level : Medium)
  2. Azure HDInsight    (Difficulty level : Easy)
* Instructions:
  1. Create the environment for Hive
     1. For Azure HDinsight, follow the instructions given in this URL to create the HDInsight cluster and run a simple query: <https://docs.microsoft.com/en-us/azure/hdinsight/hadoop/apache-hadoop-linux-create-cluster-get-started-portal>  (NOTE:  you may choose just 2 worker nodes instead of 4 worker nodes to minimize the hourly cost).
        1. Use Ambari views to create tables from the CSV files containing the baseball data
     2. With Bitnami Hadoop, use the Hive command or Beeline command to create tables and then loading data using the CSV files
     3. WIth Cloudera Hadoop, use HUE  to create tables from the CSV files containing the baseball data
  2. Create the Hive tables and then run the Hive HQL queries that you wrote in Week 6 Assignment #1. Capture the results of the query.
  3. If using CLI, don't forget to exit out of the Hive CLI and close the server connection
  4. Terminate the Azure HDInsight cluster to avoid using up the available credit