# Assignment 1

## Discussion Questions (5 points)

Post your response to Week 1 discussion board.

1. Share your past experience with Hadoop, either previous coursework, applications that you've built, or any other self-directed learning with Hadoop or Hadoop-related projects. What are you most hoping to learn from this course?
2. Describe the data warehousing technologies that you are using currently (or have used in the past), and the use cases, performance, data magnitude/scale, and data access requirements you have for your data warehouse. Does your data warehouse serve both OLTP and OLAP use cases?
3. What kinds of data warehouse applications are suitable for Hive? Why is Hive not a replacement for a relational database management system?

## Assignment (15 points)

The Resources section contains a set of CSV data containing batting and pitching statistics from 2012, plus fielding statistics, standings, team stats, managerial records, post-season data, and more[1]. Included in the ZIP file is a README (readme 2012.txt) that describes the data contained in each of the CSVs.

(5 points for creating the database/tables and loading the data.)

Create a database called 'baseball_stats' and in that database create Hive tables to hold the data in Masters.csv, which contains a master list of players and their information, Teams.csv (Teams data), Batting.csv (batting statistics) and Salaries.csv (salary statistics). Load the CSV data into the tables you created, and implement the following queries:

[1] Lahman's Baseball Statistics: http://seanlahman.com/baseball-archive/statistics/

1. List of top 50 players by highest number of home runs in a season.  List the players by Player ID, First Name and Last Name, number of home runs and season year. (5 points)
2. List average salaries for each team in 2012.  List the teams by Team ID, League ID, Team Name, and average salary amount in descending order. (5 points)

Submit your CREATE TABLE statements, query statements, and results for both queries.