**Big Data Hadoop—Real Time Project—Insurance**

**Analyze health reports across years for the US market**

A US-based insurance provider has decided to launch a new medical insurance program targeting various customers. To help this customer understand the current realities and the market better, you have to perform a series of data analytics tasks using Hadoop. The customer has provided pointers to the data set you can use. The following are the goals for the project:

* Calculate the average % of people, aged between 18 and 64, who have obtained insurance from private players from 2001 to 2011.
* Calculate the average percentage of people, aged 65 years or more, who are solely covered by public insurance from the year 2001 to 2011.

You can download the required data set in XLS format from the URL below:

<http://www.census.gov/hhes/www/hlthins/data/utilization/tables.html>

Note: You have to download all the data sets of this page to perform your analysis.

The total time provided by your manager to complete this task is 4 hours.

Hint: Use the techniques taught in lesson 03 to load data sets in HDFS. For using Pig/Hive, recall the techniques mentioned in Lesson 07/08.

***This project should be performed using CloudLab.***