Problem Description 2:

The 2009 Data Expo features a data set which consists of commercial flights within the USA from October 1987 to April 2008. The data set consists of nearly 120 million rows and takes up 12 GB [1]. Performing analyses on this data set in R presents a significant challenge due to its sheer size. It also provides an opportunity to explore different techniques for managing large data sets and provides motivation for understanding and implementing parallel statistical algorithms.

In general, there are two approaches taken to manage large data sets in R. The first involves using a relational database. The database is created and an R package facilitates interaction with the database. SQL statements are created in R to retrieve subsets of the database and analyses are performed on these subsets.

The second approach involves using an R package designed to handle large datasets. These packages implement functionality similar to matrix or data.frame objects, with a few caveats. These objects have the advantage of requiring an R user to understand the SQL syntax. However, their caveats must be understood and there is often a pre-processing step required when using them.

The 2009 Data Expo airline data set can be managed using either of these two approaches. Using each of the two approaches to provide solutions to these challenges allows for a direct comparison between the two techniques. The solutions also allow a user to explore the strengths and weaknesses of each approach.

Another difficulty in dealing with large data sets is that statistical algorithms often do not scale well with memory. In general, there are two options to mitigate this problem. The first is chunking. Analyses are performed on chunks of data and each of these processed chunks are combined to give a result similar to that of an analysis performed on the entire data set. A second approach involves parallelizing an algorithm. For a given algorithm, independent tasks are run simultaneously thereby reducing the amount of time taken to run the algorithm.

The challenges provided at the Data Expo web page provide applications for both chunking and parallelization. Answering the question “Do older planes suffer more delays?” can provide an example of creating a linear model in chunks. Answering the question “When is the best time of day to fly to minimize delays?” can provide an example of parallelization.

[1] http://stat-computing.org/dataexpo/2009/