Name: Niko Kortström
Student number: 014154573

Exercise 1

a.

Wikipedia (https://en.wikipedia.org/wiki/Big_data) states that big data means data sets that are too large or complex to be processed by traditional data processing applications. Oracle (https://www.oracle.com/big-data/index.html) uses the four Vs approach: volume, velocity, variety and value that seems to sum the subject up pretty well.

When someone mentions big data, I always think of Google. Their various internet services no doubt collect huge amounts of data. However, big data is by no means limited to large IT companies that often come to mind first. Big data is big because it is collected from so many places.

Most companies that have the possibility are, or at least should be, collecting their own big data for internal usage or selling it. For example, pretty much any application you download to your smart phone asks for permissions to use I/O devices that seemingly have no effect on the operation of the application itself. They are using various sensors and user inputs to form their data. Big data can be pretty much any information collected to the huge remote clouds for processing and analyzing.

b.

i. Find out which destinations were popular in a certain location. Use this data to deduce what travel commercials to show around that location.

ii. Based on which applications and settings were popular, you could develop the next version of the operating system to make usage of these settings and applications easier and faster.

iii. Analyze current interests of people and possibly the English writing patterns dominating at the moment.

Exercise 2

PageRank is an algorithm that estimates web pages' importance based on how many and how valuable links there are to it. The better ranked based site the link is on, the more valuable it is. This algorithm is used by Google to rank websites in results of their search engine. The challenge with this algorithm is probably the really massive amount of data to go through. If you truly iterate over all web pages to find the links, required time must be enormous even with a powerful computing cluster. Therefore I think if this computing needs to be repeated often, it should probably try to sample smaller amounts of pages and try to estimate values based on this.

Exercise 3

a.

With this amount of data there is no chance to fit the data in memory. This causes a lot of memory swapping which slows down computation. Application developers should focus on trying to distribute the data so that you can perform as many operations as possible on the amount of data that can be fit in memory at once.

Most likely many disks are needed to store the data too. I think the data should probably be stored in a way that is preferable for most used operations. Maybe in some kind of ordered way to help iterating.

b.

The limited amount of time to collect the desired data must be the biggest challenge. At least you have to keep the process of acquiring the data simple enough. Another possibility is that you bargain from the accuracy of the data. I think it would be preferable to iterate over the data once at most. Otherwise the time constraints might be too demanding to get too much useful analyzing done.