Project Brief – Big Data

Background and aim
The project is based on a proposal from a small project group working with innovation at the company Cybercom. The aim is to build an infrastructure for big data, analyse it and draw meaningful conclusions. The project will be divided into two similar projects which will be run on different platforms to explore their different possibilities and advantages. The platforms will be IBM open platform and Cloudera.

Project “Fallolyckorna” where data from hospital environments will be analysed is placed on the IBM open platform. The aim of the analytics is to be able to predict and reduce fall accidents. This project will be referred to as **Healthcare**.

The second project is called **Virtual Assistance System**. It will provide additional functionality to an existing product made by Cybercom. The product is called ‘Machine Book’ and is an application for showing and visualising data. This project aims to add analytics to extract new data and insights from datasets of this product.

The concept of big data is a new venture for Cybercom, and as such could lead to the development of future products to solve a variety of issues and to add increased value propositions to customers.

Solution
Both projects aim to transfer a dataset to its chosen big data platform for the purpose of analysing it. Two likely issues and their solutions have been identified:

- Hardware limitations
- Analysis selection

Both platforms used in the project require the use of a main computer with sizeable amounts of primary memory. Additionally, because the distribution of workload and storage is an integral advantage of a big data system, the analysis of the datasets should preferably be performed in a distributed fashion on a larger set of computers, called a cluster. Cybercom will supply the necessary hardware.

The two projects ultimately intend to produce new sets of data and to reach conclusions conducive to their respective stipulated objective. This hinges on the ability to choose appropriate analysation methods, and to apply them to the correct subsets of data. Success will depend on to which extent the group has familiarised themselves with the available software and tools of the respective platforms. Steps will be taken to ensure members have practiced the useage of them in solving likely tasks.

Organization
The group is divided into two separate smaller groups where each smaller group have one project as their responsibility. The aim is still for everyone to be involved in both projects however, and as such the role of leader for the entire project will be alternated throughout the weeks.

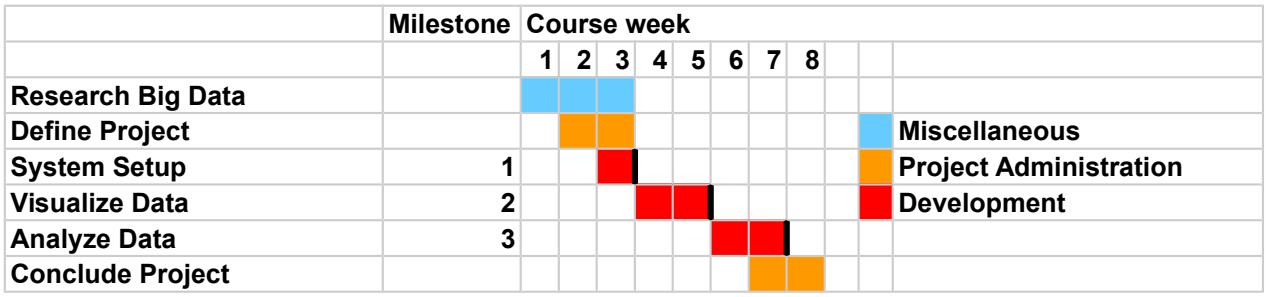
Healthcare	Virtual Assistance System
Jimmie Berger	William Björklund
Alexander Erenstedt	Victor Dahlberg
Philip Laine	Sebastian Lind

The project's students divided into their corresponding group.

Course week	1	2	3	4	5	6	7	8
Project Leader	Alex	Alex	Victor	Sebastian	Jimmie	William	Philip	Victor

Each weeks corresponding project leader.

The two projects will make use of different big data platforms, and the group members will acquaint themselves appropriately with these through reading and practice.
There is a variable hardware requirement in regards to the project. Preferably each of the two projects will have access to a cluster of computers to distribute its storage and computing on, but this is no strict requirement. Inescapably however, the two groups do require access to a computer with large amounts of primary memory in order to utilize the big data platforms.



The timeframe for the project is eight weeks, with the tasks to be accomplished being:

1. System Setup: Ingesting the respective dataset into the appropriate big data platform.
2. Visualise the data
3. Perform analytics to generate new data