The rise of data sets of massive scale, and the need to have quick and efficient analysis results, have led to a need for new data mining architectures, specifically cluster-environments. We plan on building a system that will allow users to run queries and manage data mining jobs on the UCT ICTS cluster. This will be done through a web-based user interface. Initially the system will only feature algorithms specifically designed for processing social media data, but ideally it will scale and allow people to easily build additional functionality into it in following years. Additional functionality includes the ability to pre-process different kinds of data, as well as implementing different algorithms. We'd like to get your input specifically with regards to what the user interface's features should be, as well as the back-end requirements. Please insert your answers below each question, as well as your name in the space below:

Name:

Context

1. What field of research are you in? Please give a brief description of the research you are doing now.

2. Have you worked with a similar system before? Please indicate the name of the system if you have.

Front-end requirements

1. What kind of info would you expect to see on the home screen of such an interface? For example, a log of previously executed jobs, a list of available data sets, etc.

2. If you have worked with a similar system, what was the usual process when 1. Running a query, and 2. Kicking off a data mining job? Please describe in detail.

3. What would be your ideal method for doing the above mentioned tasks? Please describe in detail, ex. Click Tools, select Job Type, input parameters, select data set, click Start.

4. How are your data mining results generally displayed after executing a job? How would you ideally like these results to be displayed? For example, list format, tables, graphs (please specify which type), visualized linked nodes.

5. What kind of formats would you like to be able to export your results to? For example, Excel file, PDF format, XML.

Back-end requirements

1. What information are you usually trying to mine from the data? For example: deriving a result through calculations, classification, clustering, etc.

2. What is the magnitude of the data sizes you expect to work with in your field?

3. Do you perform any preprocessing on the data before running the algorithms on it? If so, please give a short overview. Please specify how long it usually takes for different data set sizes.

4. Do the systems you normally use make use of some sort of pipelining, where intermediate outputs are generated and given to the next step to process and output a result?

5. What configurations do you usually process your data on? e.g. a cluster of 20 computers with 4 core CPU's running at 4.0GHz each.

6. Do you often change this configuration depending on the situation? If so, under which circumstances are you willing to change the configuration to cope with different situations?

7. How long does it take for a typical query to compute a result? Please indicate the type and magnitude of the query, and a rough indication of the time taken.