**Progress Report and Updated Milestones**

|  |
| --- |
| **Milestones** |
| 1. Back-end Software Stack    1. Research and choose software stack **DONE**    2. Build and Configure Hadoop, Hive, Shark, Spark, and BlinkDB **DONE**    3. Load data and test/benchmark BlinkDB **DONE** |
| 1. Hierarchical Clustering    1. Determine method and metric for hierarchical clustering **DONE**    2. Prototype clustering in Map Reduce **DONE**    3. Dealing with multi-variate data (different ranges, counts, categorical data) |
| 1. Visualization    1. Choose appropriate visualization techniques **DONE**    2. Implement visualization interface **DONE**    3. Scale the visualization to multi-dimensional data **DONE**    4. Zoom in/out procedure **PARTIAL** |
| 1. Integration    1. Integrate clustering into BlinkDB **PARTIAL**    2. Build interface from BlinkDB to front end visualization **PARTIAL**    3. Final refinements to UI after end to end inegration    4. Scale from 12GB data set to multi TB data set |

Additional Notes:

1. Plan to scale project up using EC2. The research papers we are referencing used this method for demos.  
2. We have zoom in/out capability right now but have decided to make some changes to it.

3. Integration between BlinkDB and front end is partially built using PHP Thrift calls to BlinkDB server.