# Progress Report and Updated Milestones

| Milestones |
| --- |
| 1. Back-end Software Stack<br>1.1. Research and choose software stack **DONE**<br>1.2. Build and Configure Hadoop, Hive, Shark, Spark, and BlinkDB **DONE**<br>1.3. Load data and test/benchmark BlinkDB **DONE** |
| 2. Hierarchical Clustering<br>2.1. Determine method and metric for hierarchical clustering **DONE**<br>2.2. Prototype clustering in Map Reduce **DONE**<br>2.3. Dealing with multi-variate data (different ranges, counts, categorical data) |
| 3. Visualization<br>3.1. Choose appropriate visualization techniques **DONE**<br>3.2. Implement visualization interface **DONE**<br>3.3. Scale the visualization to multi-dimensional data **DONE**<br>3.4. Zoom in/out procedure **PARTIAL** |
| 4. Integration<br>4.1. Integrate clustering into BlinkDB **PARTIAL**<br>4.2. Build interface from BlinkDB to front end visualization **PARTIAL**<br>4.3. Final refinements to UI after end to end inegration<br>4.4. 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.