**Under Armour Engineering Project**

**Data Extraction, Processing, and Analysis**

**Ian Robertson**

**What I need:**

1. Description of Problem (to get it in written form)
2. Keywords and phrases that are related to the information/videos you would like to extract from YouTube. *Note: it is possible to enhance the algorithms to identify synonyms and relevant topics centered around key words and phrases such that the system will become “smart”, but as we discussed that feature is not part of the initial prototype.*
3. Network Security Team to open port 22 for remote connection via putty if they haven’t already. Additionally, I may need them to relax network restrictions on port 5601, 9200-9400 (or just 9200 and (private) 9300 if ranges are not allowed). Possibly other ports as well, I will use SSL encryption to connect to applications via remote connection.
4. I am requesting a minimum of 2-3 servers with 64GB RAM and 1-2 TB HD apiece; however, I can work with less but it would mean the systems will not perform as well.
5. Preferably, I would like to be able to develop from a Linux distribution such as Red Hat/CentOS as some of the work will involve running jobs from the OS e.g. “cron tabs”. If the Operating system(s) are new or ‘fresh’ it will be easier to come in and begin development; however, if the IT team over there has a CM tool installed (Puppet/Chef) there will be some initial pre-configuration required.
6. The Servers and OS should be in a sandbox environment with a means to access the world-wide web (for youtube access)

**Services Provided:**

Engineering support for the development, troubleshooting, documentation, demonstration, and testing of a YouTube video data aggregation, analysis, and dissemination capability hosted on the Elasticsearch open-source platform that will present a list of relevant videos based upon a scoring metric.

**Timeline:**

80 hours to develop initial prototype with data extraction and analysis for proof-of-concept, an additional 80 hours for prototype enhancements including improvements to the scoring algorithm, search functionality, data analysis, documentation, among other improvements/enhancements.

**Milestones:**

1. Weekly (40 hour equivalence) Progress Reports.
2. Development of Data Extraction, storage, and processing schema.
3. YouTube Data extraction into Elasticsearch data storage.
4. Elasticstorage search functionality with relevant search results.
5. Adapt search functionality into 24-reporting events.
6. Feedback solicitation and incorporation (post any and all meetings).
7. Demonstration.

**Deliverables:**

1. Weekly (40 hour equivalence)\* Briefings and/or Email Reports
2. Bi-weekly (80 hours equivalence)\* Demonstrations.
3. Software code written in support of project
4. Operating System configuration
5. Basic Use Guide/outline
6. YouTube video data extraction into Elasticsearch distributed big data platform hosted on virtual or physical environment based upon needs and data schema.
7. YouTube video data analysis search capability based upon user needs and stated requirements.
8. Data results reporting based on continual time interval (i.e. hourly, daily, etc…)

Note: the terms “weekly” and “bi-weekly” refer to the full time equivalence; since this project will be part time those deliverables will be due once the 40 and 80 hour milestones are reached.