**Literature review extras**

Taylor, R. (2010). An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics, BMC Bioinformatics, [online], **11**(Suppl 12), pp.S1 [Accessed 15 March 2013] Available at: <http://www.biomedcentral.com/content/pdf/1471-2105-11-S12-S1.pdf>

This source details an overview of several technologies such as Hadoop and other technologies that are reliant on Hadoop such as MapReduce, HBase, Hive, Pig etc in the concept of ultra large scale data set analysis. This source mentions how this analysis is a developing issue and that open source software has been developed to be able to scale to petabyte level in order to deal with this issue. The source also gives an overview of the current uses of hadoop and its reliant technologies, as well as a conclusion on the usefulness of Hadoop and its reliant technologies, stating them to have a substantial base and is cost effective based analysis on commodity clusters. This source is similar to the work of Mone, G. (2013) who also details on Hadoop and technologies that add on to Hadoop, however this source focuses more on data storage add-ons while the previously referenced source deals more with distributed, cloud based add-ons. This source will help answer the academic question and achieve the aims by using this source when creating the artefact and in the final report during the evaluation of the artefact.

Conclusions:

Hadoop and the MapReduce programming paradigm already have a substantial base in the

bioinformatics community, especially in the field of next-generation sequencing analysis, and such use is increasing. This is due to the cost-effectiveness of Hadoop-based analysis on commodity Linux clusters, and in the cloud via data upload to cloud vendors who have implemented Hadoop/HBase; and due to the effectiveness and ease-of-use of the MapReduce method in parallelization of many data analysis algorithms.