**Topic: Big data case study in life sciences**

**Using Cassandra to retrieve evolutionary sequence conservation scores for Cancer mutation annotation**

**Problem:** Sequence conservation scores for each base pair of the human genome amounts to a large amount of data in the order of several gigabytes (~20) and regular RDBMS are not able to handle this amount of data in the speed that is needed to annotate hundreds or thousands of mutation data being generated by the increasing number of tumour samples being sequenced. The common method for analyzing this kind of data is keeping them in text files and using scripts for retrieving data. But when annotating large amounts of mutations from large amounts of samples, the time taken file read writes in the data retrieval process adds huge delays to data analysis pipelines.

**My goal:** Fast and accurate retrieval of sequence conservation scores to speed up pipelines for annotating mutations obtained from whole genome sequencing.

**Overview of steps:**

1. Installing and configuring Cassandra on Ubuntu
2. Using Cassandra Bulk Loader for inserting large datasets (> ~10,000,000 lines)
3. Implementing annotation solution in Java

**Big Data Set:** Evolutionary sequence conservation scores (PhastCons) for each base pair (only chromosome 1 was used for demo)

<http://hgdownload.cse.ucsc.edu/goldenPath/hg19/phastCons46way/placentalMammals/chr1.phastCons46way.placental.wigFix.gz>

Size : 1.3GB in wigFix format (4.5GB in csv format)

Format : .wigFix (This needs to be converted to a .csv format)

**Hardware:** Ubuntu 64 bit

**Software:** Cassandra 2.5.1, Eclipse and Datastax DevCenter

**Benefits (Pros/Cons):**

Pros: Compared to RDBMS, NoSQL DBs are able to handle storing, managing and retrieval of a large data source by spreading it over a cluster. Since data is in column format and mainly consists of read operations Cassandra is better suited that other NoSQL databases.

**Summary:** Spreading out per base pair conservation scores across a cluster in a per-chromosome (one node contains conservation scores of one chromosome, ideal # of nodes n > 23) manner allows for parallel retrieval of conservation scores for fast annotations thousands of single nucleotide variations and indels in cancer samples.