Online Analytical Processing on Big Sequence Data (PI: Dr. Lo Chi Lik Eric; 2012/13)

Many kinds of real-life data exhibit logical ordering among their data items and are thus sequential in nature. Examples of sequence data include web query logs, stock data, archived data streams and various kinds of RFID logs such as those generated by a commodity tracking system in a supply chain, and smart-card-based electronic payment systems like the Octopus system in Hong Kong. Similar to conventional data, there is a strong demand to warehouse and to analyze the vast amount of sequence data in a user-friendly and efficient way.

This project aims to research and develop an OnLine Analytical Processing (OLAP) engine for sequence data analysis. Building such an OLAP engine poses new research challenges, comparing with building a traditional OLAP engine for relational data. Specifically, while traditional OLAP engines support data grouping based on attribute values, an OLAP engine for sequence data should support *pattern-based aggregation*, i.e., the data are grouped by the various *patterns* they possess and an aggregate for each group is computed. Furthermore, an OLAP engine for sequence data should be prepared for handling very "big data", since nowadays it is not uncommon to see enterprises collecting and storing terabytes or petabytes of sequence data (e.g., logs) for analytical use. These factors make the design and implementation of an OLAP engine for sequence data analysis challenging. Interesting research questions include "*how should a pattern-based aggregate query be defined?*", "*what data structures should be pre-computed in order to support online evaluation of pattern-based aggregate queries?*", and "*what are the special issues if the sequence data are very big?*"

Although there is ample work on online analytical processing (OLAP) and recently on "big data analytics" (e.g., MapReduce), to our knowledge, there is little

work that discusses about building an OLAP engine for analyzing big sequence data. This project aims to research and develop such an engine. The project will consist of several phases: (i) formal definition of pattern-based aggregate queries and investigation of the usefulness of their results, (ii) design and implementation of proper data structures and algorithms for efficient evaluation of pattern-based aggregate queries, (iii) prototype an engine, which evaluates pattern-based aggregate queries and presents their results to users at interactive speed, and (iv) experimental evaluation of the performance and scalability of the suggested methods and implementation. Successful completion of this project will bring significant advancements to the area of OLAP on sequence data.