STSM Scientific Report

STSM title: Trajectory data management using data-intensive and cloud computing techniques
Reference: ECOST-STSM-IC0903-190513-027686
Grantee: Dragan Stojanovic, Faculty of Electronic Engineering
University of Nis, Serbia
Host: Apostolos Papadopoulos, Department of Informatics, Aristotle
University of Thessaloniki, Greece
Duration: June 09, 2013 – June 16, 2013

1. Purpose of the STSM

- Exchange previous works in the area of trajectory data management, query processing and analysis on high-performance distributed infrastructure that have been done by Faculty of Electronic Engineering, University of Nis and DElab, Aristotle University of Thessaloniki.
- Integrate trajectory and mobility data management, query processing and analysis over cluster/cloud computing infrastructure using advanced data-intensive computing frameworks and tools and the application of MapReduce (Hadoop) framework on trajectory data management and processing of spatio-temporal queries over massive trajectory data sets.

2. Description of the work carried out during the STSM

- Studied the adaptability of spatio-temporal algorithms to the MapReduce framework and its Hadoop implementation.
- We consider ideas to implement the spatial-join of a set of trajectories and a set of spatial objects, as well as the similarity search within the set of trajectories using MapReduce/Hadoop framework.
- The main challenge is how to deal with the skewed distribution of trajectory data, and perform the partition of the data set to support load-balancing that is needed to obtain performance gains over distributed computing infrastructure.
- We consider partition the trajectory segments in partitions in such a way that locality is preserved and partitions are balanced.
- We investigate indexing methods and algorithms for spatial-join of a trajectory data set partitioned and distributed in Hadoop Distributed File System (HDFS), and a set of spatial objects, e.g. point of interests (POI) or street/road segments for the purpose of symbolic annotation of trajectories.
- We also investigate indexing structures and algorithms for similarity search of trajectory data set for the purpose of trajectory clustering.
- When the size of the second dataset is small comparable to the size of trajectory data set (which is the common case) we consider using distributed cache mechanism provided by MapReduce/Hadoop to make this data available to all nodes in the cluster.

3. Description of the main results obtained

The main results of the visit is the method and algorithms developed for processing of spatial join queries between massive trajectory data set and the set of spatial objects (point of interests, street/roads segments, etc.). Also, we consider the method for similarity search of trajectory data set to support trajectory clustering and classification.

The method consists of three consecutive MapReduce phases:

The first phase has a massive trajectory data set as an input. The input data set is represented as a stream of positions, actually the trajectory segments reported by moving objects at specific timestamps and is ordered by timestamps in ascending order. Thus, input data is not actually structured and stored as continuous trajectories. We consider two variants for the input data for MapReduce processing: either the whole trajectory data set, or the sampled set performed by sampling over trajectory segments. The trajectory segments are split and distributed using a tiling scheme based on a regular grid to generate z-order values. At the end of the first Reduce phase, the histogram file containing the number of trajectory segments for each z-order value is generated. By performing cumulative counts (out of MapReduce) of numbers of trajectory segments for different z-order values, the split points of z-order values are detected and thus partitions are determined. The partitions determined include (almost) equal number of trajectories to provide load balancing for skewed rajectory data sets.

In the second MapReduce phase, each trajectory segment for all trajectories in the set is partitioned according to its z-value and partition splitting detected in the previous phase. Each part of the original trajectory data set is stored in a separate file forming the partition of almost the exact size. These files represent the input for the third MapReduce phase, which actually performs spatial join processing using a plane-sweep algorithm. Since the second dataset is small comparing to the size of trajectory data set, it can be made available through the distributed cache mechanism provided by MapReduce/Hadoop to all nodes in the cluster in order to perform spatial join processing.

A demo Hadoop application is developed related to spatial join of trajectory data set and street/road segments data set for the purpose of map matching and are currently tested for massive trajectory data and road/street network data on the DElab cluster computing infrastructure, as well as on IaaS cloud infrastructure at the Faculty of Electronic Engineering, both with Hadoop setup.

4. Future collaboration with host institution (if applicable)

- We are currently implementing and evaluating the algorithms described previously related to partition of the data set to support load-balancing between cluster nodes while preserving locality of trajectory data in each partition.
- We plan to complete an experimental evaluation of implemented method and algorithms on real massive trajectory data set and publish this work to a conference/journal in the next several months.
- The future research and collaboration will include adaptation of proposed method and developed algorithms to other trajectory related problems, such as similarity search to support trajectory clustering, analysis and classification.
- Preparation of the proposal for the new COST action related to Big spatio-temporal data management, query processing and analysis.

5. Foreseen publications/articles resulting or to result from the STSM (if applicable)

- A paper to be prepared for a journal or conference in the domain of Big spatio-temporal data management, processing and querying in 2013/2014
- The position paper for the COST MOVE final event in 30 Sept/1 Oct 2013.

6. Confirmation by the host institution of the successful execution of the STSM

Dragan Stojanovic's mission in our laboratory has been very successful. The collaboration was excellent and as can be seen from the first part of "STSM Scientific Report", the results achieved are very encouraging. Dragan have worked closely with several members of our group in the domain of trajectory data management, query processing and analysis on high-performance distributed computing infrastructure. Both our laboratories share common interest in this field and have previous research results. During his stay Dragan successfully combined these results creating added value to research performed by both laboratories. During his STSM Dragan proactively took part in research activities of our lab. Long term collaboration has been agreed on in the form of joint publications and development of demo applications on real massive trajectory data sets which is a certain proof of success of Dragan's STSM.

Dr. Dragan Stojanovic, Associate professor

Computer Science Department
Faculty of Electronic Engineering, University of Nis
Aleksandra Medvedeva 14, 18000 Nis
Serbia
Phone: + 381 18 529 235
Fax: + 381 18 588 399
E-mail: dragan.stojanovic@elfak.ni.ac.rs