**Parallel and Continuous Join Processing for Data Stream**

Présidente du Jury

M. Abdelkader Hameurlain, Professeur, IRIT, Rapporteur

M. Johan Montagnat, Chargé de Recherche, CNRS, Rapporteur

M. Lei Yu, Professeur, École Centrale Pékin, Examinateur

Mme. Bich-Liên Doan, Professeur adjoint, CentraleSupelec, Examinateur

M. Frédéric Magoulès, Professeur, CentraleSupelec, Directeur de thèse

M. Fabrice Huet, Professeur assistant, I3S, Co-Encadrant de thèse

Hello everyone, my name is Song Ge, I am a Ph.D. candidate from MICS library of CentraleSupelec University of Paris-Saclay. Thank you all for coming to my Ph.D. oral defense. I hope that you will enjoy my presentation.

My dissertation title is Parallel and Continuous Join Processing for Data Stream. And this is a large topic, so I divide it into two parts. The first part is Data Driven Stream Join which we take the algorithm of finding the nearest neighbors as our use case. And the second part is Query Driven Stream Join, for this one we take Semantic Web as an example. And I will have a short introduction about the background of big data and data stream at the beginning. And a conclusion and future research directions at the end.

We are all familiar with the words “Big Data”, because this is one of the hottest topics in the IT industry for the past 5 years. But how big is the data? Let’s have a short review. Google processes 24 PB of data every day. Facebook has 10 million photos and 3 billion “likes” each data. YouTube has 800 million visitors every month. And Twitter is doubling its size every year. So the most significant issue for processing these data comes from the size of data. And the flip side of size is speed. Normally, these data can not be processed in a single machine, so it often requires a parallel and distributed solution. But the network communication in the parallel and distributed processing system may cause a bottleneck of the system. The last issue is the dynamic of data. Data is changing all the time, and usually the newest ones have the most useful values. So the best way of describing data is data streams.

Different from a persistent static relation which requires a batch oriented data processing, the transient dynamic data streams require a real-time stream processing. In order to provide a stream processing, we need to be able to add or remove computational nodes based on the current load in the architecture level; and I also want to be able to withdraw old results and take new coming data into account in the application level.

The purpose of this dissertation is to study about the process for join operation in a parallel and continuous way for data streams. Join is a very popular and often used operation in the big data area. And when the data becomes large, we usually require a parallel join. According to the two different types of parallelism, we divide the joins studied in this dissertation into two different types, the first one is Data Driven join based on Data Parallelism and the second one is Query Driven join based on Task Parallelism.

For the data driven stream join we will take the kNN algorithm as our use case. This is the outline of this part. I would like to start with a short introduction about kNN. Given a set of query points R and a set of reference points S, a k nearest neighbor join is an operation which, for each point in R, discovers the k nearest neighbors in S. This algorithm is often used for classifications and classifier, it has many applications in social network such as finding similar users, recommendation systems like finding the nearest restaurants, or search engines etc.

The reason we call it a Data driven join is because that the query never changes. But the format of data does, especially the dimension of data. For example, in our experiences, we have used GPS data which is in 2 dimensions, twitter data which is in 77 dimensions, and also images data which is in 128 dimensions, etc.

The basic idea to solve this problem is to adopt a nested loop. The idea is straightforward, it is for every element in R calculate the distance in S, the complexity for this step is O n squired. Then sort the distances for each element in R to find the top k smallest ones. The step of sort often has a log complexity, for example log(n) for priority queues and nlog(n) for quick sort or merge sort.

This algorithm is both data intensive and computation intensive. So when the data grows bigger, one machine can not execute it and return the results in a reasonable time. That is why a parallel workflow is required.

We divide the parallel workflow into 3 steps. The first step is data preprocessing, this step is used to reduce the dimension of data, or to select the central points of data clusters.

The strategies of reducing the dimension of data rely on the projection of the high dimension data into a low dimension space. We introduced two methods in this dissertation. The first one is called Z-Value which is a kind of space filling curve. The z-value is a value in one dimension. It is calculated by interleaving the binary representation of the coordinates to form a new number. The Z in this figure indicates the neighborhood after the projection.

Another projection strategy is Locality Sensitive Hashing. This method uses several hash functions to map the closer data into the same bucket by using the hash collisions.

The other thing which can be done in the preprocessing step is to select the central points of data clusters. We call them the pivots. This selection will be further used for data partitions. We introduced 3 selection methods. The random selection generates a set of samples, then calculates the pairwise distance of the points in the sample. The sample with the biggest sum of distances is chosen as the set of pivots. The furthest selection method randomly chooses the first pivot, and calculates the furthest point to this chosen pivot as the second pivot. And repeat this step until having the desired number of pivots. The k-mean selection applies the k-means method on a data sample to update the centroid of each cluster as the new pivots until the set of pivots stabilizes.

The second step is data partitioning. In this step, we have distance based partitioning strategy and size based partitioning strategy. Data partition is a very important step, because it affects the performance especially for the amount of data transmitted through network. If we don’t use any advanced partitioning strategy, and just partition data randomly, we will have a wastes of resources. For example, like in this figure, we randomly partition R and S, and distributed R on rows, and S on columns, then in order to make each subset of R calculate with every subset of S, we need to replicate both partitions of R and S n times, which leads to a total number of n square tasks for calculating the pairwise distances. This replication leads to low efficiency of performance.

So the key to improve the performance is to preserve spatial locality of objects when decomposing data for tasks. More precisely, what we want is: for every partition Ri, find the only corresponding partition Sj, where the kNN join of Ri and S equals to the kNN join of Ri and Sj. And at the same time, we also need to make sure that the kNN join of R and S, is the union of the kNN join of all the partitions of R and S. So after using the advanced partitioning strategies, we can reduce the number of partitions from n square to n.

The first type of advanced partitioning strategy we want to introduce is the Distance Based Partitioning Strategy. This strategy wants to have the most relevant points in each partition. This method is based on Voronoi using the pivots we have chosen in the pre-processing step. Basically, we will