Individual Contribution Report - CSE 511

Ajay Kannan

We have been hired by a major peer-to-peer taxi cab. This is a project where is we tasked with building a geo-spatial database and running multiple spatial queries on their large database that contains geographic data as well as real-time location data of their customers. Using Spark-SQL on a geo-database, the goal of the project is to extract data from this database that will be used by your client for operational (day-to-day) and strategic level (long term) decisions.

Reflection

The spark project was tough to understand and implement. We had to go through the process of learning Scala and spark. We helped each other to do the same. I contributed to the overall work flow of the project. We developed two project codebases for both phases, out of which the best or more effective one was selected. Also, I was part of the project development and discussion.

In **phase 1**, I had to understand the four queries, namely range query, range join query, distance query and distance join query. Both range queries use ST_Contains which gives the rectangle and the point which need to be checked. It refreshed my geometry skills and I wrote the pseudo-code in the planning stage itself. To be specific, in the phase 1, I developed the code and tested the same. I also changed the test cases of the first phase and check the output for same. It turned out to be as expected. Also, Boundary cases for the code is an important factor which I realised in the alpha stage.

Just like the phase, I began to understand the work-flow step-by-step in the **phase 2**. My peers and I, began to divulge in problem of hot-spot analysis. The evaluation of the analysis is quite complex. Evaluation can be done using two metrics:

A. correctness of the result (do the identified spatiotemporal hot spots match those returned by the reference implementations), and
B. computation time when running against the specified input dataset across multiple node.

We had study the codebase for implementing it. For the phase 2, though my code was not used in the final stage, I was a part of testing the codebase and validating it. These functions were rather challenging to understand and code.

Analysis

The first phase was simple and we had to implement only two functions. ST_Within and ST_Contains are the two user defined function that calculate the distance between two points and also to check whether the points are contained within a rectangle or not. This is done using four stand-alone queries which is already provided in the project phase one. Understanding, I did execute the implementation of the two functions.

For the phase two, I found it hard to understand it at first. But, through the discussion with my peers, I understood and I implemented it. Though, my code was not used the at end for phase two, I learnt a great deal from the discussions and code reviews. There are four modules for the phase, which can be classified as hot-cell analysis functions and hot-zone analysis function (function and its respective utility function). I helped the making of the hot-cell functions and the testing for the same.

Lessons Learned

We separated ourselves into to two groups and executed the two phases of the project. I learnt that the best way to execute the project is by discussing every new findings and having a review every now & then. That way, we won't miss any important details and nobody will be left behind.

Future Applications

First of all, I learnt how to work using Apache-spark and Scala. Just like javascript, Scala is is a strong statically typed general-purpose programming language which supports both object-oriented programming and functional programming. It is mainly used for databases and its utilities.

I also learnt the value of geo-Spatial databases and its use-cases such as hard terrain use-case or some simple cab application. And since we live in a world where observational data is being collected at an increasing rate. This data is often considerably valuable to the organisations, both from a real-time analytic perspective (e.g., geofencing. In order to handle these large collections of observational data (e.g., numbering in the billions), distributed processing techniques are required. Over the past few years, interest in the such framework has exploded, both in government, industry, and academia. So, I can use geo-spark for my future endeavours in geo-databases.

I refined my team building skill-sets also. Since this is my first semester, I believe that values I learnt in this course will be put to good use later in the future endeavours.

Assessment

- Yes, I made an honest effort to learn from this project experience. The best of my Knowledge, I have worked well and mingled with peers to fulfil the criteria of the project.
- I have implemented most of my necessary learnings of Data processing at scale in the Scala project. This project adds to my knowledge in NoSQL databases as I have already worked with MongoDB and CassandraDB.
- The course also helped me with the general understanding of the data processing and this project has solidified it. This has also been a key course to strengthen my learning and interest in the field of computer science.