QoS-Aware Access Control for Big Geospatial Data Management

Abstract –

In the last decade or so, the proliferation of ubiquitous GPS-enabled devices and Internet of Things (IoT) have resulted in an accumulation of huge number of datasets. Most of this data is geo-referenced and real-time. Managing such datasets is challenging, which motivated several works to provide QoS-aware optimizations for current big data management systems (for example, Apache Spark and MongoDB). However, works from relevant literature have mainly focused on supporting data processing capabilities for big data management systems, accounting for the three well-known *Vs* (velocity, volume and variety). However, interesting real-life applications allow many authorization-related access-patterns by many users. Consider a scenario where many users issue similar requests to a disk-resident big data set. Even though final results may differ, it is often the case that they share a common subset. Hence, retrieving the same subset several times for different user requests (depending on their access level) challenges system's capacity (for example, main memory and CPU) and may cause the system to come into a halt. To aggravate the challenge, consider that disk-resident data is georeferenced (tagged with spatial reference – GPS data, multidimensional data represented by longitude and latitude) and query requests are proximity-alike that seek answers revealing pairwise relationship among spatial dataset objects. For example, a range-based query requires finding a set of spatial objects that surround a focal point. For instance, finding all people who were around a person in a specific time in a city center. Answering this kind of queries is compute-intensive because of the geometric calculations involved (consider that locations are represented using a multidimensional way – longitude and latitude). The challenge is doubled if we consider that the same query (or subset of it) need to be computed many times for several people as per their access rules, which means scanning the disk many times searching for the same subset of result set. If we also consider that those results need to be joined with a data set coming from a streaming source, the problem may turn intractable. The aim of this proposal is designing an efficient QoS-aware Rule Based Access Control (RBAC) framework for big geospatial data processing. This may include custom spatial indexing, a novel caching scheme, a statistical front-stage preprocessing (for the stream data) and a query profiler, to mention just a few.