ABSRTRACT

The recent evolution of ubiquitous positioning technologies has led to the rapid growth of the usage of spatial data by various systems, such as geographic information systems (GIS) and location based services (LBS). The spatial queries of these systems are complex and computationally intensive, this caused the need of handling them using high performance platforms.

MapReduce and its open source implementation Hadoop, have been extensively used by large enterprises in processing large-scale data sets in parallel and therefore, they are suitable for performing the intensive computations of spatial data. However, Hadoop is not aware with the spatial properties of spatial data, which is why computation benefits of considering the spatial properties of the data when processing, may be wasted by this limitation. This introduces the need of extending Hadoop to be aware of spatial features of the data in order to obtain better performance.

Several research work have proposed approaches to extend Hadoop to support spatial data. For example, SpatialHadoop extends Hadoop to add support to spatial data and spatial queries. This is achieved through the following: (1) using spatial indexes to efficiently access the spatial data stored on the Hadoop distributed file system (HDFS) and (2) adding an operation layer that supports spatial operation.

This thesis introduces Co-SpatialHadoop; it is a Hadoop-based spatial approach and an extension for SpatialHadoop platform. It enhances the network usage of spatial complex queries such as “spatial join” by colocating related spatial files together on HDFS, which may be joined later. It enhances the saving operation of HDFS by introducing a new spatial placement policy which colocates data using spatial attributes of spatial records.

Co-SpatialHadoop also builds inverted indexes using non-spatial attributes of data records to enhance the response time of non-spatial queries. It adds non-spatial operations to operation layer of SpatialHadoop. These modules are implemented using the functionality of MapReduce.

Co-SpatialHadoop uses real spatial data obtained from OpenStreetMap (OSM) in experiments. The experiments demonstrate good results; colocation algorithm enhances network overhead of spatial join queries by 60%. Also, inverted indexes enhance response time and network overhead of non-spatial queries.

List of Figures

List of Tables

List of Algorithms

LIST OF ACRONYMS

GFS	Google File System
GIS	Geographic Information System
GPS	Global Positioning System
HDFS	Hadoop Distributed File System
KNN	K Nearest Neighbors
LBS	Location Based Services
MBR	Minimum Bounding Rectangle
ORDBMS	Object Relational Database Management Systems
SDBMS	Spatial Database Management Systems
WKB	Well Known Binary Language
WKT	Well Known Text Language