CHAPTER 6

6. CONCLUSION AND FUTURE WORK

6.1.CONCLUSION

In this thesis, we research colocating spatial data files that are stored on a distributed file system. Queries accessing these files potentially access two or more at a time and are executed by Hadoop. Our study shows that colocating blocks of these files based on their spatial properties that are accessed together by queries can have a significant improvement on the network overhead during query execution. In this thesis, we propose an algorithm for colocating spatial files. Our proposed approaches and algorithms are encapsulated in Co-SpatialHadoop, which we implemented as extension to SpatialHadoop [1]. Additionally, we extended HDFS to add a new block placement policy that is based on spatial properties of the data. Moreover, we add an inverted index of non-spatial attributes to improve the performance of the non-spatial queries.

To evaluate the effectiveness of our proposed colocation approach implemented in Co-SpatialHadoop, we compare the execution performance of queries when we use Co-SpatialHadoop and SpatialHadoop [1]. The experiments showed that the introduced colocation algorithm that is based on spatial file properties enhances the network overhead of spatial query. Moreover, using inverted indexes on non-spatial attributes significantly enhances execution time of non-spatial queries.

Next, we discuss possible future work extensions to this thesis work.

6.2.FUTURE WORK

1. Multidimensional dynamic data

   Temporal-Spatial queries such as weather or traffic predictions are executed on files with more complex structure. Therefore, more challenges are introduced. The colocation algorithms we proposing are designed for static data only, and need to be modified to address these new challenges.

2. Placement of spatial blocks

   In Section 3.5, we have identified a challenge about file blocks that can be assigned to multiple locator items in the locator table, which we use for tracking block placement and enforcing colocation of files. We partially addressed this challenge in our proposed work and more techniques can be investigated. Locator table partitioning can be handled more efficiently to decrease the overlapping between locator items, which is expected to solve colocation overlapping problem.

3. Replicas with different partitioning techniques

   Hadoop Distributed File System (HDFS) replicates files to achieve fault tolerance. This feature can be exploited similar to HAIL [31] to store multiple partitions of the data. For example, we can partition the spatial files using different spatial index (R-Tree, R*Tree, Grid ...) and store them as replicas of the file. Spatial operations can choose the index that produces the minimal number of input splits and map functions to enhance the performance.

4. Enhance the inverted index
    - Enhance inverted index to handle multi-words queries.
    - Enhance inverted index to handle dynamic data.
    - Detect the most used attributes from old queries workload.

5. Implement integrated queries (spatial/non-spatial)

    We propose to extend our work by adding support to integrated queries, which are queries that include spatial and non-spatial operations. We can add "where" clause to spatial join query to join records based on spatial attribute after filtering them based on non-spatial attribute used by "where" clause. The non-spatial filtering is handled by inverted index, which filters blocks prepared by spatial index.

7. BIBLIOGRAPHY

[1] Ahmed Eldawy and Mohamed F. Mokbel, "A Demonstration of SpatialHadoop: An Efficient MapReduce Framework for Spatial Data". In *Proceedings of the VLDB Endowment*, volume 6, no.12, pages 1230-1233, Aug. 2013. Available at: http://spatialhadoop.cs.umn.edu/

[2] Mohamed Y. Eltabakh, Yuanyuan Tian, Fatma O¨ zcan, Rainer Gemulla, Aljoscha Krettek, and John McPherson, "Co-Hadoop: Flexible Data Placement and Its Exploitation in Hadoop". In *Proceedings of the VLDB Endowment*, volume 4, no.9, pages 575-585, Jun. 2011.

[3] OpenStreetMap (OSM), http://wiki.openstreetmap.org/

[4] J. Patel, J. Yu, N. Kabra, K. Tufte, B. Nag, J. Burger, N. Hall, K. Ramasamy, R. Lueder, C. Ellmann, J. Kupsch, S. Guo, J. Larson, D. De Witt, and J. Naughton, "Building a scaleable geo-spatial dbms: technology, implementation, and evaluation". In *Proceedings of the ACM SIGMOD Conference*, 1997, pages 336–347.

[5] A. Pavlo, E. Paulson, A. Rasin, D. J. Abadi, D. J. DeWitt, S. Madden, and M. Stonebraker, "A comparison of approaches to large-scale data analysis". In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*, 2009, pages 165–178.

[6] F. Wang, J. Kong, L. Cooper, T. Pan, K. Tahsin, W. Chen, A. Sharma, C. Niedermayr, T. W. Oh, D. Brat, A. B. Farris, D. Foran, and J. Saltz, "A data model and database for high-resolution pathology analytical image informatics". *J Pathol Inform*, 2(1):32, 2011.

[7] F. Wang, J. Kong, J. Gao, D. Adler, L. Cooper, C. Vergara-Niedermayr, Z. Zhou, B. Katigbak, T. Kurc, D. Brat, and J. Saltz, "A high-performance spatial database based approach for pathology imaging algorithm evaluation". J *Pathol Inform*, 4(5), 2013.

[8] Spatial and Geographic objects for PostgreSQL (PostGIS), http://postgis.net/

[9] Euro Beinat, Albert Godfrind and Ravikanth V. Kothuri, *Pro Oracle Spatial for Oracle Database 11g.* Apress, 2007 ISBN 1-59059-899-7

[10] Dinesh Agarwal, Satish Puri, Xi He, and Sushil K. Prasad, "Cloud Computing for Fundamental Spatial Operations on Polygonal GIS Data". In *Cloud Futures journal*, 2012

[11] Jeffrey Dean and Sanjay Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters". In *Communications of the ACM - 50th anniversary issue: 1958 - 2008*, volume 51, pages 107-113, Jan. 2008.

[12] Apache Hadoop, http://hadoop.apache.org/

[13] Big Data, analytics and cloud blog: http://thebigdatablog.weebly.com/blog/the-hadoop-ecosystem-overview

[14] *MPLSVPN blog*, Physical and logical structure of Hadoop:

http://www.mplsvpn.info/2012/11/hadoop-architecture-types-of-hadoop.html

[15] Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung, "The Google File System". In *Proceedings of the 19th ACM symposium on Operating systems principles* SOSP, 2003, pages 29-43.

[16] Tom White, *Hadoop: The Definitive Guide*, 3rd ed. O'Reilly, 2012

[17] Hive QL: https://hive.apache.org/

[18] Pig Latin: http://pig.apache.org/

[19] Hbase: http://hbase.apache.org/

[20] Antonin Guttman, "R-trees: A Dynamic Index Structure for Spatial Searching". In *Proceedings of the ACM SIGMOD international conference on Management of data*, 1984, volume 14, no.2, pages 47-57.

[21] Timos Sellis, Nick Roussopoulos, and Christos Faloutsos, "The R+-Tree: A Dynamic Index for Multi-Dimensional Objects". In *Proceedings of the 13th International Conference on Very Large Data Bases* (VLDB), 1987, pages 507-518.

[22] Renfeng A. Xu and Dingju B. Zhu, "An Approach for Processing Underground Spatial-Temporal Data by Cloud Computing". In *Journal of Automation and Control Engineering*, volume 1, no.2, pages 164-165, Jun. 2013.

[23] Yonggang Wang and Sheng Wang. "Research and Implementation on Spatial Data Storage and Operation Based on Hadoop Platform". In *Proceedings of the 2nd IITA International Conference on Geoscience and Remote Sensing* (IITA-GRS), IEEE, 2010, volume 2, pages 275-278.

[24] Ariel Cary, Yaacov Yesha, Malek Adjouadi, and Naphtali Rishe. "Leveraging Cloud Computing in Geodatabase Management". In *Proceedings of the IEEE International Conference on Granular Computing* (GrC), 2010, pages 73-78.

[25] Shoji Nishimura, Sudipto Das, Divyakant Agrawal and Amr El Abbadi, "MD-HBase: A Scalable Multi-dimensional Data Infrastructure for Location Aware Services". In *Proceedings of the 12th IEEE International Conference On Mobile Data Management* (MDM), 2011, volume 1, pages 7-16.

[26] Ablimit Aji, Fusheng Wang, Hoang Vo, Rubao Lee, Qiaoling Liu, Xiaodong Zhang, and Joel Saltz, "Hadoop-GIS: A High Performance Spatial Data Warehousing System over MapReduce". In *Proceedings of the VLDB Endowment*, volume 6, no.11, pages 1009-1020, Aug. 2013.

[27] A. Cary, Z. Sun, V. Hristidis and N. Rishe, "Experiences on Processing Spatial Data with MapReduce". In *Proceedings of the 21st international Conference on Scientific and Statistical Database Management*, 2009, pages 302-319.

[28] D. Pelleg and A.W. Moore, "X-means: Extending K-means with Efficient Estimation of the Number of Clusters". In *Proceedings of the* 17th international Conference on Machine Learning, 2000, pages 727-734.

[29] William B. Frakes and Ricardo Baeza-Yates. *Information Retrieval and Algorithms.*

[30] Ganglia Monitoring System, http://ganglia.sourceforge.net/

[31] Jens Dittrich, JorgeArnulfo, Quian eRuiz, Stefan Richter, Stefan Schuh, Alekh Jindal, and J¨org Schad, "Only Aggressive Elephants are Fast Elephants" [HAIL]. In *Proceedings of the VLDB Endowment*, volume 5, no.11, pages 1591-1602, Jul. 2012.

[32] Private Communications with SpatialHadoop team, Ahmed Eldawy, Ph.D student in university of Minnesota.

[33] Private Communications with Hadoop-GIS [26] team, Emory University.