ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 5, May 2018

Review on Map and Reduced based Semantic Parsing of Open Street Map using KNN

Poonam Devi [1], Rachna [2]

M.Tech. Student, Department of Computer Science & Engineering, N.G.F College of Engineering and Technology at Palwal, Haryana, India[1]

Assistant Professor, Department of Computer Science & Engineering, N.G.F College of Engineering and Technology at Palwal, Haryana, India[2]

ABSTRACT : Support of privileged exploration on vast volumes of spatial information turns out to be progressively critical in numerous application spaces, incorporating geospatial issues in various fields, area based administrations, and rising logical applications that are progressively information and process escalated. The rise of monstrous scale spatial information is because of the expansion of savvy and pervasive situating advances, improvement of high determination imaging advances, and commitment from countless clients. There are two noteworthy difficulties for overseeing and questioning enormous spatial information to help spatial inquiries: the blast of spatial information, and the high computational intricacy of spatial inquiries. In this paper, we present Big Data Geospatial Information System – an adaptable and elite spatial information warehousing framework for running extensive scale spatial questions on Big Data. Big Data Geospatial Information System underpins numerous kinds of spatial questions on Map-per and Reducer through spatial parsing adjustable spatial inquiry and queries for mammoth extraction and parallel spatial question execution on Map-per and Reducer, and successful strategies for correcting question comes about through taking care of limit objects. Big Data Geospatial Information System uses worldwide tract categorize and adaptable on request neighbourhood spatial ordering to accomplish productive inquiry handling. Big Data Geospatial Information System is coordinated into Hive to help decisive spatial inquiries with an incorporated engineering. Our investigations have shown the high effectiveness of Big Data Geospatial Information System on inquiry reaction and high adaptability to keep running on product groups. Our similar examinations have demonstrated that execution of Big Data Geospatial Information System is keeping pace with parallel SDBMS and beats SDBMS for figure escalated questions. Big Data Geospatial Information System is accessible as an arrangement of library for handling spatial inquiries, and as an incorporated programming bundle in Hive over Big Data Ecosystem along-with machine learning amalgamated algorithm namely K-nearest neighbour (KNN) .

.

KEYWORDS: MapReduce (Map-per Reducer) Big Data, Hadoop, HIVE, Open Street Maps (OSM), Geographic Information Systems (GIS), Machine Readable Language, Semantic Parsing, K Nearest Neighbours (KNN), SDBMS (Spatial Database Management System).

I. INTRODUCTION

The rapid growth of spatial data is driven by not only industrial applications, but also emerging scientific applications that are increasingly data- and compute- intensive. With the rapid improvement of data acquisition technologies such as high-resolution issue slide scanners and remote sensing instruments, it has become more efficient to capture extremely large spatial data to support scientific research. For example, digital pathology imaging has become an emerging field in the past decade, where examination of high resolution images of tissue specimens enables novel, more effective ways of screening for disease, classifying disease states, understanding disease progression and evaluating the efficacy of therapeutic strategies. Pathology image analysis offers a means of rapidly carrying out quantitative, reproducible measurements of micro-anatomical features in high-resolution pathology images and large image datasets. Regions of micro-anatomic objects (millions per image) such as nuclei and cells are computed through

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 5, May 2018

image segmentation algorithms, represented with their boundaries, and image features are extracted from these objects. Exploring the results of such analysis involves complex queries such as spatial cross-matching, overlay of multiple sets of spatial objects, spatial proximity computations between objects, and queries for global spatial pattern discovery. These queries often involve billions of spatial objects and heavy geometric computations.

A noteworthy prerequisite for the information concentrated spatial applications is quick inquiry reaction which requires a versatile design that can question spatial information on an expansive scale. Another prerequisite is to help questions on a practical design, for example, item groups or cloud conditions. In the interim, logical scientists and application designers regularly favour expressive inquiry dialects or interfaces to express complex inquiries easily, without agonizing over how questions are interpreted, upgraded and executed. With the fast change of instrument resolutions, expanded exactness of information investigation strategies, and the gigantic size of watched information, complex spatial inquiries have progressed toward becoming progressively figure and information escalated because of following difficulties.

1. **Big Data Confrontation**: High determination microscopy pictures from high determination advanced slide scanners give rich data about spatial items and their related highlights. For instance, entire slide pictures made by examining magnifying instrument slides at demonstrative determination are substantial: A normal WSI contains 100,000x100,000 pixels. One picture may contain a large number of items, and several picture highlights can be removed for each question. An investigation may include hundreds or thousands of pictures acquired from a huge accomplice of subjects. For vast scale interrelated investigation, there might be many calculations - with shifting parameters - to create a wide range of result sets to be thought about and merged. In this manner, got information from pictures of a solitary report is frequently in the size of several terabytes. A direct size clinic can routinely produce a huge number of entire slide pictures every day, which can prompt a few terabytes of determined logical outcomes every day, and peta-bytes of information can be effortlessly made inside multi year. For the Open-Street Map venture, there have been in excess of 600,000 enlisted patrons, and client contributed information is expanding ceaselessly.
2. **Data Separation:** Spatial information dividing is a basic introductory advance to characterize, produce and speak to parceled information. There are two noteworthy contemplations for spatial information parceling. The primary thought is to stay away from high thickness apportioned tiles. This is predominantly because of potential high information skew in the spatial dataset, which could cause stack unevenness among laborers in a bunch domain. Another thought is to deal with limit meeting objects appropriately. As Map Reduce gives its own particular activity booking to adjusting errands, the heap irregularity issue can be in part lightened at the undertaking planning level. In this way, for spatial information parceling, we principally center around breaking high thickness tiles into littler ones, and adopt a recursive apportioning strategy. For limit converging objects, we adopt the numerous task based strategy in which objects are recreated and allotted to each crossing tile, trailed by a post-preparing venture for curing question comes about the effective results and concludes thus can be produced.
3. **Map-per Reducer Supported Parallel Query Implementation**: Rather than utilizing unequivocal spatial question parallelization as condensed in mapzen, we adopt an understood parallelization strategy by utilizing Map-per Reducer. This will much streamline the advancement and administration of inquiry employments on groups. As information is spatially apportioned, the tile name or UID frames the key for Map-per Reducer, and distinguishing spatial objects of tiles can be performed in mapping stage. Contingent upon the question unpredictability, spatial inquiries can be actualized as guide capacities, decrease capacities or mix of both. In view of the question composes, diverse inquiry pipelines are executed in Map-per Reducer. The same number of spatial inquiries include high intricacy geometric calculations, inquiry parallelization through Mapper Reducer can fundamentally decrease question reaction time.
4. **Spatial Data Partitioning and Storage**: Spatial information parceling fills two noteworthy needs. In the first place, it gives two-dimensional information apportioning and produces an arrangement of tiles, which turn into a preparing unit for questioning errands. A vast arrangement of such errands can be prepared in parallel without

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 5, May 2018

information dependence or correspondence necessity. Consequently, spatial dividing gives information apportioning as well as computational parallelization. Last, spatial information dividing could be basic to relieve spatial information skew. Information skew is a typical issue in spatial applications. For instance, with a settled network dividing of pictures into tiles with size of 4Kx4K, the biggest include of items a tile is more than 20K articles, contrasted with the normal check of 4,291. For Open-Street-Map dataset, by parceling the space into 1000x1000 settled matrices, the normal check of items per tile is 993, yet the biggest include of articles a tile is 794,429. In the event that there is a parallel spatial inquiry preparing in light of tiles, such extensive skewed tile could fundamentally build the reaction time because of the straggling tiles. As Map-per Reducer gives its own activity planning to adjusting errands, for spatial information parceling, we principally center on breaking high thickness locales into little ones, and adopt a recursive dividing strategy. We either accept the information is a pre-parceled tile set with settled matrix measure, which is usually utilized for imaging examination applications, or pre-create settled lattice based tile set if no dividing exists. We include the quantity of articles each tile, and sort them in view of the tallies. We characterize an edge C_{max} as the maximal include of articles permitted a tile. We pick all tiles with question checks bigger than C_{max}, and split every one of them into two equivalent half-sized tiles in view of an ideal heading: x or y. A heading is viewed as ideal if the split along that bearing produces another tile with question check beneath the limit, or the two new tiles are more adjusted. This procedure is rehashed until the point when all tiles have checks beneath than C_{max}

5. **Integration with HIVE:** Hive is an open source Map-per Reducer based question framework that gives a decisive inquiry dialect to clients. By giving a virtual table like perspective of information, SQL like question dialect Hive-Query Language, and programmed inquiry interpretation, Hive accomplishes versatility while it extraordinarily streamlines the exertion on creating applications in Map-per Reducer. Hive-Query Language bolsters a subset of standard ANSI SQL articulations which most information examiners and researchers know about. The below figure depicts the scenario under the scheme for ready reference, perusal and consideration to gain the purpose in the related context and productive solutions.

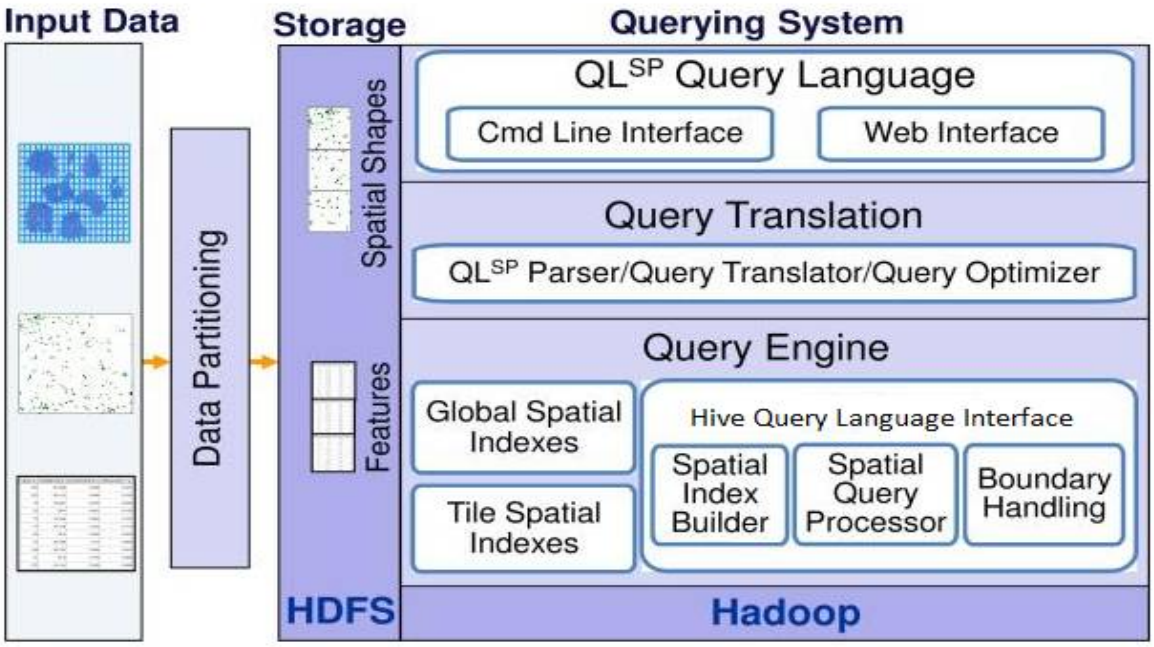

Figure 1: Data Partitioning using Big-Data (Hadoop Distributed File System and Query Translation and Query Execution using Eco System provided by Hadoop based on Map and Reduced Algorithms.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 5, May 2018

II. RELATED WORK

Spatial help has been reached out to NoSQL based arrangements, for example, neo4j/spatial [2] and GeoCouch [1]. These methodologies assemble spatial information structures and access techniques over key-esteem stores, in this manner exploit the adaptability. Nonetheless, these methodologies bolster constrained inquiries, for instance, GeoCouch underpins just jumping box questions, and there is no help of the investigative spatial inquiries for spatial information warehousing applications.

Parallel SDBMS has been utilized for overseeing and questioning huge scale spatial information in light of shared nothing design [28], for example, Greenplum, IBM Netezza, Teradata, and divided rendition of IBM DB2 Spatial Extender, Oracle Spatial, MS SQL Server Spatial, and PostGIS. These methodologies do not have the structure for spatial parceling and limit protest taking care of. Information stacking speed is a noteworthy bottleneck for SDBMS based arrangements [29], particularly for complex organized spatial information composes [35]. We have beforehand built up a parallel SDBMS based approach PAIS [34, 35, 9] in view of DB2 DPF with sensible versatility, however the approach is exceptionally costly on programming permit and equipment requirement[29], and requires complex tuning and upkeep. The goal of the work introduced in this paper is to give a versatile and savvy way to deal with help expressive and superior spatial inquiries. The Sloan Digital Sky Survey venture (SDSS) [3] makes a high determination multi-wavelength guide of the Northern Sky with 2.5 trillion pixels of imaging, and takes a vast scale parallel database approach. SDSS gives a high accuracy GIS framework for cosmology, executed as an arrangement of UDFs. The database keeps running on GrayWulf design [31] through joint effort with Microsoft.

Examinations of Map Reduce and parallel databases for organized information are talked about in [29, 20, 30]. Tight mix of DBMS and Map Reduce is examined in [10, 37]. Map Reduce frameworks with abnormal state revelatory dialects incorporate Pig Latin/Pig [27, 21], SCOPE [19], and HiveQL/Hive [32]. YSmart gives an improved SQL to Map Reduce work interpretation and is as of late fixed to Hive. Hadoop-GIS adopts a strategy that coordinates DBMS's spatial ordering and definitive question dialect capacities into Map Reduce

Segregation based approach for parallelizing spatial joins is talked about in [40], which utilizes the various task, single join approach with apportioning based spatial join calculation. The creators additionally give re-adjusting of undertakings to accomplish better parallelization. We adopt the same various task strategy for dividing, however utilize file based spatial join calculation, and depend on Map Reduce for stack adjusting. R-Tree based parallel spatial join is additionally proposed in early work [17] with a consolidated shared virtual memory and shared nothing engineering. As of late we have abused enormous information parallelism by creating GPU mindful parallel geometric calculation calculations to help spatial joins running on work area machines [36]. Incorporating GPU into our Map Reduce pipeline is among our future work.

III. PROPOSED ALGORITHM

The above proposed scheme is the amalgamation of Map-per Reducer and Machine Learning Model using KNN to extract the Corpus using Spatial parsing techniques with Hadoop (Big Data) Eco System the following are steps to extract the effective and accurate results wherein the execute time will be less and computation utilisation will be optimal for the proposed framework :

1. Download the open street map data (osm) in XML format.
2. Migrate the osm data to HDFS repository using distributed framework.
3. Spatial Data processing using Map-Reduce (MAP,SHUFFLE,REDUCE) to summarise the corpus.
4. Semantic parsing using KNN will form the relations between the nodes and ways (in relation or arbitrary).
5. Semantic Schema will be formed in HIVE.
6. OSM data will be integrated with HIVE Semantic Schema.
7. Contextual Queries will be executed for results.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 5, May 2018

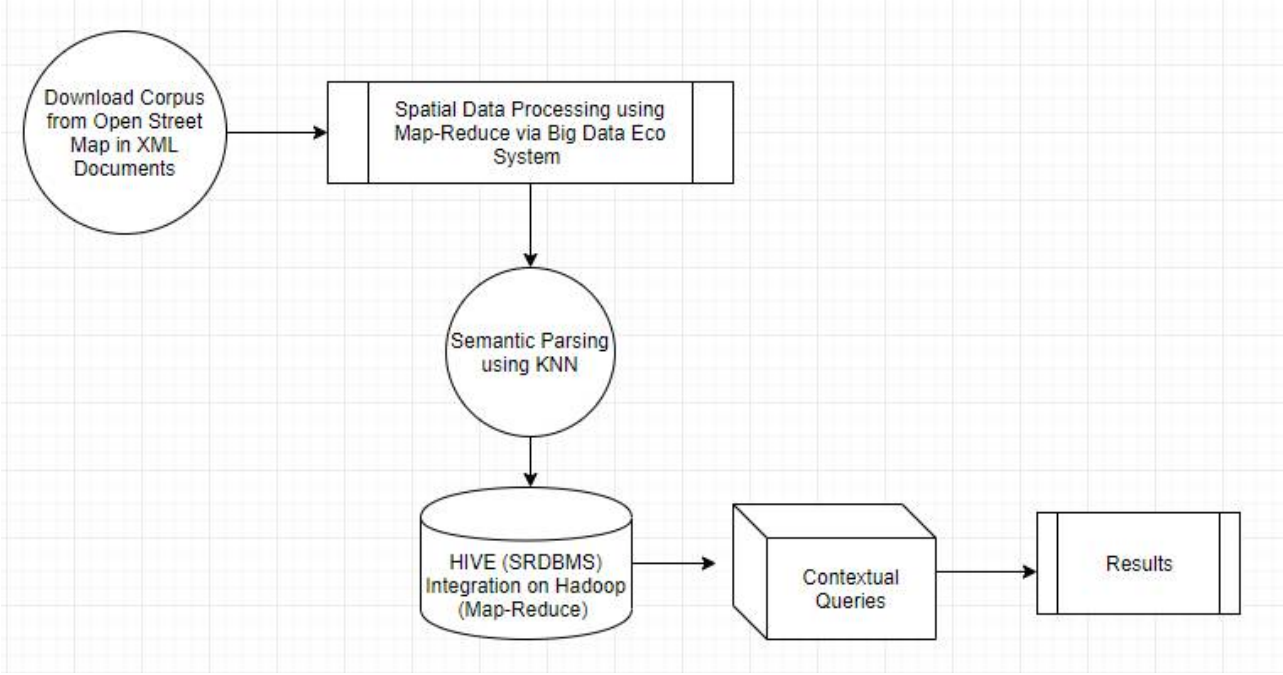

Figure 2: Proposed scheme the amalgamation of Map-per Reducer and Machine Learning Model using KNN to extract the Corpus using Spatial parsing techniques with Hadoop

REFERENCES

1. Geocouch. https://github.com/couchbase/geocouch/
2. neo4j/spatial. https://github.com/neo4j/spatial.
3. The sloan digital sky survey project (sdss) http://www.sdss.org.
4. Spatial index library. http://libspatialindex.github.com.
5. Spatialhadoop. http://spatialhadoop.cs.umn.edu/
6. Geos. 2013 http://trac.osgeo.org/geos.
7. Hadoop-GIS wiki. 2013 https://web.cci.emory.edu/confluence/display/hadoopgis.
8. Openstreetmap. 2013 http://www.openstreetmap.org.
9. Pathology analytical imaging standards. 2013 https://web.cci.emory.edu/confluence/display/PAIS.
10. Abouzeid A, Bajda-Pawlikowski K, Abadi D, Silberschatz A, Rasin A. Hadoopdb: an architectural hybrid of mapreduce and dbms technologies for analytical workloads. Proc. VLDB Endow. 2009 Aug;2:922–933.
11. Aji A. Proceedings of the on SIGMOD/PODS 2012 PhD Symposium. ACM; New York, NY, USA: 2012. High performance spatial query processing for large scale scientific data. pp. 9–14.
12. Aji A, Wang F, Saltz JH. Towards building a high performance spatial query system for large scale medical imaging data. SIGSPATIAL/GIS. 2012:309–318. [PMC free article] [PubMed]
13. Akdogan A, Demiryurek U, Banaei-Kashani F, Shahabi C. Voronoi-based geospatial query processing with mapreduce. CLOUDCOM. 2010:9–16.
14. Beckmann N, Kriegel H, Schneider R, Seeger B. The r*-tree: An efficient and robust access method for points and rectangles. SIGMOD. 1990
15. Bercken J. V. d., Seeger B. An evaluation of generic bulk loading techniques. VLDB. 2001:461–470.
16. Blanas S, Patel JM, Ercegovac V, Rao J, Shekita EJ, Tian Y. A comparison of join algorithms for log processing in mapreduce. SIGMOD. 2010
17. Brinkhoff T, Kriegel H-P, Seeger B. Parallel processing of spatial joins using r-trees. ICDE. 1996
18. Cary A, Sun Z, Hristidis V, Rishe N. Experiences on processing spatial data with mapreduce. SSDBM. 2009:302–319.
19. Chaiken R, Jenkins B, Larson P, Ramsey B, Shakib D, Weaver S, Zhou J. SCOPE: easy and efficient parallel processing of massive data sets. PVLDB. 2008;1(2):1265–1276.
20. Dean J, Ghemawat S. Mapreduce: a flexible data processing tool. Commun. ACM. 2010;53(1):72–77.
21. Gates A, Natkovich O, Chopra S, Kamath P, Narayanam S, Olston C, Reed B, Srinivasan S, Srivastava U. Building a high level dataflow system on top of MapReduce: The Pig experience. PVLDB. 2009;2(2):1414–1425.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 5, May 2018

22. Gupta H, Chawda B, Negi S, Faruquie TA, Subramaniam LV, Mohania M. Processing multi-way spatial joins on map-reduce. EDBT. 2013:113–124.
23. Kamel I, Faloutsos C. Hilbert r-tree: An improved r-tree using fractals. VLDB. 1994:500–509.
24. Lee R, Luo T, Huai Y, Wang F, He Y, Zhang X. Ysmart: Yet another sql-to-mapreduce translator. ICDCS. 2011
25. Lo M-L, Ravishankar CV. Spatial hash-joins. SIGMOD. 1996:247–258.
26. Nieto-Santisteban MA, Thakar AR, Szalay AS. Cross-matching very large datasets. NSTC NASA Conference. 2007
27. Olston C, Reed B, Srivastava U, Kumar R, Tomkins A. Pig latin: a not-so-foreign language for data processing. SIGMOD. 2008
28. Patel J, Yu J, Kabra N, Tufte K, Nag B, Burger J, Hall N, Ramasamy K, Lueder R, Ellmann C, Kupsch J, Guo S, Larson J, De Witt D, Naughton J. Building a scaleable geo-spatial dbms: technology, implementation, and evaluation. SIGMOD, SIGMOD '97. 1997:336–347.
29. Pavlo A, Paulson E, Rasin A, Abadi DJ, DeWitt DJ, Madden S, Stonebraker M. A comparison of approaches to large-scale data analysis. SIGMOD. 2009:165–178.
30. Stonebraker M, Abadi DJ, DeWitt DJ, Madden S, Paulson E, Pavlo A, Rasin A. Mapreduce and parallel dbmss: friends or foes? Commun. ACM. 2010;53(1):64–71.
31. Szalay AS, Bell G, vandenBerg J, Wonders A, Burns RC, Fay D, Heasley J, Hey T, Nieto-Santisteban MA, Thakar A, Ingen C. v., Wilton R. Graywulf: Scalable clustered architecture for data intensive computing. HICSS. 2009:1–10.
32. Thusoo A, Sarma JS, Jain N, Shao Z, Chakka P, Anthony S, Liu H, Wyckoff P, Murthy R. Hive: a warehousing solution over a map-reduce framework. Proc. VLDB Endow. 2009 Aug;2(2):1626–1629.
33. Thusoo A, Sarma JS, Jain N, Shao Z, Chakka P, Anthony S, Liu H, Wyckoff P, Murthy R. Hive: a warehousing solution over a map-reduce framework. 2009 Aug;2:1626–1629.
34. Wang F, Kong J, Cooper L, Pan T, Tahsin K, Chen W, Sharma A, Niedermayr C, Oh TW, Brat D, Farris AB, Foran D, Saltz J. A data model and database for high-resolution pathology analytical image informatics. J Pathol Inform. 2011;2(1):32. [PMC free article] [PubMed]
35. Wang F, Kong J, Gao J, Adler D, Cooper L, Vergara-Niedermayr C, Zhou Z, Katigbak B, Kurc T, Brat D, Saltz J. A high-performance spatial database based approach for pathology imaging algorithm evaluation. J Pathol Inform. 4(5):2013. [PMC free article] [PubMed]
36. Wang K, Huai Y, Lee R, Wang F, Zhang X, Saltz JH. Accelerating pathology image data cross-comparison on cpu-gpu hybrid systems. Proc. VLDB Endow. 2012 Jul;5(11):1543–1554. [PMC free article] [PubMed]
37. Xu Y, Kostamaa P, Gao L. Integrating hadoop and parallel dbms. SIGMOD. 2010:969–974.
38. Zhang S, Han J, Liu Z, Wang K, Xu Z. Sjmr: Parallelizing spatial join with mapreduce on clusters. CLUSTER. 2009
39. Zhong Y, Han J, Zhang T, Li Z, Fang J, Chen G. Towards parallel spatial query processing for big spatial data. IPDPSW. 2012:2085–2094.
40. Zhou X, Abel DJ, Truffet D. Data partitioning for parallel spatial join processing. Geoinformatica. 1998 Jun;2:175–204.