JOURNAL OF SPATIAL INFORMATION SCIENCE
Number 21 (2020), pp. 13–24
doi:10.5311/JOSIS.2020.21.719

INVITED ARTICLE

Thinking spatial

Mohamed F. Mokbel

Qatar Computing Research Institute (QCRI)
Hamad Bin Khalifa University (HBKU)
Doha, Qatar

Received: August 25, 2020; accepted: November 5, 2020

Abstract: The systems community in both academia and industry has tremendous success in building widely used general purpose systems for various types of data and applications. Examples include database systems, big data systems, data streaming systems, and machine learning systems. The vast majority of these systems are ill equipped in terms of supporting spatial data. The main reason is that system builders mostly think of spatial data as just one more type of data. Any spatial support can be considered as an afterthought problem that can be supported via on-top functions or spatial cartridges that can be added to the already built systems. This article advocates that spatial data and applications need to be natively supported in special purpose systems, where spatial data is considered as a first class citizen, while spatial operations are built inside the engine rather than on-top of it. System builders should consider spatial data while building their systems. The article gives examples of five categories of systems, namely, database systems, big data systems, machine learning systems, recommender systems, and social network systems, that would benefit tremendously, in terms of both accuracy and performance, when considering spatial data as an integral part of the system engine.

Keywords: spatial databases, big spatial data, spatial machine learning, recommender systems, social networks, location-based services

1 Introduction

Spatial data and applications have been ubiquitous. For example, ride sharing services (e.g., Uber, Lyft, Didi, and Ola) have revolutionized the transportation sector relying on the ability to locate riders, match them with nearby available drivers, and track their routes all the way. As an indication of its ubiquity, Uber had close to two billion rides in Q4 2019 [67]. Mapping services, including routing and navigation, has also been ubiquitous in our daily lives. As of April 2018, Google Maps and Waze had more than 150 and 25 Million users,

respectively [66]. Digital contact tracing [10], which relies on Bluetooth and GPS technology to identify people who were in close contact to each other, has been pointed out as a crucial application in the forefront of fighting and limiting the spread of COVID-19 [26]. Going from smart phones to larger satellite devices, a National Aeronautics and Space Administration (NASA) archive hosts more than 33 Petabytes of public spatial data, with an average daily production growth of 5 Terrabytes, extracted from satellite images [39]. The unprecedented amounts of spatial data produced from various devices, including smart phones and satellites, along with the spatial data produced from or consumed by spatial applications, e.g., ride sharing, mapping services, contact tracing, location-based social networks, and location-based games call for full-fledged efficient systems and algorithms to store, retrieve, and analyze spatial data.

Meanwhile, the systems community has been dealing with the spatial attributes of any object as just one more attribute, with not much special support. System builders mostly build general-purpose systems that are generic enough to handle any kind of attributes. Whenever there is a pressing need for spatial data support, it is considered as an afterthought problem that can be addressed by adding new data types, extensions, or spatial cartridges to existing systems. This ends up in producing over-the-counter systems that can be used by other applications, regardless of their very specific needs.

There are two arguments against designing such special-purpose systems for spatial data and applications. First, how big is the market segment that needs spatial data support? Second, would it be really different from thinking of spatial data support as an afterthought problem? For the first argument, we have already mentioned several applications with a huge market share. For the second argument, we would strongly argue that things would be really different if we start thinking spatial when building our systems. This article will list few examples of systems that would look really different if we start building them while thinking spatial.

2 Classical example: spatial databases

One can easily use a database management system (DBMS) to support a nearest-neighbor query through a simple SQL query that selects object ID from the table of objects, ordered by distance, and limit one on the answer. Yet, this is extremely inefficient due to the need of calculating the distance between the user location and each object in the table and sorting the results. Commercial DBMSs may not care much about this as having a nearest-neighbor query is not a common thing, hence its performance does not hurt much.

Thinking spatial, and considering the ubiquity of spatial data and applications in which nearest-neighbor queries are among the most important ones, we would consider having a specially designed nearest-neighbor operator that can be added to a query plan with other query operators. This also means modifying the query optimizer to consider optimizing query plans with the nearest-neighbor operator, as well as having new spatial index structures to support that important query. As a classical example, there have been tremendous efforts to build spatial databases [28, 47, 62] by enriching the database engine with spatial data types [29], spatial index structures [27], spatial query operators [4], and spatial query optimizers [5]. Such efforts were instrumental in integrating spatial data support in major commercial database system, e.g., Oracle [43] and Microsoft SQL Server [25].

3 Thinking spatial in big data

Various applications and agencies need to process unprecedented amounts of spatial data, produced from several devices such as smart phones, space telescopes, and medical devices. For example, the Blue Brain Project [38] studies the brain's architectural and functional principles through modeling brain neurons as spatial data [68]. Epidemiologists use spatial analysis techniques to identify cancer clusters [42], track infectious disease [6], and drug addiction [69]. Meteorologists study and simulate climate data through spatial data management and analysis [24]. News reporters use geotagged tweets for event detection and analysis [55]. Unfortunately, the immense need to manage big spatial data was hampered by the lack of specialized systems, techniques, and algorithms to support such data. While big (non-spatial) data is well supported with a variety of distributed systems and cloud infrastructure (e.g., Hadoop [30] and Spark [64]), none of these systems or infrastructure were designed to support spatial data. The only way to support big spatial data in such systems is to either treat it as non-spatial data or to write a set of *on-top* functions. However, doing so does not take any advantage of the distinguished properties of spatial data, hence resulting in sub-par performance.

Thinking spatial, one would need to follow a systems approach by providing a native built-in support for spatial data inside the core engine of big data systems. This will allow programs and frameworks running on top of the spatially-aware systems to make use of its embedded spatial functionality. Recent systems that strived for this goal had tremendous success in achieving orders of magnitude better performance than general-purpose big data systems. Such systems include Hadoop-GIS [2] and SpatialHadoop [17] that injected spatial awareness in the Hadoop big data system [30], GeoSpark [71] and Simba [70] that injected spatial awareness inside Spark [64], and Sphinx [23] that injected spatial awareness inside Impala [31]. See [19] for a comprehensive survey.

To clarify how these systems achieve their performance, we will take SpatialHadoop as an example. SpatialHadoop [17, 18], available as free open-source [65], is a full-fledged MapReduce framework with native support for spatial data. SpatialHadoop is a comprehensive extension to Hadoop [30] that injects spatial data awareness in each Hadoop layer, namely, the language, storage, MapReduce, and operations layers. In the language layer, SpatialHadoop employs the Pigeon language [16], which is an extension of the Pig Latin language [40], traditionally used with Hadoop. Pigeon is compatible with the Open Geospatial Consortium (OGC) standard which makes it easy to learn and use for users who are familiar with existing OGC compliant tools such as PostGIS. In the storage layer, SpatialHadoop provides standard spatial indexes, such as grid and R-tree, which are used to store the data in an efficient way in the Hadoop Distributed File System (HDFS) [14]. Indexes are organized in two-layers, one global index that partitions data across nodes, and multiple local indexes to organize records inside each node. The MapReduce layer in SpatialHadoop modifies the original Map-Reduce functionality [12] to balance the query workload over distributed nodes, taking into account the spatial distribution of data and queries. In the query processing layer, SpatialHadoop encapsulates basic spatial operations, range query, nearest-neighbor, and spatial join [49], as well as a suite of fundamental computational geometry operations [15, 34]

Injecting spatial awareness in big spatial data systems has enabled more spatial functionality to be supported. In case of SpatialHadoop, this includes: (a) adding a visualization layer that provides efficient algorithms to visualize big spatial data [21, 22]. SpatialHadoop

supports single level images, which are generated at a fixed resolution, and multilevel images, which are generated at multiple resolutions to allow users to zoom in, (b) providing a backbone support for applications that manage and visualize satellite data [20], and (c) supporting spatio-temporal data and applications [3].

4 Thinking spatial in machine learning

Data scientists and developers have been spending significant efforts applying machine learning techniques on their massive data. However, the skills and efforts needed to deploy such techniques become a major blocking factor in having a wide deployment of machine learning. Markov Logic Network (MLN) [13, 46] was recently introduced to reduce this gap. In particular, MLN is a language representation that combines first-order logic with probabilistic models, and is empowered by machine learning modules that detect patterns and conclude inference. This had a significant impact on the wide deployment of machine learning techniques in various applications, including knowledge base construction [48], data cleaning [44], and information extraction [32], among others. However, MLN is oblivious to spatial data and its distinguishing characteristics, which results in missing important results and having less accuracy in important MLN-based applications that can take advantage of the spatial properties of spatial data.

Meanwhile, as mentioned in Section 3, there is a recent tremendous increase of big spatial data and applications. As a result, various applications and agencies need to take advantage of the recent advances of Markov Logic Networks (MLN) and machine learning techniques to analyze the unprecedented amounts of spatial data. One obvious way to start with is to use the most advanced MLN technology as is with spatial data. While this would work to some extent, it will have a sub-par performance. The main reason is that MLN (and its machine learning techniques) do not have a native support for spatial data. The only way to support spatial data in MLN is to simply ignore its spatial features and deal with it as non-spatial data. However, doing so does not leverage or consider the distinguished properties of spatial data, hence resulting in sub-par performance.

Thinking spatial, one would need to adopt machine learning techniques for big spatial data and applications. This includes going for two orthogonal, but related, directions. First injecting the spatial awareness inside machine learning techniques and applications (e.g., knowledge base construction), which will result in a higher accuracy for such applications. Second, taking advantage of the recent advances in machine learning techniques, in particular Markov Logic Network (MLN), to boost the usability, deployment, scalability, and accuracy of long lasting spatial data analysis techniques.

Along the first direction, knowledge-base construction would be a prime example. Knowledge-base construction has been an active area of research over the last two decades with several system prototypes coming form academia and industry, along with vital applications. DeepDive, an MLN-based system, has emerged as one of the most popular probabilistic knowledge base construction systems [48], applied in vital applications, including geology [72] and paleontology [41]. Unfortunately, probabilistic knowledge base systems do not fully utilize the underlying spatial information, which results in less accuracy in the factual scores. Thinking spatially, the Sya system [50, 51] came as the first spatial probabilistic knowledge base construction system, based on Markov Logic Networks (MLN). Sya injects the awareness of spatial relationships inside the MLN grounding and inference

phases, which are the pillars of the knowledge base construction process, and hence results in a better knowledge base output. Sya provides a simple spatial high-level language, a spatial variation of factor graph, a spatial rules-query translator, and a spatially-equipped statistical inference technique to infer the factual scores of relations. In addition, Sya provides an optimization that ensures scalable grounding and inference for large-scale knowledge bases.

Along the second direction, same as MLN made it possible for data scientists and developers to embrace the difficulty of deploying machine learning techniques, introducing Spatial Markov Logic Networks (SMLN) can act as a backbone infrastructure to support long lasting spatial analysis techniques that lack scalability as well as suffer from difficulty of deployment. In particular, Flash [53] is introduced as a framework for generic and scalable spatial probabilistic graphical modeling (SPGM). SPGM is an important class of spatial data analysis that provides efficient probabilistic graphical models for spatial data. Existing SPGM tools are neither generic nor scalable when dealing with big spatial data. Flash exploits Markov Logic Networks (MLN) to express SPGM as a set of declarative logical rules. In addition, it provides spatial variations of the scalable RDBMS-based learning and inference techniques of MLN to efficiently perform SPGM predictions. Applications of Flash include supporting scalable and accurate execution of autologistic spatial regression that predicts missing values [52, 54].

5 Thinking spatial in recommender systems

Recommender systems make use of community opinions to help users identify useful items from a considerably large search space. For example, recommender systems have successfully been used to help users find interesting books and media from a massive inventory base (Amazon [35]), news items from the Internet (Google News [11]), and movies from a large catalog (Netflix). The technique used by many of these systems is collaborative filtering [1, 45], which analyzes past community opinions to find correlations of similar users and items. Community opinions are expressed through explicit ratings represented by the triple *(user, rating, item)* that represents a user providing a numeric rating for an item. Unfortunately, recommender systems are not friendly to spatial operations. For example, one may want to have recommendations on a restaurant in a certain area, or a tourist wants to get recommendation of items preferred by locals, e.g., "When in Rome, do as the Romans do". Trying to get such spatial recommendations from existing recommender systems would be just a spatial filter on top of existing systems, which is not accurate as this would lose the essence of the collaborative filtering method.

Thinking spatial, the LARS system [33, 60], a Location-Aware Recommender System, injects the spatial awareness inside the core functionality of collaborative filtering rather than being an on-top filter. In particular, LARS goes beyond the rating triple (*user*, *item*, *rating*), which forms the basis of current collaborative filtering methods to a new taxonomy of location-based ratings: (1) *Spatial ratings for non-spatial items*, represented as a four-tuple (*ulocation*, *user*, *rating*, *item*); for example, a user located at home rating a movie, (2) *Non-spatial ratings for spatial items*, represented as a four-tuple (*user*, *rating*, *ilocation*, *item*); for example, a user with unknown location rating a restaurant, and (3) *Spatial ratings for spatial items*, represented as a five-tuple (*ulocation*, *user*, *rating*, *ilocation*, *item*); for example, a user

at office rating a restaurant. With this taxonomy, traditional rating triples can be classified as *non-spatial ratings for non-spatial items*.

With this new taxonomy of location-based ratings, LARS provides scalable query processing that exploits user rating locations through user partitioning; a technique that influences recommendations with ratings spatially close to querying users in a manner that maximizes system scalability while not sacrificing recommendation quality. Meanwhile, LARS exploits item locations using travel penalty; a technique that favors recommendation candidates closer in travel distance to querying users in a way that avoids exhaustive access to all spatial items. LARS can apply these techniques separately, or together, depending on the type of available location-based rating. For system deployment, LARS is realized inside the database engine [56, 61].

6 Thinking spatial in social network

Social networking systems, e.g., Facebook and Twitter, are among the most popular web services nowadays. A common functionality shared by such web services is the *news feed* functionality, where users of social networks receive a feed of news/posts from their friends/groups of interest [63]. Due to the large volume of related news/posts for each user, existing news feed systems opt to select a subset of k relevant news either based on the message timestamp, i.e., most recent k messages, or based on some weighting criteria. Unfortunately, news feed systems mostly ignore the spatial aspect of related messages, and hence, users may miss important messages that are spatially related to them either because they are not so recent or do not make the weighting criteria.

Thinking spatial, one would like to see news feed more related to the current spatial location. For example, let's say that one of my friends has visited Istanbul and posted something about it. I saw the post, and ignored it as it does not really matter to me now, being in Minneapolis. Few months down the road, I am in Istanbul, looking at my news feed. The most important post I would like to see now is the one that was related to Istanbul and posted few months ago. Yet, as the social network is not designed for spatial awareness, it could not recognize that this one is much more important to me now than other posts.

The Sindbad system [57, 58] is a prototype for a location-based social network that each post (a) is associated with a location, and (b) has spatial domain of interest. Sindbad supports three new services beyond traditional social networks, namely, *location-aware news feed* [8], *location-aware ranking* [7], and *location-aware recommender* [9, 59]. These new services not only consider social relevance for its users, but they also consider spatial relevance. Once a Sindbad user logs on to the system, a location-aware news feed query is triggered to retrieve the relevant news feed, i.e., messages posted by the user's friends that have spatial extents covering the location of the requesting user. The output of the location-aware news feed module will be processed further by the location-aware ranking module to get only the top-k news feed based on the spatial and social relevance, which will be returned to the user as the requested news feed. Sindbad users can also request spatial recommendations based on: (a) user location (if available), (b) item location (if available), and (c) ratings previously posted by either the user or the user's friends.

7 Conclusion and other spatial thoughts

This paper makes the case for: (a) using general-purpose systems, where spatial operations are considered as an after-thought problem and supported by ad-hoc on-top functions, always results in sub par performance when dealing with spatial data and applications, and (b) the need to build special purpose systems, where spatial attributes are considered first class citizens and spatial operations are taken into account while building the system. Examples were given for various systems, including database systems (Section 2), Big Data systems (Section 3), machine learning systems (Section 4), recommender systems (Section 5), and social networks (Section 6).

The list of systems that can be redesigned to support spatial data can go on and on to include systems designed for microblogs analysis, crowd sourcing, data streaming, data privacy, data cleaning, and data integration, among others. For example, in microblogs analysis, the TAGHREED system [36, 37] goes beyond the idea of building a general-purpose index structure and query engine to building spatial indexing and spatial query engines for microblogs. In crowdsourcing, many of the tasks are spatially oriented, where the location of the worker plays an important role in adequately performing the task, e.g., rating a restaurant would be preferred to be done locally, geolocating an object that we have a vague idea on its whereabouts would need to be solved by people living around that object. Also, some areas in the world would have more expertise in some jobs than others, e.g., translation and sport-related tasks. In general, having the locations of workers ahead in the equation would change the way that we assign workers to crowdsourcing tasks to achieve better quality.

And, the question comes again. Is it really worth building such systems while Thinking Spatial? I would say definitely yes, it worth it. Spatial information is really special, and it should not be considered as few more attributes.

References

[1] ADOMAVICIUS, G., AND TUZHILIN, A. Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. *IEEE Transactions on Knowledge and Data Engineering, TKDE 17*, 6 (2005), 734–749.

[2] AJI, A., WANG, F., VO, H., LEE, R., LIU, Q., ZHANG, X., AND SALTZ, J. H. Hadoop-GIS: A High Performance Spatial Data Warehousing System over MapReduce. *Proceedings of the International Conference on Very Large Data Bases, VLDB 6*, 11 (2013), 1009–1020.

[3] ALARABI, L., MOKBEL, M. F., AND MUSLEH, M. ST-Hadoop: a MapReduce framework for spatio-temporal data. *GeoInformatica 22*, 4 (2018), 785–813.

[4] AREF, W. G., AND SAMET, H. Extending a DBMS with Spatial Operations. In *Proceedings of the International Symposium on Advances in Spatial Databases, SSD* (1991), pp. 299–318.

[5] AREF, W. G., AND SAMET, H. Optimization for Spatial Query Processing. In *Proceedings of the International Conference on Very Large Data Bases, VLDB* (1991), pp. 81–90.

[6] AUCHINCLOSS, A., GEBREAB, S., MAIR, C., AND ROUX, A. D. A Review of Spatial Methods in Epidemiology: 2000-2010. *Annual Review of Public Health 33* (2012), 107–22.

[7] BAO, J., AND MOKBEL, M. F. GeoRank: An Efficient Location-Aware News Feed Ranking System. In *Proceedings of the ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM SIGSPATIAL GIS* (Orlando, Florida, 2013).

[8] BAO, J., MOKBEL, M. F., AND CHOW, C.-Y. GeoFeed: A Location-Aware News Feed System. In *Proceedings of the International Conference on Data Engineering, ICDE* (Washington D.C., 2012).

[9] BAO, J., ZHENG, Y., AND MOKBEL, M. Location-based and Preference-Aware Recommendation Using Sparse Geo-Social Networking Data. In *Proceedings of the ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM SIGSPATIAL GIS* (Redondo Beach, CA, 2012).

[10] CENTERS FOR DISEASE CONTROL AND PREVENTION. Digital Contact Tracing Tools. https://www.cdc.gov/coronavirus/2019-ncov/php/contact-tracing/contact-tracing-plan/digital-contact-tracing-tools.html. Last Accessed July 25, 2020.

[11] DAS, A., DATAR, M., GARG, A., AND RAJARAM, S. Google News Personalization: Scalable Online Collaborative Filtering. In *Proceedings of the International Conference on World Wide Web, WWW* (Banff, Canada, 2007), pp. 271–280.

[12] DEAN, J., AND GHEMAWAT, S. MapReduce: Simplified Data Processing on Large Clusters. *Communications of ACM 51* (2008), 107–113.

[13] DOMINGOS, P., AND LOWD, D. *Markov Logic: An Interface Layer for Artificial Intelligence*. Morgan and Claypool Publishers, 2009.

[14] ELDAWY, A., ALARABI, L., AND MOKBEL, M. F. Spatial Partitioning Techniques in SpatialHadoop. In *Proceedings of the International Conference on Very Large Data Bases, VLDB* (2015), pp. 1602–1605.

[15] ELDAWY, A., LI, Y., MOKBEL, M. F., AND JANARDAN, R. CG_Hadoop: Computational Geometry in MapReduce. In *Proceedings of the ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM SIGSPATIAL GIS* (2013), pp. 284–293.

[16] ELDAWY, A., AND MOKBEL, M. Pigeon: A Spatial MapReduce Language. In *Proceedings of the International Conference on Data Engineering, ICDE* (Chicago, IL, 2014).

[17] ELDAWY, A., AND MOKBEL, M. F. A Demonstration of SpatialHadoop: An Efficient MapReduce Framework for Spatial Data (System Demo). In *Proceedings of the International Conference on Very Large Data Bases, VLDB* (Riva del Garda, Italy, 2013).

[18] ELDAWY, A., AND MOKBEL, M. F. SpatialHadoop: A MapReduce Framework for Spatial Data. In *Proceedings of the International Conference on Data Engineering, ICDE* (2015), pp. 1352–1363.

[19] ELDAWY, A., AND MOKBEL, M. F. The Era of Big Spatial Data: A Survey. *Foundantion and Trends in Databases 6*, 3-4 (2016), 163–273.

[20] ELDAWY, A., MOKBEL, M. F., ALHARTHI, S., ALZAIDY, A., TAREK, K., AND GHANI, S. SHAHED: A MapReduce-based System for Querying and Visualizing Spatio-temporal Satellite Data. In *ICDE* (2015), pp. 1585–1596.

[21] ELDAWY, A., MOKBEL, M. F., AND JONATHAN, C. A Demonstration of HadoopViz: An Extensible MapReduce System for Visualizing Big Spatial Data. *Proceedings of the International Conference on Very Large Data Bases, VLDB 8*, 12 (2015), 1896–1907.

[22] ELDAWY, A., MOKBEL, M. F., AND JONATHAN, C. HadoopViz: A MapReduce Framework for Extensible Visualization of Big Spatial Data. In *Proceedings of the International Conference on Data Engineering, ICDE* (2015), pp. 601–612.

[23] ELDAWY, A., SABEK, I., ELGANAINY, M., BAKEER, A., ABDELMOTALEB, A., AND MOKBEL, M. F. Sphinx: Empowering Impala for Efficient Execution of SQL Queries on Big Spatial Data. In *Proceedings of the International Symposium on Advances in Spatial and Temporal Databases, SSTD* (2017), pp. 65–83.

[24] FAGHMOUS, J., AND KUMAR, V. *Spatio-Temporal Data Mining for Climate Data: Advances, Challenges, and Opportunities*. Advances in Data Mining, Springer, 2013.

[25] FANG, Y., FRIEDMAN, M., NAIR, G., RYS, M., AND SCHMID, A. Spatial Indexing in Microsoft SQL Server 2008. In *Proceedings of the ACM International Conference on Management of Data, SIGMOD* (2008), pp. 1207–1216.

[26] FERRETTI, L., WYMANT, C., KENDALL, M., ZHAO, L., NURTAY, A., ABELER-DORNER, L., PARKER, M., BONSALL, D., AND FRASER, C. Quantifying SARS-CoV-2 Transmission suggests Epidemic Control with Digital Contact Tracing. *Science 368* (2020), 621–626.

[27] GAEDE, V., AND GÜNTHER, O. Multidimensional Access Methods. *ACM Computing Surveys 30*, 2 (1998), 170–231.

[28] GÜTING, R. H. An Introduction to Spatial Database Systems. *VLDB Journal 3*, 4 (1994), 357–399.

[29] GÜTING, R. H., AND SCHNEIDER, M. Realms: A Foundation for Spatial Data Types in Database Systems. In *Proceedings of the International Symposium on Advances in Spatial Databases, SSD* (1993), pp. 14–35.

[30] HADOOP. Apache Hadoop. http://hadoop.apache.org/. Last Accessed July 25, 2020.

[31] IMPALA. Apache Impala. https://impala.apache.org/. Last Accessed July 25, 2020.

[32] JIANG, S., LOWD, D., AND DOU, D. Learning to Refine an Automatically Extracted Knowledge Base Using Markov Logic. In *Proceedings of the IEEE International Conference on Data Mining, ICDM* (2012).

[33] LEVANDOSKI, J. J., SARWAT, M., ELDAWY, A., AND MOKBEL, M. F. LARS: A Location-Aware Recommender System. In *Proceedings of the International Conference on Data Engineering, ICDE* (2012), pp. 450–461.

[34] LI, Y., ELDAWY, A., XUE, J., KNOROZOVA, N., MOKBEL, M. F., AND JANARDAN, R. Scalable computational geometry in MapReduce. *VLDB Journal 28*, 4 (2019), 523–548.

[35] LINDEN, G., SMITH, B., AND YORK, J. Amazon.com Recommendations: Item-to-Item Collaborative Filtering. *IEEE Internet Computing 7*, 1 (2003), 76–80.

[36] MAGDY, A., ALARABI, L., AL-HARTHI, S., MUSLEH, M., GHANEM, T., GHANI, S., AND MOKBEL, M. F. Taghreed: A System for Querying, Analyzing, and Visualizing Geotagged Microblogs. In *Proceedings of the ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM SIGSPATIAL GIS* (Dallas, TX, 2014).

[37] MAGDY, A., ALARABI, L., AL-HARTHI, S., MUSLEH, M., GHANEM, T., GHANI, S., AND MOKBEL, M. F. Demonstration of Taghreed: A System for Querying, Analyzing, and Visualizing Geotagged Microblogs. In *Proceedings of the International Conference on Data Engineering, ICDE* (Seoul, South Korea, 2015).

[38] MAKRAM, H. The Blue Brain Project. *Nature Reviews Neuroscience 7* (2006), 153–160.

[39] NATIONAL AERONAUTICS AND SPACE ADMINISTRATION (NASA). Earth Science Data and Information System Annual Metrics Reports. https://earthdata.nasa.gov/eosdis/system-performance/eosdis-annual-metrics-reports. Last Accessed July 25, 2020.

[40] OLSTON, C., REED, B., SRIVASTAVA, U., KUMAR, R., AND TOMKINS, A. Pig Latin: A Not-so-foreign Language for Data Processing. In *Proceedings of the ACM International Conference on Management of Data, SIGMOD* (2008), pp. 1099–1110.

[41] PETERS, S. E., ZHANG, C., LIVNY, M., AND RÉ, C. A Machine Reading System for Assembling Synthetic Paleontological Databases. *PLoS One 9*, 12 (2014).

[42] PICKLE, L., SZCZUR, M., LEWIS, D., , AND STINCHCOMB, D. The Crossroads of GIS and Health Information: A Workshop on Developing a Research Agenda to Improve Cancer Control. *International Journal of Health Geographics 5*, 1 (2006), 51.

[43] RAVADA, S., AND SHARMA, J. Oracle8i Spatial: Experiences with Extensible Databases. In *Proceedings of the International Symposium on Advances in Spatial Databases, SSD* (1999), pp. 355–359.

[44] REKATSINAS, T., CHU, X., ILYAS, I. F., AND RÉ, C. HoloClean: Holistic Data Repairs with Probabilistic Inference. *VLDB Journal 10*, 11 (2017), 1190–1201.

[45] RESNICK, P., IACOVOU, N., SUCHAK, M., BERGSTROM, P., AND RIEDL, J. GroupLens: An Open Architecture for Collaborative Filtering of Netnews. In *Proceedings of the Conference on Computer Supported Cooperative Work, CSWC* (Chapel Hill, NC, 1994), pp. 175–186.

[46] RICHARDSON, M., AND DOMINGOS, P. M. Markov Logic Networks. *Machine Learning 62*, 1-2 (2006), 107–136.

[47] RIGAUX, P., SCHOLL, M., AND VOISARD, A. *Spatial Databases with Application to GIS*. Morgan Kaufmann, 2002.

[48] SA, C. D., RATNER, A., RÉ, C., SHIN, J., WANG, F., WU, S., AND ZHANG, C. DeepDive: Declarative Knowledge Base Construction. *ACM SIGMOD Record 45*, 1 (2016), 60–67.

[49] SABEK, I., AND MOKBEL, M. F. On Spatial Joins in MapReduce. In *Proceedings of the ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM SIGSPATIAL GIS* (2017), pp. 1–10.

[50] SABEK, I., AND MOKBEL, M. F. Sya: Enabling Spatial Awareness inside Probabilistic Knowledge Base Construction. In *Proceedings of the International Conference on Data Engineering, ICDE* (2020), pp. 1177–1188.

[51] SABEK, I., MUSLEH, M., AND MOKBEL, M. F. A Demonstration of Sya: A Spatial Probabilistic Knowledge Base Construction System. In *Proceedings of the ACM International Conference on Management of Data, SIGMOD* (2018), pp. 1689–1692.

[52] SABEK, I., MUSLEH, M., AND MOKBEL, M. F. TurboReg: a framework for scaling up spatial logistic regression models. In *Proceedings of the ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM SIGSPATIAL GIS* (2018), pp. 129–138.

[53] SABEK, I., MUSLEH, M., AND MOKBEL, M. F. Flash in Action: Scalable Spatial Data Analysis Using Markov Logic Networks. *Proceedings of the International Conference on Very Large Data Bases, VLDB 12*, 12 (2019), 1834–1837.

[54] SABEK, I., MUSLEH, M., AND MOKBEL, M. F. RegRocket: Scalable Multinomial Autologistic Regression with Unordered Categorical Variables Using Markov Logic Networks. *ACM Transactions on Spatial Algorithms and Systems 5*, 4 (2019), 1–27.

[55] SANKARANARAYANAN, J., SAMET, H., TEITLER, B. E., LIEBERMAN, M. D., AND SPERLING, J. TwitterStand: News in Tweets. In *Proceedings of the ACM Symposium on Advances in Geographic Information Systems, ACM GIS* (2009), pp. 42–51.

[56] SARWAT, M., AVERY, J. L., AND MOKBEL, M. F. RECATHON: A Middleware for Context-Aware Recommendation in Database Systems. In *Proceedings of the International Conference on Mobile Data Management, MDM* (2015), pp. 54–63.

[57] SARWAT, M., BAO, J., ELDAWY, A., LEVANDOSKI, J. J., MAGDY, A., AND MOKBEL, M. F. The Anatomy of Sindbad: A Location-Aware Social Networking System. In *In Proceedings of the ACM SIGSPATIAL International Workshop on Location-Based Social Networks, LBSN, co-located with SIGSPATIAL GIS* (Redondo Beach, CA, 2012).

[58] SARWAT, M., BAO, J., LEVANDOSKI, J. J., ELDAWY, A., AND MOKBEL, M. F. Sindbad: A Location-aware Social Networking System (System Demo). In *Proceedings of the ACM International Conference on Management of Data, SIGMOD* (Scottdale, AZ, 2012).

[59] SARWAT, M., ELDAWY, A., MOKBEL, M. F., AND RIEDL, J. Plutus: Leveraging Location-Based Social Networks in Recommending Potential Customers to Venues. In *Proceedings of the International Conference on Mobile Data Management, MDM* (Milan, Italy, 2013).

[60] SARWAT, M., LEVANDOSKI, J. J., ELDAWY, A., AND MOKBEL, M. F. LARS*: An Efficient and Scalable Location-Aware Recommender System. *IEEE Transactions on Knowledge and Data Engineering, TKDE 26*, 6 (2014), 1384–1399.

[61] SARWAT, M., MORAFFAH, R., MOKBEL, M. F., AND AVERY, J. L. Database System Support for Personalized Recommendation Applications. In *Proceedings of the International Conference on Data Engineering, ICDE* (2017), pp. 1320–1331.

[62] SHEKHAR, S., AND CHAWLA, S. *Spatial Databases: A Tour*. Prentice Hall, 2003.

[63] SILBERSTEIN, A., TERRACE, J., COOPER, B. F., AND RAMAKRISHNAN, R. Feeding Frenzy: Selectively Materializing User's Event Feed. In *Proceedings of the ACM International Conference on Management of Data, SIGMOD* (Indianapolis, IN, 2010), pp. 831–842.

[64] SPARK. Apache Spark. https://spark.apache.org/. Last Accessed July 25, 2020.

[65] SPATIALHADOOP. SpatialHadoop. http://spatialhadoop.cs.umn.edu/. Last Accessed July 25, 2020.

[66] STATISTA. Most popular mapping apps in the United States as of April 2018, by monthly users. https://www.statista.com/statistics/865413/most-popular-us-mapping-apps-ranked-by-audience/. Last Accessed July 25, 2020.

[67] STATISTA. Number of rides Uber gave worldwide from Q2 2017 to Q1 2020. https://www.statista.com/statistics/946298/uber-ridership-worldwide/. Last Accessed July 25, 2020.

[68] TAUHEED, F., BIVEINIS, L., HEINIS, T., SCHURMANN, F., MARKRAM, H., AND AILAMAKI, A. Accelerating Range Queries for Brain Simulations. In *Proceedings of the International Conference on Data Engineering, ICDE* (2012), pp. 941–952.

[69] THOMAS, Y., RICHARDSON, D., AND CHEUNG, I. Geography and Drug Addiction. *Springer Verlag* (2009).

[70] XIE, D., LI, F., YAO, B., LI, G., ZHOU, L., AND GUO, M. Simba: Efficient In-Memory Spatial Analytics. In *Proceedings of the 2016 International Conference on Management of Data (ACM SIGMOD 2016)* (2016), pp. 1071–1085.

[71] YU, J., SARWAT, M., AND WU, J. GeoSpark: A Cluster Computing Framework for Processing Large-Scale Spatial Data. In *Proceedings of the ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM SIGSPATIAL GIS* (Seattle, WA, 2015).

[72] ZHANG, C., GOVINDARAJU, V., BORCHARDT, J., FOLTZ, T., RÉ, C., AND PETERS, S. GeoDeepDive: Statistical Inference Using Familiar Data-processing Languages. In *Proceedings of the ACM International Conference on Management of Data, SIGMOD* (2013).