Jia Yu

Assistant Professor

jia.yu1@wsu.edu | jiayuasu.github.io | jiayuasu | jia-yu-27632182

Research Interests

Database Systems, Geospatial Data Management

Education

Arizona State University *Tempe, Arizona, U.S.A*

Doctor of Philosophy in Computer Science *Sept. 2013 - Summer 2020*

- Advisor: Assistant Professor Mohamed Sarwat (started in January 2015)
- Thesis: System Support for Large-scale Geospatial Data Analytics
- Thesis committee: Mohamed Sarwat, Kasim Selcuk Candan, Ming Zhao, Wenwen Li (from ASU geography department)

Northwest Agriculture and Forestry University *Yangling, Shaanxi, China*

Bachelor of Engineering in Software Engineering *Sept. 2009 - Jul. 2013*

- Outstanding Graduate (200 / 5600)

Employment History

Washington State University *Pullman, Washington, U.S.A*

Tenure-track Assistant Professor, School of Electrical Engineering and Computer Science *Aug. 2020 - Present*

Microsoft Research *Redmond, Washington, U.S.A*

Research Intern, Database Group *Jun. 2019 - Aug. 2019*

- Mentor / Collaborators: Umar Farooq Minhas, David Lomet, Jaeyoung Do, Yinan Li, Chi Wang, Badrish Chandramouli, Johannes Gehrke, Donald Kossmann

IBM Almaden Research Center *San Jose, California, U.S.A*

Research Intern, Exploratory Databases Group *May. 2018 - Aug. 2018*

- Mentor / Collaborators: Vijayshankar Raman, Yingjun Wu, Yuanyuan Tian, Ronald Barber, and Richard Sidle

Apple *Cupertino, California, U.S.A*

Software Development Intern, Apple Map Team *Jun. 2016 - Aug. 2016*

- Mentor / Collaborators: Huang-Hsiang Cheng, Alex Radeski
- I worked on cluster computing frameworks and resource management systems such as Apache Spark and Apache Mesos. I developed internal evaluation tools to assist in large-scale spatial analysis.

Publications

As of 01/10/2022, Google Scholar citations: 730, h-index: 9

Based on both selectivity and impact, computer science conferences are considered as important as journals (source: National Academies Press)

Peer-reviewed journal publications

J1 **Jia Yu** and Mohamed Sarwat. GeoSparkViz: A Cluster Computing System for Visualizing Massive-Scale Geospatial Data. *The VLDB Journal*, 30(2):237--258, 2021a. doi:10.1007/s00778-020-00645-2 (pdf)

J2 **Jia Yu**, Zishan Fu, and Mohamed Sarwat. Dissecting GeoSparkSim: a scalable microscopic road network traffic simulator in Apache Spark. *Distributed Parallel Databases Journal*, 38(4):963--994, 2020a. doi:10.1007/s10619-020-07306-x (pdf)

J3 **Jia Yu**, Zongsi Zhang, and Mohamed Sarwat. Spatial Data Management in Apache Spark: the GeoSpark Perspective and Beyond. *GeoInformatica Journal*, 23(1):37--78, 2019a. doi:10.1007/s10707-018-0330-9 (pdf)

Peer-reviewed conference publications

C1 Yiqun Xie, Xiaowei Jia, Han Bao, Xun Zhou, **Jia Yu**, Rahul Ghosh, and Praveen Ravirathinam. Spatial-Net: A Self-Adaptive and Model-Agnostic Deep Learning Framework for Spatially Heterogeneous Datasets. In *International Conference on Advances in Geographic Information Systems (SIGSPATIAL)*, pages 313--323, 2021. doi:10.1145/3474717.3483970

C2 **Jia Yu**, Kanchan Chowdhury, and Mohamed Sarwat. Tabula in action: A sampling middleware for interactive geospatial visualization dashboards. *Proceedings of the VLDB Endowment (PVLDB)*, 13(12):2925--2928, 2020b. doi:10.14778/3415478.3415510 (Demo paper) (pdf)

C3 Jialin Ding, Umar Farooq Minhas, **Jia Yu**, Chi Wang, Jaeyoung Do, Yinan Li, Hantian Zhang, Badrish Chandramouli, Johannes Gehrke, Donald Kossmann, David B. Lomet, and Tim Kraska. ALEX: An Updatable Adaptive Learned Index. In *Proceedings of the ACM International Conference on Management of Data (SIGMOD)*, pages 969--984, 2020. doi:10.1145/3318464.3389711 (16-page version, 21-page MSR technical report)

C4 **Jia Yu** and Mohamed Sarwat. Turbocharging Geospatial Visualization Dashboards via a Materialized Sampling Cube Approach. In *Proceedings of the International Conference on Data Engineering (ICDE)*, pages 1165--1176, 2020. doi:10.1109/ICDE48307.2020.00105 (pdf)

C5 **Jia Yu** and Mohamed Sarwat. Spatial data wrangling with geospark: A step by step tutorial. In *International Conference on Advances in Geographic Information Systems Spatial API Workshop, ACM SIGSPATIAL Spatial API Workshop*, pages 3:1--3:2, 2019a. doi:10.1145/3356394.3365589 (Tutorial) (pdf)

C6 Zishan Fu, **Jia Yu**, and Mohamed Sarwat. Building Microscopic Road Network Traffic Simulators in Apache Spark. In *Proceedings of the International Conference on Mobile Data Management (MDM)*, pages 320--328, 2019a. doi:10.1109/MDM.2019.00-42 (pdf)

C7 Zishan Fu, **Jia Yu**, and Mohamed Sarwat. Demonstrating GeoSparkSim: A Scalable Microscopic Road Network Traffic Simulator Based on Apache Spark. In *Proceedings of the International Symposium on Advances in Spatial and Temporal Databases (SSTD)*, pages 186--189, 2019b. doi:10.1145/3340964.3340984 (Demo paper) (pdf) **[Best Demo Paper Runner-Up]**

C8 Yingjun Wu, **Jia Yu**, Yuanyuan Tian, Ronald Barber, and Richard Sidle. Designing Succinct Secondary Indexing Mechanism by Exploiting Column Correlations. In *Proceedings of the ACM International Conference on Management of Data (SIGMOD)*, pages 1223--1240, 2019a. doi:10.1145/3299869.3319861 (pdf)

C9 Yingjun Wu, **Jia Yu**, Yuanyuan Tian, Richard Sidle, and Ronald Barber. HERMIT in action: Succinct secondary indexing mechanism via correlation exploration. *Proceedings of the VLDB Endowment (PVLDB)*, 12(12):1882--1885, 2019b. doi:10.14778/3352063.3352090 (Demo paper) (pdf)

C10 **Jia Yu** and Mohamed Sarwat. Geospatial Data Management in Apache Spark: A Tutorial. In *Proceedings of the International Conference on Data Engineering (ICDE)*, pages 2060--2063, 2019b. doi:10.1109/ICDE.2019.00239 (Tutorial) (pdf, website)

C11 Yuhan Sun, **Jia Yu**, and Mohamed Sarwat. Demonstrating Spindra: A Geographic Knowledge Graph Management System. In *Proceedings of the International Conference on Data Engineering (ICDE)*, pages 2044--2047, 2019. doi:10.1109/ICDE.2019.00235 (Demo paper) (pdf)

C12 **Jia Yu**, Anique Tahir, and Mohamed Sarwat. GeoSparkViz in Action: A Data System with built-in support for Geospatial Visualization. In *Proceedings of the International Conference on Data Engineering (ICDE)*, pages 1992--1995, 2019b. doi:10.1109/ICDE.2019.00222 (Demo paper) (pdf)

C13 **Jia Yu**, Zongsi Zhang, and Mohamed Sarwat. GeoSparkViz: a scalable geospatial data visualization framework in the Apache Spark ecosystem. In *Proceedings of the International Conference on Scientific and Statistical Database Management (SSDBM)*, pages 15:1--15:12, 2018. doi:10.1145/3221269.3223040 (pdf)

C14 **Jia Yu** and Mohamed Sarwat. Indexing the Pickup and Drop-Off Locations of NYC Taxi Trips in PostgreSQL - Lessons from the Road. In *Proceedings of the International Symposium on Advances in Spatial and Temporal Databases (SSTD)*, pages 145--162, 2017 (pdf)

C15 **Jia Yu**. SRC: Geospatial Visual Analytics Belongs to Database Systems: the BABYLON approach. *International Conference on Advances in Geographic Information Systems (SIGSPATIAL)*, 9(3):2--3, 2017. doi:10.1145/3178392.3178394 (Extended Abstract) (pdf) **[Third Place of ACM SIGSPATIAL Student Research Competition)**

C16 **Jia Yu**, Raha Moraffah, and Mohamed Sarwat. Hippo in Action: Scalable Indexing of a Billion New York City Taxi Trips and Beyond. In *Proceedings of the International Conference on Data Engineering (ICDE)*, pages 1413--1414, 2017. doi:10.1109/ICDE.2017.201 (Demo paper) (pdf)

C17 **Jia Yu** and Mohamed Sarwat. Two Birds, One Stone: A Fast, yet Lightweight, Indexing Scheme for Modern Database Systems. *Proceedings of the VLDB Endowment (PVLDB)*, 10(4):385--396, 2016. doi:10.1109/ICDE.2016.7498357 (pdf)

C18 **Jia Yu**, Jinxuan Wu, and Mohamed Sarwat. A demonstration of GeoSpark: A cluster computing framework for processing big spatial data. In *Proceedings of the International Conference on Data Engineering (ICDE)*, pages 1410--1413, 2016. doi:10.1109/ICDE.2016.7498357 (Demo paper) (pdf)

C19 **Jia Yu**, Jinxuan Wu, and Mohamed Sarwat. GeoSpark: a cluster computing framework for processing large-scale spatial data. In *International Conference on Advances in Geographic Information Systems (SIGSPATIAL)*, pages 70:1--70:4, 2015. doi:10.1145/2820783.2820860 (Short paper) (pdf)

C20 Zijiang Yang, Bei-Bei Yin, Junpeng Lv, Kai-Yuan Cai, Stephen S. Yau, and **Jia Yu**. Dynamic Random Testing with Parameter Adjustment. In *IEEE Annual Computer Software and Applications Conference, COMPSAC Workshops*, pages 37--42, 2014. doi:10.1109/COMPSACW.2014.10

C21 Lei Zhang, Bei-Bei Yin, Junpeng Lv, Kai-Yuan Cai, Stephen S. Yau, and **Jia Yu**. A History-Based Dynamic Random Software Testing. In *IEEE Annual Computer Software and Applications Conference, COMPSAC Workshops*, pages 31--36, 2014. doi:10.1109/COMPSACW.2014.9

Book chapter

B1 **Jia Yu** and Mohamed Sarwat. Big Geospatial Data Processing Made Easy: A Working Guide to GeoSpark. In *Handbook of Big Geospatial Data*, pages 35--53. Springer, 2021b

Patent

P1 Jia Yu, Zongsi Zhang, and Mohamed Sarwat. Systems and methods for an end-to-end visual analytics system for massive-scale geospatial data, November 30 2021. US Patent 11,188,557

Extramural Funding

ICESpark: An Open-Source Big Data Platform for Science Discoveries in the New Arctic and Beyond

National Science Foundation
EarthCube program
2021

TOTAL BUDGET $1,249,289, MY SHARE $293,452

- Lead PI: Jia Yu; Co-PI: Yiqun Xie (UMD PI), Sinead Farrell (UMD), George Hurtt (UMD), Jacob Wenegrat (UMD)
- Funding agency: National Science Foundation (NSF)
- NSF Award page 1 (WSU), NSF Award page 2 (UMD)

AI Institute: Agricultural AI for Transforming Workforce and Decision Support (AgAID)

National Science Foundation
National Artificial Intelligence (AI) Research Institutes program
2021

TOTAL BUDGET $20,000,000, MY SHARE $200,000

- Role: Co-PI; PI: Ananth Kalyanaraman
- Funding agency: NSF & United States Department of Agriculture
- NSF announcement, USDA-NIFA announcement, WSU News

Teaching Experience

SP 2022	**Instructor**, CPT_S 233 Advanced Data Structures Java (14 students), I redesigned the course materials	*WSU*
FA 2021	**Instructor**, CPT_S 415 Big Data (88 students)	*WSU*
SP 2021	**Instructor**, CPT_S 223 Advanced Data Structures C++ (100 students), I redesigned the course materials	*WSU*
FA 2020	**Instructor**, CPT_S 415 Big Data (70 students), I redesigned the course materials especially the course project	*WSU*
2019	**Instructor**, CSE511 Data Processing at Scale (graduate, introduction)	*ASU*
2019	**Teaching Assistant**, CSE412 Database Management (undergraduate)	*ASU*
2018	**Coursera course designer**, ASU MS Degree of CS: Data Systems, over 10K learners (statistics)	*ASU*
2018	**Teaching Assistant**, CSE412 Database Management (undergraduate)	*ASU*
2016	**Teaching Assistant**, CSE512 Distributed Database Systems (graduate)	*ASU*
2015	**Teaching Assistant**, CSE512 Distributed Database Systems (graduate)	*ASU*
2014	**Teaching Assistant**, CSE543 Information Assurance (graduate)	*ASU*
2014	**Teaching Assistant**, CSE240 Introduction to Programming Languages (undergraduate)	*ASU*

Students

Abrar Akhyer Abir

Washington State University

PhD student in Computer Science

Spring 2022 - present

- Scalable spatial query processing

Shengya Zhang

Washington State University

PhD student in Computer Science

Fall 2021 - present

- Distributed geospatial stream processing

Congying Wang

Washington State University

PhD student in Computer Science

Spring 2021 - present

- Lightweight geospatial data structures

Mentoring experiences at Arizona State University

PhD students: Ankita Sharma, Kanchan Chowdhury

Master students: Zishan Fu, Anique Tahir, Zongsi Zhang, Jinxuan Wu

Software Artifacts

Apache Sedona (formerly GeoSpark): a cluster computing framework for processing big spatial data

Main developer

- I am leading the Apache Sedona (incubating), a cluster computing framework for processing big spatial data, and collaborating with more than 50 contributors from the community. In particular, I contributed to the following components in Sedpna ecosystem: (1) distributed computation engine (Sedona) (2) distributed visualization engine (SedonaViz) (3) distributed data simuator (SedonaSim).
- I have published several papers featuring different components of Apache Sedona ecosystem.
    1. Sedona - research paper: Geoinformatica Journal (J3); demo paper: ICDE 2016 (C18); tutorial: ICDE 2019 (C10), SIGSPATIAL 2019 (C5); short paper: SIGSPATIAL 2015 (C19).
    2. SedonaViz - research paper: VLDB Journal 2021 (J1), SSDBM 2018 (C13); demo paper: ICDE 2019 (C12); short paper: SIGSPATIAL 2017 Student Research Competition (C15).
    3. SedonaSim - research paper: MDM 2019 (C6); demo paper: SSTD 2019 (C7, Best Demo Paper Runner-Up).
- I maintain Apache Sedona GitHub source code and Apache Sedona website. Sedona ecosystem receives 300K downloads per month. Users and contributors are from Facebook, Uber, MoBike, SafeGraph and numerous startups. Databricks (the tech unicorn behind Apache Spark) featured Sedona in its blog post and provides an interactive Sedona notebook for its Spark Cloud.

Tabula: turbocharge geospatial visualization dashboards via a materialized sampling cube

Main developer

- Tabula is a sampling middleware that sits between the data system and the geospatial visualization dashboard to accelerate the interactive visual analysis.
- Research paper: ICDE 2020 (C4), demo paper: VLDB 2020 (C2).
- Implemented Tabula in SparkSQL, open-sourced the system on GitHub and released a video to demonstrate how Tabula works.

Hippo: a fast yet lightweight database indexing scheme

Main developer

- Hippo is a fast yet lightweight database indexing scheme. It significantly shrinks the index storage and mitigates maintenance overhead without compromising much on the query execution performance.
- Research paper: VLDB 2016 (C17), SSTD 2017 (C14); demo paper: ICDE 2017 (C16).
- Hippo is a PostgreSQL 9.6 built-in index. I open-sourced Hippo on GitHub and released a video to demonstrate how Hippo works.

ALEX: an updatable adaptive learned index

Team member

- Collaborated with researchers in Microsoft Research Database group to design a new set of model-based learned index structures called ALEX, that work beyond static data. ALEX reorganizes storage structures for index nodes that support efficient data insertion.
- Research paper: SIGMOD 2020 (C3).
- Alex is open-source on GitHub

Professional Services

Program Committee Chairs:

● SpatialAPI 2021: ACM SIGSPATIAL Workshop on APIs and Libraries for Geospatial Data Science (SpatialAPI workshop)

● SpatialEpi 2021: ACM SIGSPATIAL Workshop on Spatial Computing for Epidemiology (SpatialEpi workshop)

● COVID 2020: ACM SIGSPATIAL Workshop on Modeling and Understanding the Spread of COVID-19 (COVID workshop)

Program Committee members:

● ACM International Conference on Management of Data (SIGMOD): 2023

● International Conference on Very Large Data Bases (VLDB): 2021 (demo track)

● ACM International Conference on Advances in Geographic Information Systems (SIGSPATIAL): 2020, 2021

● Symposium on Spatial and Temporal Databases (SSTD): 2021

● IEEE International Conference on Mobile Data Management (MDM): 2022

Journal reviewers (publons profile):

●ACM Transactions on Spatial Algorithms and Systems (ACM TSAS)

●VLDB Journal

● IEEE Transactions on Knowledge and Data Engineering (TKDE)

●International Journal of Geographical Information Science (IJGIS)

●Geoinformatica

● IEEE Transactions on Visualization and Computer Graphics (TVCG)

● Distributed and Parallel Databases (DAPD)

●IEEE Transactions on Cloud Computing

● Journal of Supercomputing

●Computers and Geosciences (CAGEO)

- IEEE Transactions on Parallel and Distributed Systems
- Geo-Spatial Information Science
- IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)
- Frontiers in Big Data

Professional Society Memberships

Association for Computing Machinery (ACM)

Institute of Electrical and Electronics Engineers (IEEE)

Honors & Awards

2019	**Best Demo Paper Runner-Up**, SSTD	*Vienna, Austria*
2019	**Engineering Graduate Fellowship**, Ira A. Fulton Schools of Engineering	*ASU, U.S.A*
2016 - 19	**NSF Student Travel Grant (3 times)**, IEEE ICDE	
2015 - 19	**NSF Student Travel Grant (4 times), Microsoft Student Travel Grant**, ACM SIGSPATIAL	
2017	**Third Place of Student Research Competition**, ACM SIGSPATIAL	*Los Angeles, U.S.A*
2013	**Outstanding Graduate**, Northwest A & F University	*Yangling, China*
2011 - 12	**First-class Scholarship (2 times)**, Northwest A & F University	*Yangling, China*
2011 - 12	**Merit Student (2 times)**, Northwest A & F University	*Yangling, China*

Presentation

- Conference talks (9 times): VLDB, ICDE, SIGSPATIAL, SSTD, MDM, ApacheCon (Apache Software Foundation annual conference)
- Company talks (7 times): Microsoft Research, IBM Almaden Research Center, Apple, NVidia, StateFarm, Vocareum
- The slides of my talks are usually available. But some talks are confidential due to NDA.

Turbocharging Geospatial Visualization Dashboards via a Materialized Sampling Cube Approach — *Dallas, Texas, U.S.A*

ICDE 2020 — *April. 2020*

- Talk slides

Spatial Data Wrangling With GeoSpark: A Step-by-Step Tutorial — *Chicago, Illinois, U.S.A*

ACM SIGSPATIAL SpatialAPI Workshop — *Nov. 2019*

- Talk slides and coding examples

GeoSpark and Geospatial Data Management in Apache Spark — *Las Vegas, Nevada, U.S.A*

ApacheCon 2019 North America — *Sept. 2019*

- Talk slides

ALEX: An Updatable Learned Index — *Redmond, Washington, U.S.A*

Microsoft Research — *Aug. 2019*

- Slides not available due to NDA

Designing Succinct Secondary Indexes by Exploiting Column Correlations — *Redmond, Washington, U.S.A*

Microsoft Research — *Jul. 2019*

- Presentation video

Building a Large-Scale Microscopic Road Network Traffic Simulator in Apache Spark — *Hong Kong, China*

MDM 2019 — *Jun. 2019*

- Conference presentation slides

Geospatial Data Management in Apache Spark: A Tutorial *Macau, China*

ICDE 2019 *Apr. 2019*

- Tutorial website

Spatial Data Management in Apache Spark - The GeoSpark Perspective and Beyond *Arizona, U.S.A*

NVidia *Aug. 2018*

- Talk slides

Code-generation for Fast Queries on Compressed Data *San Jose, California, U.S.A*

IBM Almaden Research Center *Aug. 2018*

- Slides not available due to NDA

Deploy Distributed Database Course Project on Vocareum *Arizona, U.S.A*

Vocareum *Feb. 2018*

- Talk slides

Geospatial Visual Analytics belongs to Database Systems *Redondo Beach, California, U.S.A*

ACM SIGSPATIAL 2017 Student Research Competition *Nov. 2017*

- Conference presentation slides

Interactive and Scalable Exploration of Geospatial Data *Arizona, U.S.A*

State Farm *Sept. 2017*

- Talk slides

Two Birds, One Stone: A Fast, yet Lightweight, Indexing Scheme for Modern Database Systems *Munich, Germany*

VLDB 2017 *Aug. 2017*

- Conference presentation slides

Indexing the Pickup and Drop-off Locations of NYC Taxi Trips in PostgreSQL – Lessons from the Road *Washington D.C., U.S.A*

SSTD 2017 *Aug. 2017*

- Conference presentation slides

Data affinity for computation on Spark and Mesos *Cupertino, California, U.S.A*

Apple *Aug. 2016*

- Slides not available due to NDA

GeoSpark: A Cluster Computing Framework for Processing Large-Scale Spatial Data *Seattle, Washington, U.S.A*

ACM SIGSPATIAL 2015 Fast Forward Session *Nov. 2015*

- Conference presentation slides