YASH GOVIND

yashgovind@gmail.com · 703-798-5998 · linkedin.com/in/ygovind

INTERESTS

Data management, data integration, entity matching, schema matching, machine learning, crowdsourcing

EDUCATION

University of Wisconsin - Madison, WI Aug 2014 - Feb 2020
Ph.D. Computer Sciences Advisor: Professor AnHai Doan
Dissertation: CloudMatcher: Toward a Cloud Service for Entity Matching
M.S. Computer Sciences GPA: 3.85 / 4.0

Pt. Ravi Shankar Shukla University, India Aug 2003 - May 2007
Bachelor of Engineering, Computer Science GPA: 8.67 / 10.0
Secured 2nd rank in the department

RESEARCH EXPERIENCE

University of Wisconsin - Madison, WI Feb 2016 - Feb 2020
Research Assistant Advisor: Professor AnHai Doan
Developed CloudMatcher, a large-scale cloud/crowd based entity matching solution. Served as the lead designer and developer. Identified research challenges and documented lessons learned in building such a solution. Wrote ˜20K Python, ˜3K Javascript, ˜3K HTML5/CSS3 and some Java MapReduce lines of code.

- Developed and evaluated an end-to-end matching solution for lay users.
- Divided the end-to-end workflow into individual services, scaled the system to handle multiple workflows and users, and evaluated the system for many use cases.
- Designed and developed a blocking rule execution solution for entity matching task that could easily scale to tables with millions of tuples.

University of Wisconsin - Madison, WI Sep 2014 - May 2015
Research Assistant Advisor: Remzi Arpaci-Dusseau
Examined data/metadata management challenges for crashes in Copy-on-Write (CoW) based file systems (B-tree/LFS).

INDUSTRIAL EXPERIENCE

Informatica LLC, Madison, WI 08/17/2020-Present
345 W Washington Ave, Suite-333, 334, Madison, WI-53703
Principal Software Engineer
Designing and developing large scale solutions for problems in the entity matching and schema matching space.

- Developed a solution to classify columns based on metadata only.
- Developed and deployed schema matching solution using Spark to production that could identify similar columns, source recommendations for joinable and unionable tables.
- Deployed the CM system developed at GreenBay Technology, Inc. and evaluated on many datasets. Developed a novel entity matching solution that scaled the existing solution by 10x in data size, and slashed deployment cost per time unit from $20K to $350. The solution was called “a breakthrough with massive savings” by the division VP and was singled out by the CTO in a monthly meeting with all engineers.

GreenBay Technologies, Inc., Madison, WI 08/12/2019-02/28/2020 (Intern) & 03/02/2020 - 08/14/2020
51 Bagley Ct., Madison, WI-53705
Principal Software Engineer Manager: AnHai Doan
Lead the design and development effort of the CM system to do EM at large scale using machine learning techniques.

- Developed the CM system using Spark and MongoDB. Wrote ˜4000 lines of code in Python.

- Evaluated the system to do identity, non-identity, and product matching.

American Family Insurance, Madison, WI 05/14/2018-08/24/2018 (FT) & 09/04/2018-05/17/2019 (PT)
6000 American Pkwy, Madison, WI 53783
Data Analytics Intern Manager: Mingju Sun
Deployed the CloudMatcher system to solve entity matching problems in the insurance domain.

- Helped six business teams prepare and clean data, and perform entity matching using CloudMatcher.
- Added cross validation, ability to cluster matching records, data management portal to view data in a project. Wrote ˜1500 lines of code in Python, Javascript, and HTML5.
- Optimized the run time of the end-to-end workflow.

Futurewei Technologies, Santa Clara, CA 05/18/2015 - 08/28/2015
2330 Central Expy, Santa Clara, CA 95050
Systems Software Intern Manager: John Wei
Extended the IceFS solution to isolate metadata in the Ext3 file system dynamically based on the size of the file system. Wrote ˜2000 lines of code in C.

- Implemented space isolation to remove read-only problems in case a particular block corrupts.
- Enhanced user-level tools (mke2fs, dumpe2fs, e2fsck) and added a monitoring tool to view metadata stats.

Syntel Ltd. 10/05/2009 - 08/01/2014
Plot B1/B2 Software Technology Park, Dehu-Alandi Road, Talawade, MH, INDIA-412114
Software Engineer/Project Lead Reference: Penni Kramer/Neeraj Sinha
In 2011, deputed at client side HUMANA INC., GREENBAY WI where I was the lead developer for the agent reporting/management application system (3 on-site/15 off-shore)

- Developed and maintained a service-oriented architecture to check agent compliance, bonus, medicare. Solved more than 50 priority-one business issues related to agent compliance.
- Led production support maintenance for old legacy system written in mainframe (COBOL, CICS, MQ series, VSAM, DB2).
- Optimized the runtime for ˜20 batch jobs, and improved SQL performance in 5 SSIS packages.

Mahindra Satyam 09/10/2007 - 09/25/2009
Hitech City Madhapur, Hyderabad - 500081, India
Software Developer
Worked as a mainframe developer and production support analyst for vendor fee reimbursement applications.

PUBLICATION

- *Deep learning for blocking in entity matching: a design space exploration*, in Proceedings of the Conference on Very Large Databases (VLDB), 2021, with many others.
- *Deep Learning for Semantic Matching: A Survey*, in Proceedings of the Journal of Computer Science and Cybernetics, 2021, with many others.
- *Magellan: Toward building ecosystems of entity matching solutions*, in Proceedings of the Communications of the ACM 63(8), 83-91), 2020, with many others.
- *Entity Matching Meets Data Science: A Progress Report from the Magellan Project*, in Proceedings of the Conference of the Special Interest Group on Management of Data (SIGMOD), 2019, Yash Govind et al.
- *Entity Matching Meets Data Science: A Progress Report from the Magellan Project*, in Proceedings of the Conference of the Special Interest Group on Management of Data (SIGMOD), 2019, Yash Govind et al.
- *CloudMatcher: A Hands-off Cloud/Crowd Service for Entity Matching (demo)*, in Proceedings of the Conference on Very Large Databases (VLDB), 2018, Yash Govind et al.
- *Toward a System Building Agenda for Data Integration (and Data Science)*, IEEE Data Engineering Bulletin, 2018, with many others.
- *Magellan: Toward Building Entity Matching Management Systems*, **SIGMOD Research Highlight, 2018**, with many others.

- *CloudMatcher: A Cloud/Crowd Service for Entity Matching*, in Proceedings of the Workshop on Big Data Analytics as a Service: Architecture, Algorithms, and Application in Health Informatics, KDD, 2017, with many others.
- *Human-in-the-Loop Challenges for Entity Matching: A Midterm Report*, in Proceedings of the Workshop on Human-in-the-Loop Data Analytics, SIGMOD, 2017, with many others.

PROFESSIONAL ACTIVITIES

- Program Committee Member & Reviewer: Reviewed over 40 academic papers. SIGMOD 2020, HILDA 2019, CODS-COMAD 2021, CODS-COMAD 2022, HILDA 2019, SYSML 2019, DI2KG 2020, VLDB 2020, EDBT 2021, SIGMOD 2022.
- Reviewer: IEEE TKDE, ACM JDIQ.

SKILLS

- Programming languages & tools: Python, C++, C, HTML, Javascript, SQL, Mainframe, Git.
- Data science tools: Pandas, Scikit-learn, Networkx, Numpy, Dask, Anaconda.
- Experience working with Spark and MapReduce.

COURSEWORK

- Big Data and Databases: Big Data Systems, Advanced Databases, Data Models and Languages
- Systems: Introduction to OS, Advanced OS, Distributed Systems
- Others: Introduction to AI/ML, Networks, Optimization, HR Strategy, Understanding Data in Psychology

ACTIVITIES

- *Chair*, ACM student chapter, UW-Madison: organized social and technical events for CS students
- *Technical volunteer*, Clean Lakes Alliance, WI: helped building an eco-yard tool to score your yard and give recommendation (won the SharkTank challenge as an intern at American Family Insurance)
- *Placement coordinator head*, SSTC: organized mock technical and non-technical interviews for CSE students
- *Member*, Graduate Studies Committee, SSTC: organized mock technical and non-technical interviews for CSE students