

YASH GOVIND

544 W Main St, Apt 107, Madison, WI 53703

ygovind@cs.wisc.edu · 703-798-5998 · linkedin.com/in/ygovind · cs.wisc.edu/~ygovind/

INTERESTS

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

EDUCATION

University of Wisconsin - Madison, WI

Aug 2014 - May 2020

Ph.D. Computer Science GPA: 3.85 / 4.0

Advisor: Professor AnHai Doan

Topic: Building hands-off cloud/crowd services for entity matching

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 - Present

Research Assistant

Advisor: Professor AnHai Doan

Building CloudMatcher, a cloud/crowd based entity matching solution. Serving as the lead designer and developer. Identifying research challenges and lessons learned. 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.

University of Wisconsin - Madison, WI

Sep 2014 - May 2015

Research Assistant

Advisor: Remzi Arpaci-Dusseau

Examining data/metadata management challenges for crashes in Copy-on-Write (CoW) based file systems (B-tree/LFS).

INDUSTRIAL EXPERIENCE

American Family Insurance, Madison, WI

May 2018 - May 2019

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.

Huawei Technologies, Santa Clara, CA

May 2015 - Aug 2015

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 - Humana Inc., Greenbay, WI

Oct 2009 - Aug 2014

Software Engineer/Project Lead

Reference: Penni Kramer/Neeraj Sinha

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.

Tech Mahindra (Client: CIGNA)

Sep 2007 - Oct 2009

Software Engineer

Worked as a mainframe developer and production support analyst for vendor fee reimbursement applications.

PUBLICATION

- *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.

SKILLS

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

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 nontechnical interviews for CSE students