William Spoth

Website: willspoth.com | Email: wmspoth@gmail.com | GitHub: willspoth

Education

University at Buffalo — Buffalo, NY
Doctor of Philosophy in Computer Science — *Expected May 2021*
Master of Science in Computer Science; GPA: 3.95/4.0 — December 2018
Bachelor of Science in Computer Science; GPA: 3.64/4.0 — May 2016
Bachelor of Arts in Psychology; GPA: 3.64/4.0 — May 2016

Technical Skills

Languages: Scala/Java, Python, SQL, C
Frameworks: Spring, Flask
Technologies: Spark, Kubernetes, GCP, Git, Docker
Tools: Maven, Jupyter Notebook, Microcontrollers

Focus Area

Databases

- Created working database from scratch in Java including optimizer, indexes, and file formatter
- Created various functionality in PostgreSQL, Oracle, and SQLite using PL/SQL
- Implemented and refactored significant systems to include functionality such as Grafana and Apache Spark as core components

Machine Learning

- Extensive use of TensorFlow, Keras, SparkML, Scikit-learn, and Smile libraries to create data models and visualizations
- Implemented a novel clustering technique to sort JSON datasets using Apache Spark and compared it to multiple Scikit-learn clustering algorithms
- Created simple multi-layered neural network example in TensorFlow and demonstrated uses and core concepts to non-domain audience

Data Science

- Presented differentially private type-system with an intuitive Jupyter Notebook Demo
- Experience creating Bayesian models and visualizing data with predictive distributions
- Experience creating unique data mining tools using Map-Reduce and production database systems like PostgreSQL and Oracle

Invited Talks — Company

Model based compression for JSON Collections — **Snowflake (2019)**
Adaptive Schemas and Elastic Query for No-SQL Data — **Oracle (2019)**
Entity extraction for heterogenous JSON collections — **Datometry (2019)**
Why data staging is the hardest part of TensorFlow — **Stark & Wayne (2018)**

Work Experience

CUBRC: Summer Internship — May 2019 - August 2019

- Implemented non performance degrading support for historical queries in PostgreSQL
- Created automatic system deployment scripts using Kubernetes and Docker
- Updated existing systems to Spring Boot microservices
- Managed data pipelines and refactored RPC calls to external systems

University at Buffalo: Database Teaching Assistant — Spring 2018

- Facilitated learning and implementation of operator pipelines, optimizer, and indexes
- Troubleshot logical and performance issues using java profiler
- Created auto-grading system using RaspberryPi's and Docker

University at Buffalo: Research Assistant — Fall 2017 - Current

- Incorporate research projects into existing production systems
- Co-review research papers
- Assist in preliminary grant research and writing

Projects

Mimir: Probabilistic data explorer

- Migrated backend machine learning models and core query executor from SQLite to Apache Spark
- Created core probabilistic schema resolver
- Implemented functional dependency based entity resolver

json-schema-scala: JSON schema validator with precision metrics

- Parses JSON schemas into easily traversable Scala objects
- Validates JSON records against schema
- Reports helpful schema metrics like tightness of attributes, object decision trees, and schema diff
- Utilizes FastParse, performant parser library

Json Explorer: JSON entity extractor

- Extracts ER-style entities an relationships from mixed JSON collections
- Discovers structure misuse through structure entropy analysis
- Utilizes Apache Spark for parallelization
- Exports schemas using json-schema specification

FBtition: Publicly verifiable petition signature protocol

- Block-chain based authentication
- PGP and web-of-trust validation
- Automatic removal of bot votes

Publications — Conference

Your notebook is not crumby enough, REPLace it — **CIDR 2020**
Michael Brachmann, William Spoth, Oliver Kennedy, Boris Glavic, Heiko Mueller, Sonia Castelo, Carlos Bautista, Juliana Freire

Data Debugging and Exploration with Vizier — **SIGMOD 2019**
Mike Brachmann, Carlos Bautista, Sonia Castelo, Su Feng, Juliana Freire , Boris Glavic, Oliver Kennedy, Heiko Mueller, Remi Rampin, William Spoth, Ying Yang

SchemaDrill: Interactive Semi-Structured Schema Design — **HILDA 2018**
William Spoth, Ting Xie, Oliver Kennedy, Ying Yang, Beda Hammerschmidt, Zhen Hua Liu, Dieter Gawlick

Adaptive Schema Databases — **CIDR 2017**
William Spoth, Bahareh Sadat Arab, Eric S. Chan, Dieter Gawlick, Adel Ghoneimy, Boris Glavic, Beda Hammerschmidt, Oliver Kennedy, Seokki Lee, Zhen Hua Liu, Xing Niu, Ying Yang

Achievements & Activities

ACM Member
ACM-SIGMOD Member
Best Graduate Teaching Assistant Award 2018
BSA Eagle Scout Excellence Scholarship