Zach Hafen-Saavedra

z.hafen.saavedra@gmail.com || zhafen.github.io || Chicago, IL ||

Data Scientist and Manager with over 12 years of experience leading efficient solution development for complex problems, including 12 years of Python experience, 4 years technical team leadership, 10 years enterprise-scale relational database experience, and 6 years of natural language processing experience. Led the migration of two teams to cloud-based solutions.

Education

Northwestern University 2020
PhD, MS, Physics and Astronomy Evanston, IL
Specialization: Astrophysical Data Analysis

University of Northern Colorado 2014
BS, Mathematical Physics Greeley, CO

Skills

Technical Skills: full-stack data science (incl. ingestion, system design, software engineering, ML/AI solution development, data engineering), healthcare analytics, NLP (incl. LLMs, vector databases, NER), relational databases, mathematical modeling and statistics
Interpersonal skills: technical leadership and management, stakeholder relations, storytelling, mentoring
Tools: Python (incl. PySpark, pytorch), Databricks (incl. asset bundles), Azure, AWS, SQL, Power BI, Streamlit, Docker, parallel computing, Git (2000+ commits/year), C/C++, and whatever gets the job done

Experience

Manager of Specialty Analytics - Research Analytics June 2024 - Present
Northwestern Medicine Chicago, IL

- Worked closely with the Enterprise Data, Cloud Ops, and Identity Access Management teams and Databricks support to bring our Azure framework from conception to production within a year of my hire.
- Managed five direct reports to deliver 10+ data products per week to Neurology, Cardiology, Urology, and Clinical Trials institutes at Northwestern Medicine (NM).
- Organized and facilitated weekly seminars, daily office hours, biweekly hack days, a three-day Python camp, system documentation, and a tutorial library, upskilling our team for a cloud environment.
- Re-envisioned from the ground up our on-premises reporting framework, designing a framework that uses Azure Databricks, Azure DevOps, Power BI, and Service Now to deliver and maintain 1300+ active reports developed by 15+ analytics developers.
- Dissected the ML Engineering team's deployment framework and adapted it to our needs, creating a robust framework that is NM's first infrastructure-as-code framework for an analytics team.
- Developed a PySpark library tailored to our team's routine tasks, enabling analysts with deep SQL experience but limited Python experience to build reports using as little or as much Python as preferred.
- Navigated organizational and technical challenges across five teams to establish a data channel between Azure and on-prem data sources, enabling seamless blending of cloud and legacy data.
- Worked closely with Identity Access Management and Cloud Security on NM's first use of B2B accounts, preparing the infrastructure for 60+ citizen developers to use our Azure resources.
- Automated the analysis of 3000+ MySQL queries and derived database usage statistics to estimate the lift to migrate to Azure, informing contractor hiring needs.
- Hired and trained analytics staff to extend the number of supported institutes from three to four.
- Directed the adoption of Agile project management over the course of six months, increasing the visibility of team efforts to over 80% of work hours.
- Addressed human resource issues spanning work authorization, stakeholder relations, and resource provisioning, with all direct reports explicitly vouching that they feel supported.
- Fielded 10+ inquiries per week from policy to data engineering to crisis management, ensuring smooth operation of our team while protecting patient privacy.

- Organized the NM NLP working group and held quarterly meetings of 15+ NLP experts from across the organization, addressing gaps in NLP knowledge and establishing a community of practice.
- Collaborated with the ML/AI and Cloud Ops teams to establish secure resources for LLM processing of sensitive data, enabling the use of LLMs in healthcare data analysis.
- Mentored team members in the development of an LLM-based classification model, automating the assessment of 2000+ faculty reviews and saving thousands of hours of manual work.
- Instructed data engineers in re-tooling the team's medical text de-identification approach, enabling the efficient parallel processing of 1 million+ healthcare notes with improved accuracy.
- Guided analysts in the development of new named entity recognition (NER) models that draw on labeling software and LLMs to identify 100s of patients missed by human classification.

Far Horizons Data Scientist September 2023 - June 2024
Adler Planetarium Chicago, IL

- Utilized tools including CodeBuild, Docker, ECR, and PostgreSQL to deploy an end-to-end data analytics pipeline on AWS, enabling Adler staff to ingest and process data from AWS S3 with the push of a button.
- Developed an automated ELT and image-registration pipeline using Python and shell scripting, dramatically increasing the georeferencing speed from 4 manual images/hour to 5000 images/hour.
- Developed documentation, an intuitive user interface, a suite of 40+ code tests, and a stable, containerized computing environment, preparing for 3 years of minimal-maintenance use by stakeholders.
- Generated a high-resolution aerial map of 10 km around Indianapolis and released it to partner organizations, enabling predictions including light-based income estimation.
- Directed the adoption of Agile project management, attaining 70%+ community-driven development.
- As a museum resident scientist, educated and collaborated with non-technical educators to deliver life-changing deep-impact programs for 20+ high-school students and 4 interns.

Group Lead and McCue Prize Postdoctoral Fellow July 2020 - June 2023
University of California–Irvine, Department of Physics and Astronomy Irvine, CA

- Advised and mentored team members for the Dean of Physical Science's research group, leading to the advancement of 7 astrophysics and 1 computational linguistics PhD students.
- Developed a Python-frontend, C++-backend code to perform NLP embeddings of scientific text, and presented at AI4Science on a clustering-related metric correlated with a 150%+ increase in citations.
- Automated data retrieval from NASA APIs, extracting metadata for more than a million papers.
- Orchestrated a mock data challenge spanning nine international institutions, quantifying where statistical models attained 90%+ accuracy, informing technical stakeholders' modeling decisions.
- Led and organized a workshop of twenty key community leaders, fostering cross-specialty dialogue to discern high-value targets.
- Built an end-to-end workflow linking three disparate sources of structured and semi-structured data and predicting focus areas for adjacent disciplines.

National Science Foundation Graduate Fellow in K-12 Education June 2014 - July 2020
Northwestern University, Department of Physics and Astronomy Evanston, IL

- Used remote resources to apply event-tracking to 20+ TB of relational data, and isolated predictive parameters stakeholders could use to predict future behavior with 99%+ certainty.
- Performed time-series decision-tree classification to predict the cosmic origins of the atoms we are made of, delivering testable hypotheses to guide collaborators.
- Responsibly used remote resources to operate 100,000-CPU-hour simulations, generating data used by 30+ stakeholders to increase statistical power and realism.
- Crafted award-winning visualizations displayed throughout Chicago libraries and museums, advertising the beauty of science to a wide audience.
- Partnered with local schools to pioneer a high-school data-science program, reaching over 100 students.
- Collaborated with 100+ researchers, leading to 36 published papers, 7 as lead author.