Aaron Batilo

https://www.aaronbatilo.dev

(720) 207-7973

AaronBatilo@gmail.com

EXPERIENCE

- **Microsoft AI**
  Member of Technical Staff *March 2024 - Present*
  - Leading technical transition for subset of LLM pretraining infrastructure and assets from Inflection AI
- **Inflection AI**
  Member of Technical Staff *January 2024 - March 2024*
  - Kubernetes cluster administration and debugging for LLM pretraining and inference use cases
- **Cohere AI (Remote)**
  Member of Technical Staff *May 2023 - January 2024*
  - Contributed improvements for managing GPU compute for LLM inference in capacity constrainted environments
  - Added multi-node supercomputing scale GPU support for LLM training pipeline code
- **Color Health (Remote)**
  Staff Software Engineer - Tools and Infrastructure *January 2021 - May 2023*
  - Lead technical implementation for shifting to containers off of pets / sticky application infrastructure
  - Integrated tracing and profiling across Python applications leading to to reduced latency by upwards of 80%
  - Re-paved manually provisioned infrastructure with GitOps self service based infrastructure as code
- **iStreamPlanet (Remote)**
  Technical Lead / Senior Software Engineer - Product team, Tools and Infrastructure team *August 2019 - January 2021*
  - Lowered AWS cost by 6 digits monthly by improving applications to use spot instances with zero downtime
  - Developed internal GitOps Kubernetes PaaS for self service and auditable infrastructure to over 70 developers
  - Owned performance testing and deployment of Golang services to support $>$10k QPS and $>$1Tbps of traffic
- **Conga**
  Technical Lead / Senior Software Engineer - Machine Learning / Platform Engineering *January 2018 - August 2019*
  - Led full-stack development, and launch of new product for ML inferences on contracts (Conga AI Analyze)
  - Administered TeamCity server and championed company wide adoption of CI / CD best practices
- **FullContact Inc**
  Software Engineer II - Data Platform *July 2017 - January 2018*
  - Designed, trained, and deployed a text classifier for determining the organization and department for a job title
  - Implemented Schema.org support in data scraping pipeline, resulting in 14% lift in data extracted from websites
- **Josh.ai**
  Software Engineer - Full-stack / Android *June 2015 - July 2017*
  - Created Python / Flask full stack solution for interfacing with home automation system, deployed to Heroku
  - Created native Android app for real time communication with home automation system over a WebSocket
- **Northrop Grumman Corporation**
  DevOps Engineer (TS SCI with CI Poly / SSBI) *June 2014 - June 2015*
  - Wrote Python scripts for system monitoring across 40+ virtualized hosts on mission critical and secured networks
  - Automated silent installation of COTS products for live deployments saving hours of manual labor
- **University of Colorado, Colorado Springs**
  Student Web Developer / Research Assistant / Teaching Assistant *January 2013 - December 2015*
  - Wrote perl / MSSQL / HTML5 / CSS3 responsive site to render class schedules
- **Northrop Grumman Corporation**
  Software Engineer Intern *Summers of (2011 - 2013)*
  - Various summer internship projects working on unclassified cybersecurity related tools

PROJECTS

- **https://www.aaronbatilo.dev**: My website has a full list of non-work projects. Some of these projects have:
  - Been featured on the Wall Street Journal, Colorado Sun, Denver 7, Denver 9News, more
  - Used by Microsoft, Docker, GitHub, MGM Resorts, Tidal Music