Josh Arnold

Knoxville, TN
-Email me on Indeed: http://www.indeed.com/r/Josh-Arnold/173c9bd05244875e

Work Experience

Machine Learning Engineer

Oak Ridge National Laboratory - Oak Ridge, TN
December 2019 to September 2021

Served as co-PI for Data Science for ORNL partnership with Centers for Medicare & Medicaid Services (CMS) Center for Program Integrity.

• Architected on-prem data lake for CMS data using PySpark, DeltaLake, Airflow.
• Applied machine learning algorithms such as XGBoost, Transformers, et al. to provide novel insights regarding waste, fraud, and abuse.
• Communicated findings via biweekly presentations.
• Mentored junior data scientists.

Served as Data Engineer for the REACHVET project, a human outreach program within the Department of Veterans Affairs to identify veterans at risk of suicide and opioid abuse.

• Designed and implemented HDFS-SQLServer-PySpark hybrid solution for speeding up daily ETL pipelines by 25x over existing approach.
• Designed WebUIs for interactive data visualization.
• Proposed entity-centric data model for improving feature and cohort selection.

Research Scientist I - Data Engineer

Oak Ridge National Laboratory - Oak Ridge, TN
December 2017 to December 2019

Served as Data Engineer for the Department of Veterans Affairs Open Source Lab for the evaluation of state-of-the-art open-source technologies on the basis of security, usability, and performance within the context of VA research and production workflows.

• Provisioned, deployed, tested, and benchmarked a myriad of open source storage solutions, including Solr, Elasticsearch, Hadoop, HBase, HA PostgreSQL, MongoDB, Kafka, et al. using Ansible and Helm.
• Deployed and maintained Kubernetes cluster using `kubespray-ansible` on top of ORNL's on-prem OpenStack cloud for rapid development and testing.
• Wrote os-fiddle, a thin Scala wrapper for Apache jclouds for programmatic provisioning with OpenStack.

Application Developer

Vanderbilt University - Nashville, TN
August 2016 to November 2017

Served as Big Data and Machine Learning expert at the Advanced Computing Center for Research and Education.

• Spear-headed ML pipeline implementation of sentiment analysis from earnings calls data in collaboration researchers at the Owen Graduate School of Management.
• Created "context_bot", a Python bot for scraping content from ACCRE's website and responding to user tickets with website content recommendations.
• Designed ACCRE user database using PostgreSQL and SQLAlchemy.
• Developed scripts for provisioning Spark jobs on the ACCRE supercomputer.

Materials Engineer

National Institute of Standards and Technology - Gaithersburg, MD
September 2014 to December 2015

Served as Research Scientist for the Inorganic Materials Group (IMG)

• Created C++/Python machine learning pipelines for automated classification of hyper-dimensional image features, allowing for reliable classification of cement phases.
• Added mean-field aqueous solution equilibrium approximation to cellular automata simulations of cement microstructure evolution using C++/MPI.
• Developed Monte Carlo model of nanoscale cement morphology using C++/CUDA GPU parallelization, allowing for the estimation of critical microscale model parameters.
• Coordinated IMG’s participation in the Materials Data Curator System, part of the Materials Genome Initiative.

Education

Doctorate in Environmental Engineering

Vanderbilt University - Nashville, TN
August 2009 to February 2014

Master's degree in Environmental Engineering

Vanderbilt University - Nashville, TN
August 2006 to August 2009

Bachelor's degree in Civil Engineering

Vanderbilt University - Nashville, TN
August 2002 to May 2006

Skills

- Git (6 years)
- Rust (1 year)
- Spark (5 years)
- Python (6 years)
- GitHub (6 years)
- Hadoop (4 years)
- AWS (Less than 1 year)
- Microsoft SQL Server (2 years)
- Linux (6 years)

- SQL (5 years)
- Elasticsearch (Less than 1 year)
- Solr (Less than 1 year)
- Software development
- Scala (4 years)
- Machine learning (10+ years)
- Big data (6 years)
- CI/CD (3 years)
- Kubernetes (2 years)
- Natural language processing (5 years)
- Docker (5 years)
- C++ (2 years)
- Deep learning (5 years)
- Torch (5 years)
- NumPy (6 years)
- Pandas (6 years)
- Lucene (Less than 1 year)
- Ansible (4 years)
- Data visualization (10+ years)
- HDFS (5 years)
- MATLAB (10+ years)
- ETL
- Data modeling
- C/C++ (2 years)
- SLURM (1 year)
- CUDA (6 years)
- ArcGIS (1 year)
- Data Warehouse (5 years)
- Technical writing
- Kafka (1 year)

Certifications and Licenses

The Data Incubator - 77477

March 2016 to Present

Completed a vigorous eight-week bootcamp for leveraging SOTA data analytics tools, including MySQL, MapReduce, Spark, Pandas, Scikit-learn, et al.

Assessments

Analyzing data — Expert
June 2021

Interpreting and producing graphs, identifying trends, and drawing justifiable conclusions from data
Full results: Expert

Advanced mechanical knowledge — Expert
September 2021

Understanding and applying mechanical concepts and processes
Full results: Expert

Search engine optimization — Proficient
September 2021

Interpreting online website performance metrics and understanding search engine optimization tactics
Full results: Proficient

Basic computer skills — Highly Proficient
September 2021

Performing basic computer operations and troubleshooting common problems
Full results: Highly Proficient

Numerical reasoning skills — Expert
September 2021

Quickly and accurately performing basic mathematical operations, recognizing numerical sequences, and interpreting graphs
Full results: Expert

Groups

Society for Industrial and Applied Mathematics
September 2006 to Present