Nate Dire

P: 425.405.5263
E: nate.dire@gmail.com
W: http://www.linkedin.com/in/ndire

Summary

- Flexible, analytical, and proven data science engineer with strong systems and statistics background.
- Pre-series A experience at two successful Seattle startups.
- Ability to independently research, apply, and productize complex algorithms.
- Experience applying machine learning and NLP end-to-end in a B2B SaaS product.
- Key contributor to successful high-performance POSIX/ACID clustered file system that scaled to 144 nodes and 15 PB.
- Advanced knowledge of file systems, distributed system, and *nix server programming.
- Production contributions with AWS, Chef, Clojure, Java, MongoDB, Postgres, Python, Ruby, Solr, and Spark.

Work History

Highspot, Seattle, WA
Principal Software Engineer May 2013 - June 2018

Data Science contributions:

- Evaluated LDA topic model on office documents in both Spark and AWS Comprehend.
- Implemented matrix factorization recommendations using Spark mllib ALS with time-decayed item affinity. Added vector support in Solr to compute preference scores based on user and item vectors. Plumbed crude A/B testing model to evaluate lift.
- Built out Relevance KPI Dashboards in AWS QuickSight using custom Postgres views.
- Provisioned and configured production snapshot environment for optimizing MRR via query log replay experiments.
- Implemented custom suggested query scheme with fuzzy matching and personalization. Seeded with entity title N-grams. Used for 20% of queries.
- Improved media file relevance with content annotation. Implemented hinted transcription for audio and video using Google Cloud Speech API; 10-color text search and facets using GraphicsMagick quantization; image topic and OCR search using Google Cloud Vision API.
- Researched and resolved search pathologies with spell correction, multi-term synonyms, exact match boosting, and query sanitization. Reduced re-query rate by more than 10%.
- Worked with Data Scientist to productize slide similarity feature based on TF-IDF and cosine similarity of text and images. Independently productized full document similarity with hierarchical agglomerative max-link clustering for summarizing engagement analytics by document cluster.
- Planned multi-language search roadmap. Added basic tokenization and normalization support with testing by native speakers. Implemented language identification by combining CLD library and Google Cloud Translation API.
- Built session analysis pipeline from clickstream logging to DAU queries. Joined client-side and server-side logging. Decorated links for CTR and MRR analysis.
- Built business intelligence pipeline from MongoDB to RDS Postgres. Currently used for business metrics, billing, and to deliver custom analytics to customers.

Infrastructure contributions:

- Eliminated downtime deployments by researching and proposing crude rolling deployment scheme. Established git branching and tagging model for releases, along with corresponding application version recorded in clickstream log.

- Measured and improved email deliverability using Sendgrid and 250ok. Implemented open and click tracking via webhook updates. Prototyped recipient tracking via external server. Designed DKIM/SPF scheme for sending with customer email in from header.
- Redesigned Solr commit model along with background work queue to be more NRT-friendly, reducing average latency by 50% and eliminating a class of outages. Implemented atomic updates for 100x faster updates to numeric fields.
- Planned application concurrency roadmap; refactored data layer for tombstoning, idempotency, and optimistic updates to MongoDB and Solr, including HTTP 409 and 410 responses.
- Deployed 8-node Solr cluster behind private ELB using VPC ClassicLink. Deployed attendant Zookeeper cluster to replace single instance. Upgraded Solr approximately yearly from 4.2 to 6.5. Wrote tasks to validate index, balance replicas, and run daily rolling restart.
- Expanded MongoDB server to replica set. Migrated deployment to CloudManager. Set up backup scheme with tasks for developer restores of production. Planned upgrades from 2.4.x to 3.2.x, including Ruby, Java, and Clojure library updates. Proposed and implemented functional partitioning scheme with second replica set.

EMC, Isilon Division (formerly Isilon Systems), Seattle, WA

Consultant Software Engineer April 2011 - March 2013

- Presented clustered file system architecture at EMC World 2011 and 2012.
- Wrote 200+ page product internal architecture document in LaTeX.
- Led large cluster scalability roadmap.
- Designed and implemented new BSD vnode operation locking to enable file system filter API.
- Member of product architecture team; met with customers and consulted on product deployment configurations.

Lead Software Engineer August 2008 - April 2011

- Independently formed multi-department group to create field product configuration tool used by all sales engineers.
- Led design and implementation of multithreaded, distributed job engine which runs critical file system jobs on clusters up to 144 nodes.
- Led design and implementation of file system SSD strategy.
- Planned and executed roadmaps for tiering and integrity features.

Software Engineer March 2003 - August 2008

- Researched and evaluated erasure codes. Implemented Reed-Solomon codes for 4-failure protection, including hand-optimized x86 assembly with SSE2 instructions.
- Designed and implemented mark-and-sweep collection for orphaned file system structures.
- Independently researched reliability analysis techniques. Implemented Monte Carlo simulator in C++.
- Coordinated Mean-Time-To-Data-Loss (MTTDL) estimation with Marketing, QA, and Operations.

Education

Whitman College
B.A. Mathematics, *cum laude*, May 1998

- Phi Beta Kappa
- Presented inverse eigenvalue numerical analysis research at 1998 AMS/MAA Annual Meeting

University of Washington
M.S. Computer Science, June 2005
Certificate in Natural Language Technology, March 2016

Coursera
Recommenders Specialization, April 2018