AMIR YOUSSEFI

Single page highlights followed by resume

HIGHLIGHTS

Principal engineer and technical leader with expertise in Data, Software and Machine Learning Engineering:

- **Eng:** Data Systems (Spark/Druid/Hadoop/Kafka), App Dev (Java/Python), DB/Data Warehousing, Cloud, DevOps
- **ML:** ML App Dev (Conversational LLM+Speech), Training Data prep/gen, open model tuning/training, MLOps

ROLES

- Principal Software Eng, **Hippocratic AI (LLM+Speech Startup)**
- Sr Architect & Chief Data Engineer, Data Technology department, **PayPal**
- Engineering Manager & Sr Software Eng, the first (2007-10) **Hadoop** team**, Yahoo**
- Principal Software Eng, (real time) Query and Data Storage team, **Conviva**
- CTO & Data Scientist, **OnTopic (NLP/ML startup)**
- Hadoop Architect & Applications Lead (IC), iCloud and IS&T teams, **Apple**
- Software/BI Eng II, Data Warehouse team, **Amazon**

SKILLS

- **Data Systems:** Druid, Spark, Hadoop, Kafka, Hive, Elastic, Databases (SQL/KV/Vector), Data Warehousing, ETL
- **ML Eng:** ML data lifecycle, LLM/RAG/Prompt Eng, classic NLP/ML algorithms, MLOps, some VAD/ASR
- **Engineering Leadership:** Led architecture/dev/R&D/POCs. Developed organizational eng. practices/processes
- **Cloud/DevOps:** AWS (SageMaker, RDS), GCP (GKE,), DevOps (Kubernetes, Helm, TF), Linux (power user)
- **Programming:** Java, Python, Scala (limited), various ML/DL/Web Services frameworks, libraries and tools

SELECT PROJECTS

- PayPal: led dev of Data Privacy Platform with ML; company's 1st chatbots and Kafka/Spark Streaming; GPU evals
- Yahoo's Hadoop team: Managed Utilization/Performance team and built DevOps tools for 38,000 node Hadoop clusters. Personally built 7x performant solution and won competition benchmark vs Oracle/MPP/columnar databases.
- Conviva: made large customer facing realtime Druid clusters 3x faster in query and 10x more reliable (added a 9)
- Hippocratic AI: custom ASR fine-tuning; VAD/Speaker ID; ML inference infra w/ SageMaker, LLM/Multimodal apps
- OnTopic: built NLP/Semantic Search systems with end-to-end information retrieval, text extraction, search ranking

EDUCATION

- MS (& incomplete PhD) in Computer Science (research on ML and Data Systems), Rensselaer (RPI), New York
- BS in Electrical Engineering, Sharif University of Technology

PUBLICATIONS/PRESENTATIONS

- Multiple publications and conference presentations on systems (Hadoop, Druid) and Machine Learning, 2003-24
- Finalist/category winner in 20+ hackathons & programming competitions, 2007-20

AMIR YOUSSEFI

Linkedin: http://linkedin.com/in/youssefi e-mail: amir.youssefi@gmail.com Phone: (408)372-6477

PROFESSIONAL EXPERIENCE

Principal Software Eng, Hippocratic AI, Palo Alto, 2023-2024
Developed conversational (LLM+Speech) Apps. Curated & generated medical terminology datasets/audio for custom model training. Led VAD/Medical ASR Fine Tuning/Speaker ID projects. MLOps/infra (SageMaker & EC2 GPU), led Nvidia(NeMo) team collaboration on medical ASR & deploying on-prem, Multimodal (initial Audio+LLM R&D)

Principal Software Eng, Conviva, Foster City, 2020-2023
Performance optimization/dev/ops of (near) real-time Streaming Analytics processing trillions of events per day:
- Led optimization of customer facing data systems with Druid, Spark, Hadoop, Kafka, ClickHouse, MySQL and Redis running on hybrid cloud (on-premise data-center plus GCP & AWS clouds)
- Optimized performance of the system achieving ~3 times faster queries. Enhanced Reliability of the system by 10x (a 9) and improved OPS Runbooks/observability. Upgraded infrastructure to Kubernetes/GKE, Helm, Docker, Terraform

Chief Data Engineer & Sr. Architect, Data Technology Department, PayPal, San Jose, 2015-2019
Led Data Infrastructure and a few ML projects with architecture design, R&D, hands-on prototyping/dev and scaling in production. Team provides horizontal data platform services to the company. Projects I led as architect/ML Eng:
- **Chatbots:** Led dev/POC of the first company chatbots for Customer Service/QA/MetaData Knowledge Base with Deep Learning Language Models (LSTM), Vector/Similarity Search(Elastic), Topic Modeling, classification (fastText)
- **Privacy ML Platform:** Led architecture design and implementation of ML Pipeline and made critical government GDPR compliance deadline. Developed Text Classifiers, Computer Vision(Face Detection), Image OCR/Transcription (using deep learning Faster R-CNN and TensorFlow). Implemented full MLOps lifecycle of data-set preparation, labeling, augmentation, training, testing, validation, scoring, production deployment/monitoring.
- **Data Systems/Apps:** Hands on development of various production with Hadoop, Spark/Spark Streaming(Scala), Kafka, Hive, HBase, ML Libraries (scikit-learn, TensorFlow), Data Warehousing(TD/Oracle)
- **Kafka:** Developed company's first data streaming platform for cross Data Center deployment of ML Models (~2016)
- **Speech Recognition/ASR:** Led speech-to-text POC for PayPal Call Centers comparing open-source vs vendors (Nuance, Google ML) on metrics of WER, Classifier (PR/F1-Score/AUC-ROC), latency, throughput, price
- **GPU:** Evaluated Deep Learning/Analytics performance of GPUs (V100, DGX-V100, P100) on HW/SW/Network. Conducted POC for real-time GPU Analytics (Kinetica, OmniSci/MapD/Heavy) as build/buy scenarios
- **Text scanning:** Co-developed a system for government financial regulatory compliance using NLP and Hadoop
- **ML Workshop:** Taught an internal Machine Learning tutorial

Principal Architect, Independent Contractor, ATAP, Google, Q4 2014-Q1 2015
Member of Project Abacus w/ researchers from 16 universities building data and training Multi Modal ML Algorithms (Sensors, Vision, Voice, Text) for password-less authentication on mobile phones. We built an air-gapped Hadoop cluster(HW&SW), ML infrastructure and algorithms. I/O video: https://www.youtube.com/watch?v=lGrRYnqHegc

CTO & founder, OnTopic, Santa Clara, CA, 2013-2014
Designed & implemented end-to-end NLP/Retrieval/Classification/Search/Rec Sys with Mobile and Voice interfaces
- Machine Learning: Topic Modeling (custom LDAs), Search/Ranking, Recommender System, classification
- Data Processing: Elasticsearch, MySQL, Java (crawler, data pipelines, Text-to-Speech integration), Spark, Hadoop
- Backend: Java and Scala (Spring and Play Frameworks, StanfordNLP), Python (NLTK), AWS & on-premise, Linux

Hadoop Architect & Applications Lead, Independent Contractor, Apple, Cupertino, 2010-'12, 2012-'13
- Developed distributed data processing systems solutions for IS&T and iCloud teams with Hadoop, Hive, Pig, HBase, Oozie and ZooKeeper. Also POCs with Spark and neo4j Graph Database. Applications: Enterprise Data Warehouse (with Hive+ORC & Table Format for delta/compaction), iCloud Fraud Detection/Analytics with Hadoop/HBase

Engineering Manager & Sr. SW Eng, Hadoop Team, Yahoo, Sunnyvale, 2006-2010
- Early member of Hadoop Team (2007-'10). Led Hadoop DevOps Team as Eng. Manager and Apache Pig Solutions.
- Eng Mgr for GMS (Grid Management & Monitoring System) for Hadoop Infrastructure (~38,000 servers, 24 clusters

and 12 datacenters). Anomaly detection and failure prediction using Machine Learning on collected time-series data.

- Sole Solutions Architect for Pig Team in 2007 for all Yahoo's ~350 Pig Developers. Taught Pig/Hadoop classes.
- Represented Hadoop team in a million dollar competition vs Oracle vs Greenplum MPP DB vs a Columnar DB. Developed and performance tuned solutions for a TPC-H like benchmark with 7x better performance with dozens of improvements e.g. custom InputFormat/LineReader/Partitioner Hash Algorithm/Shards/LZO Compression, predicate push-down, locality design & advanced multi parameter optimizations
- Participated in development of TFile, an open-source File Format partially adopted by HBase HFile & Hive ORC
- Co-authored Ad Targeting/Marketing Datawarehouse of Yahoo with Oracle (Dimensional Modeling) & ETL

Software/BI Eng II, Amazon Inc., Seattle WA, 2006

Data Warehousing using Dimensional Modeling, SQL and ETL on Oracle (a world top 10 Data Warehouse in 2006)

Sr. Software Eng, Data Warehouse Developer, Plateau Systems, Arlington VA, 2004-2005

- Java Enterprise (Spring/Struts) development, Oracle Data Warehousing (Dimensional Modeling)/ETL, ORM

CTO & Lead Software Engineer, Pars Expo Co., outsourced from San Jose to Tehran, 2001-2002

Built and managed dev team in the 35 person startup. Developed a large e-commerce with Java J2EE, Oracle, JS.

Engineering Manager & Software Eng, R&D Computing Center, Sharif University, Tehran, 2000-2001

Developed registration system using Java J2EE and Oracle, serving 10,000+ university students

TECHNOLOGY

Data Systems and Cloud Computing

- Distributed Data Processing: Spark, Druid, Kafka, Hive, Hadoop, HBase, Elastic, Pig, ClickHouse, ZooKeeper
- Data Warehousing: Dimensional Modeling, ETL/ELT, Oracle Analytics SQL Development and Performance Tuning
- Databases: Oracle, Elastic, MySQL, PostgreSQL (limited), Teradata (limited), Redis, MongoDB, neo4j(limited)
- Microservices/OPS: Kubernetes (w/ Helm, Terraform), Airflow, Monitoring/Observability Sys, Linux(power user)
- Cloud: GCP(GCS, CE, GKE+TF, Big Query, Pub/Sub, Dataflow, ML API), AWS(SageMaker, EC2,S3,RDS)

Software Engineering

- Programming Languages: Java, Python, SQL, Scala (limited), C++/C (limited), Shell Scripting (limited)
- Middle-Tier/SOA Frameworks: Play Framework(Java/Scala), Spring Framework, Hibernate/Alembic ORMs
- Engineering Leadership: Led Architecture/Dev/R&D/POCs. Developed organizational Eng. Practices/Processes

Machine Learning

- Open model tuning/training, ML data prep-to-prod, LLM/RAG/Prompt Eng, classic ML/NLP, MLOps (GPUs)
- Data centric ML Full cycle from collection, training, production, inference, model monitoring, visualization
- Classic ML algorithms: NLP and text (classification, Topic Modeling), Text/Vector Search, RecSys (limited)
- Tools (limited): Pandas, NumPy, scikit-learn, PyTorch, NLTK, LangChain, LlamaIndex, PyAnnote

EDUCATION

MS (& incomplete PhD), Computer Science, Rensselaer Polytechnic Institute (RPI), New York, 2002-03

M.S. Thesis & PhD Research: Machine Learning on the web data

BS, Electrical Engineering, Sharif University of Technology (SUT), Tehran, 1995-2000

PUBLICATIONS/PRESENTATIONS

- Finalist/category winner in 20+ hackathon & programming competitions, 2007-2020
- Mukherjee et al., Polaris: A Safety-focused LLM Constellation Architecture for Healthcare, 2024
- Youssefi, Ranjan, Apache Druid Optimizations for Scaling Realtime Customer Facing Analytics, Druid Summit, 2021
- Youssefi, Malek-Madani, Getting the Most from Your Hadoop Nodes, Storage Developer Conference, San Jose, 2012
- Youssefi, Qi, Hadoop Performance Tuning at Scale: How to win a benchmark competition!, Yahoo Tech Report, 2007
- Youssefi, Duke, Zaki, Visual Web Mining, World Wide Web Conference (WWW), New York, 2004
- Youssefi, Duke, Zaki, Glinert, Toward Visual Web Mining, IEEE Int'l Conf. on Data Mining (ICDM), 2003