Surya Dutta

708-712-9625 · surya@suryadutta.com · linkedin.com/in/suryadutta · github.com/suryadutta

Education

University of California, Berkeley Berkeley, CA
Master of Information and Data Science (MIDS) | GPA 3.88/4.00 *Jan 2019 – Apr 2021*

Yale University New Haven, CT
Bachelor of Science, Physics (Intensive) | GPA 3.74/4.00 *Aug 2014 – May 2018*

Experience

McMaster-Carr Supply Company Elmhurst, IL

Senior Data Engineer, Digital Analytics *Feb 2021 – Present*

- Architected, developed, and productionized data warehouse platform that combines 5 years of customer experience observations and transactions from our website with internal customer and product data. Used to report aggregate North Star metrics to company leaders and to empower project teams to self-serve data exploration and insight
- Guided conversations with 3 VPs and 8 project teams to translate stakeholder needs into pipelines & data models
- Leveraged data build tool (DBT) to prototype and document transformed tables and views that make complex datasets more accessible while keeping the data platform performant and extensible to future additions & changes
- Managed and mentored remote team of 7 new engineers (with little/no previous technical experience) in building and orchestrating robust pipeline jobs to extract, load, and transform data in the warehouse in a timely manner
- Ensured high-availability, veracity, and completeness of data through appropriate infrastructure provisioning, extensive data monitoring, and intelligent alerting through automated statistical tests

Data Engineer, Digital Analytics *Feb 2020 – Feb 2021*

- Designed and implemented ETL data pipelines to capture customer interactions and content exposure data on mcmaster.com and tie back to specific internal publishing data within a Neo4j graph data lake
- Developed suite of post-processing jobs to extract standardized OKRs from diverse customer search patterns. Built a REST API to dynamically retrieve these metrics, filtering by keywords, datetime range, and metadata

Software Engineer, Performance *Aug 2018 – Feb 2020*

- Decreased load time of mcmaster.com by 31% by implementing predictive algorithms to pre-fetch resources
- Orchestrated ongoing A/B tests to experimentally verify success of performance optimizations made by the team

Projects

Cognition Tracker | *Tensorflow, TFX, Python, AWS, S3, Docker, React Native* | Website

- Developed mobile application that measures cognition and progression of dementia through spontaneous speech
- Achieved world-leading accuracy with audio ensemble ML model (LSTM + speech-to-text BERT embeddings)
- Scaled backend architecture in AWS using Docker and Tensorflow Serving to return instantaneous score to app

Predicting Airline Delays with Weather Data | *Spark, Databricks, Python* | GitHub Repository

- Utilized BTS dataset of 31.75 million domestic US flights over 3 years, combined with an NOAA weather dataset of 631 million records, to predict the probability of a substantial (15+ minute) flight delay 2 hours beforehand
- Engineered features including optimized windowing joins and distributed PageRank of airline travel between cities
- Achieved 0.74 AUROC score using Spark MLlib gradient-boosted tree model (best-in-world: 0.9, random: 0.5)

Mitigating Gender and Racial Bias in BERT | *Tensorflow, Python, Docker* | GitHub Repository | Paper

- Created new metric to measure multiclass bias of contextualized embedding models like BERT in NER tasks
- Implemented adversarial debiasing deep learning model to significantly reduce gender and racial bias

Technical Skills

Proficient:	SQL, Python, C#/.NET, Docker, Git, MS SQL Server (data warehousing), Neo4j/Cypher (data lake), data build tool (DBT), NumPy, pandas, scikit-learn, pandas, Tensorflow, spaCy, matplotlib
Experienced:	Spark, Kafka, AWS, Linux, PostgreSQL, Redis, MongoDB, React/React Native, TFX, Jenkins
Familiar:	Presto/Trino, Apache Airflow, Google BigQuery, Google Cloud, FastAPI