Sungwook Yoon, sungwook.yoon@gmail.com

Positions

present	Principal Engineer	Comcast
2016 - 2018	Principal Data Scientist	MapR
2014 - 2015	Data Scientist	MapR
2013 - 2014	Sr. Data Scientist	Vectra Networks
2012 - 2013	Architect	Seven Networks
2012 - 2012	Data Scientist	Identified
2008 - 2012	Research Scientist	PARC (Palo Alto Research Center)
2006 - 2008	Assistant Research Professor	ASU (Arizona State University

Data Engineering Experiences

Databricks	Workflows, Unity Table
AWS	EMR, Lambda, EC2 and many other services
Data Bases	Postgres, MSSQL
ML Systems	H2O, Spark MLLib, R, TensorFlow, Python ML
Hadoop Systems	MapR, MapReduce, Hive, Spark, HBase, Kafka, OpenTSDB

Language Experiences

Experienced	C, Scala, SQL, Python, R, Lisp, Scheme
Medium Exposure	Bash, C++, Java, Ruby
Novice	Perl, Prolog

Cloud System Deliveries

Move AWS to Databricks	Moved jobs in AWS into Databricks
Move On Prem to AWS	Moved on prem spark/Hadoop jobs to AWS services
AWS Data Processing	Built and delivered numerous data processing and data quality

Spark Specific System Deliveries

Advertisement	Implemented complex business logic on OnPrem, AWS and Databricks
Real time app log analysis	Spark Core + Spark MLLib + Spark Stream

Awards

2011	Best Paper Award Runner Up	Journal of Artificial Intelligence
2011	Best Paper Award Runner Up	International Conference on Automated Planning and Scheduling
2009	Best Machine Learning	International Learning and Planning Competition
2006	(Unofficial) Winner	International Probabilistic Planning Competition
2004	Winner	International Probabilistic Planning Competition

Education

Ph.D.	Computer Engineering	Purdue University	AI, Machine Learning, Planning
M.E.	Electrical Engineering	Seoul National University	Video Compression, ATM
B.E.	Electrical Engineering	Seoul National University	Control and Instrumentation

Machine Learning and Data Science Projects Summary

Data Science Deployment	Comcast	Deployed DS Projects into Production
Data Science Engagements	MapR	Various Customer Engagements / Teachings
Malware and Fraud Detection	Vectra Networks	Production Code Dev
Mobile App Data Analysis	Seven Networks	BI and Algorithm Dev
Career Analysis	Identified	Data Analysis
GILA (Machine Learning)	ASU	Machine Learning Dev

Publications: 20+ Publications at only top journals, **JAIR**, **JMLR** and top conferences like AAAI, **NIPS**, **ICML**, **UAI**, ICAPS, IJCAI.

Data Experiences

Advertising	Impression and Reach Data Product and Analysis Audience Data Product and Analysis Various troubleshooting on data and remediations Data pipelines of various forms Data Quality
Tech Stacks	EMR/Lambda/Redshift/Athena/Databricks
Security	Anomaly detection on Pcap data Real time detection and Alert Coded the algorithm that works directly on packets
Tech Stacks	Wireshark/C++
Android	Phone app data analysis Analysis on how apps signaling pattern affects the power Visualization
Tech Stacks	R/Postgres
Social Network	User location analysis Associated users based on their connection and geo location Visualization
Tech Stacks	Ruby/Postgres/R

Selected Commercial/Government Development Experiences

Real-Time Network Anomaly Detection System @ Multiple Customers	We worked with a few fortune 500 companies for their IT security data analysis projects for several months. We used Spark to enrich the streaming data and used ES to Spark to ingest into Elasticsearch Visualization We developed machine learning system using Spark MLLib for the baseline analysis and traffic pattern We also used Spark GraphX to develop consistent network topology of the customer network. PageRank algorithm and Connected Component analysis in GraphX help the customer easily find significant lateral data movement from the network data without manual effort and automatically We used Scala for Spark development
Data Ingestion Into MapR @ Multiple Customers	We performed several data ingestion services for multiple customers. Mostly from existing databases or streaming log text data, We ingested into either MapR FS, MapR DB or OpenTSDB. The tools used are, Sqoop, Spark Streaming, Logstash or Bash codes
Real-Time App Data Processing @ Multiple Customers	We performed for several customers on their raw application data log processing, ingestion and visualization. Depends on the type of the data, we used the most fitting methods. Be it, Logstash, Bash processing, Spark, Drill or Sqoop For visualization, we used Elasticsearch + Kibana solution to show real time data ingestion
Data Science Engagements @MapR	Use Case Discovery with several customers Machine Learning code developed in Scala, Spark, H2O Lead successful workshop with customer on Machine Learning on Hadoop Delivered successful engagements in Use Case Discovery and Code development Developed Machine Learning on Hadoop course and lab material

Selected Commercial/Government Development Experiences - continued

Malware Expression Detection @Vectra	Vectra Networks specializes in detecting malware expression on packet flow. The product sits on the client's network, sniff packets then find malware expression in the packet flow. It is impossible to detect the infection, but the malware expressions are pretty limited. Among many expressions of malware, I specialized in DDoS detection (detecting client's asset joining DDoS attack). I analyzed, developed and produced the algorithm to the production level
Mobile App Data Analysis @Seven	Seven Networks optimizes out unnecessary traffics from mobile data use. I developed metrics for internal performance measure. I developed several visualization techniques for our products field performance. My visualization created unique views on our products behavior as well as mobile apps behavior in relation to phone's screen activity as well as radio activity. This led to development of novel optimization ideas and implementations. I used Hive to access Big Data and I used R for visualization. For preprocessing the data for R, I used python and C-Sharp
Career Analysis @Identified	From massive data on people's career (over 40 million people), we constructed people's career path graphs. Used Java/SQL/Python/R. Modeled as HMM with LDA. I used NGramDistance to model emission probabilities from Job function to regularized Job Title. Developed mechanism for learning transition probabilities.
Generalized Integrated Learning Architecture @ASU	DARPA GILA project, 2007. In military campaign through the air, the air-space is scare resource that every unit needs to share. We have access to the data that was recorded from the air-space managers operation. We learned from the data on how he/she selected particular missions and when she/he asked for modification. I made the knowledge representation framework. I designed the machine learning algorithm for the highly skewed data distribution. I coded and delivered in Java