Jason (Jia) Teoh

jiateoh at gmail | https://jiateoh.github.io | linkedin.com/in/jiateoh
Seattle, WA | Open to Remote Work

Industry Experience

Databricks - Spark Structured Streaming Jan ’24 - Present
Software Engineer

- TBD as I only just started - reach out for details!

Acryl Data - Data Platform Sept ’23 - Jan ’24
Software Engineer

- Built multi-region, multi-cloud data ingestion processes for the data lake.
- Built scalable Airflow pipelines across dozens of customers to aggregate aspect-level metadata into structured entity datasets, available for analytics consumption.
- Worked with customer success team and customers to deliver standalone structured entity dataset process as well as data inquiries.
- Planned long-term roadmap for data lakehouse including priorities, use cases, and risks.

Twitter - Streaming/Beam on Flink | Large Data Collider + DataAPI July ’22 - Aug ’23
Senior Software Engineer

- Built new on-premises Beam-on-Flink infrastructure for streaming+batch computation.
- Migrated key use cases and implemented scalable on-premises Beam-on-Flink solutions.
- Developed and supported configuration-driven streaming+batch aggregation and serving platform that supports critical revenue use cases (DataApi)
- Managed analytics query server with pluggable backends (e.g., Bigtable, BigQuery) as well as general-purpose data loading workflows for loading data (Large Data Collider)

Google - Cloud Spanner (https://cloud.google.com/spanner/) June ’19 - Sept ’19
Software Engineering Intern

- Designed and implemented an automated rule-based diagnostics framework to analyze query executions and identify performance problems or anomalies
- Integrated analysis framework with the existing diagnostics infrastructure to visually bind detected issues with corresponding query execution information and repair suggestions

LinkedIn - Thirdeye + LinkedIn Segmentation and Targeting Tool June ’13 - Sept ’16
Senior Software Engineer (5/15- 9/16), Software Engineer (6/13 - 5/15)
ThirdEye (now maintained by StarTree): *10/15-9/16*

- Designed open source anomaly detection framework with support for ML algorithms.
- Onboarded new use cases from data bootstrapping to server setup and infrastructure support for new anomaly detection techniques.

LinkedIn Segmentation and Targeting Tool (LISTT): 6/13-9/15

- Developed self-service application to target member segments via custom user attributes
- Built and supported large-scale multidimensional (500+) member and company datasets derived from user-defined data definitions in Teradata SQL, Hive, MapReduce, and Pig.

Research Experience

Software Engineering and Analysis Laboratory (SEAL) March ’18 - June ‘22

- Investigated debugging precision of big data systems with a focus on root cause analysis
- Combined latency instrumentation with data provenance to compute record-level latency and identify input subsets responsible for causing computation skew
- Improved data provenance trace precision of suspicious or faulty output records in dataflow applications by merging taint analysis with influence-based aggregation provenance
- Generated workloads which reproduce desired performance symptoms by leveraging targeted user-defined function fuzzing and skew-inspired mutation operators

Scalable Analytics Institute (ScAi) Sept ’16 - March ’18

- Designed and built a mobile dataflow computing system that seamlessly integrates cloud and on-device data storage and computation
- Developed Android framework integrated with Spark Catalyst Optimizer to automatically export sub-plans and coordinate with Spark jobs through Apache Kafka

Education

Ph.D., Computer Science, University of California, Los Angeles Sept ’16 - June ’22
Research Focus: Debugging in Big Data Systems, Advisor: Dr. Miryung Kim

M.S., Computer Science, University of California, Los Angeles Sept ’16 - March ’19

B.A., Computer Science, University of California, Berkeley Aug ’10 – May ’13

Publications

[SoCC 2020] Jason Teoh, Muhammad Ali Gulzar, and Miryung Kim. 2020. Influence-Based Provenance for Dataflow Applications with Taint Propagation. *ACM Symposium on Cloud Computing (SoCC ‘20)*. 24.4% acceptance rate

[SoCC 2019] Jason Teoh, Muhammad Ali Gulzar, Guoqing Harry Xu, and Miryung Kim. 2019. PerfDebug: Performance Debugging of Computation Skew in Dataflow Systems. *ACM Symposium on Cloud Computing (SoCC ‘19)*. 24.8% acceptance rate

Skills and Teaching

Programming: Scala, Java, Python, SQL, Bash, TypeScript, Apache Pig, Apache Hive, Javascript, Ruby, R, C, LaTeX
Technology: Spark, Beam, Flink, Kafka, Thrift, Hadoop, ElasticSearch, Lucene, Spring, Kubernetes, Helm, AngularJS, jQuery, Sigma (js), Datafu Hourglass
Teaching Assistantships (UCLA): Database Systems, Software Engineering