

Yan Chen

✉ +1 (647) 766-5566 | 📩 yan.chen.im@gmail.com

Skills

Programming Languages Java, Scala, Shell script, Python, SQL, C, C++, Swift, Objective-C

Frameworks Spark (Core/Streaming/SQL/MLLib), Keras, Tensorflow, MLeap, Hadoop, Spring, Spring Boot, Kafka

Databases MySQL, Oracle, PostgreSQL, MongoDB, Redis, SQLite, SQL Server, Teradata, HBase

Cloud Technologies Amazon Web Services (AWS), Google Cloud Platform (GCP)

Dev Tools & Others IntelliJ, Eclipse, XCode, Git, SVN, Jenkins, Maven, Gradle, sbt, Docker, Vim

Experience

Financial Industry Regulatory Authority (FINRA)

Toronto, Canada

Software Engineer (Contract)

09/2017 - Present

- Responsible for building the entire life cycle of a machine learning system, including problem definition, data collection, data preprocessing, model training/selection, model serving, as well as metrics monitoring, with Spark MLLib, Keras and MLeap to help with the review process of financial disclosures submitted by stockbrokers.
- Worked on building a framework to be used as the underlying library of Spark-based ETL jobs, which is able to handle data sources/destinations including HDFS, S3, RDS (Oracle, PostgreSQL, etc), NoSQL (MongoDB, DynamoDB, etc), ElasticSearch, etc.
- Built an annotation-based data validation framework for field level data validation, utilizing both Spark and Hibernate Validator.
- Worked on a disclosure review system, with Spring Boot microservices and AWS infrastructures including Lambda, SNS, SQS, S3, ECS, ECR, etc.

The Bank of Nova Scotia (Scotiabank)

Toronto, Canada

Data Engineer

02/2017 - 09/2017

- Worked as a team lead on building a tool with Spark Streaming for ingestion of data in CSV and COBOL format into HDFS and Hive tables, and integrated with HPE Data Security to provide real-time data encryption.
- Improved open-source Java-based COBOL parsing libraries to handle more variants of COBOL formats, and provided easy-to-use COBOL parsing APIs to other internal services.
- Led the proof of concept of HPE Data Security for data encryption, tokenization and masking.
- Engaged in proof of concepts of using custom NiFi processors to parse COBOL data in real-time.

Royal Bank of Canada (RBC)

Toronto, Canada

Big Data Developer

02/2016 - 02/2017

- Worked as the main developer on designing and building data pipelines with Spark Streaming for real-time data processing (with complex business logic) using a mix of Scala (mainly) and Java.
- Leveraged HBase as NoSQL database in the application; designed the initial version of the whole HBase schema.
- Used NiFi and its custom processors as a part of the whole data pipeline.
- Deployed and managed a temporary separate 3-node HDP cluster with KDC for proof of concepts purposes. Researched and solved the problem of kerberos authentication with two different clusters with their own KDC's in the same client.
- Engaged in proof of concepts of several technologies including NiFi, HBase, Akka, etc.
- Hosted seminars on Spark execution mechanisms, performance optimization, etc, for internal training purpose.

Data Mining Lab, York University

Toronto, Canada

Research Assistant

07/2016 - 03/2018

Research Assistant

05/2015 - 03/2016

Teaching Assistant

09/2014 - 04/2015

Teaching Assistant

09/2013 - 04/2014

- Designed, implemented and evaluated a sampling strategy and a distributed data mining algorithm on Apache Spark for high utility itemset mining.
- Deployed and managed a cluster of 21 instances on AWS and another cluster of 8 machines in the Data Mining Lab for Hadoop and Spark research environments.
- Implemented an algorithm in contrast pattern mining to mine a dataset, in order to find interesting differences among different groups of people and built a recommendation system with Play framework, Akka and Redis.
- Worked as a Software Developer intern on-site for Dapasoft Inc. for 3 months under a project of the BRAIN (Big Data Research, Analytics, and Information Network) Alliance.

Insigma Hengtian Software Ltd.

Hangzhou, China

Software Engineer (Intern)

03/2013 – 06/2013

- Worked as a consultant intern on-site for Cisco.
- Designed and developed workflows for data ETL (Extraction, Transformation and Loading) from multiple data sources to Teradata on Informatica PowerCenter.
- Implemented MapReduce data processing procedures with Apache Hadoop in Java to process web logs data.
- Developed Python scripts for faster transformation from the workflow scheduling design to shell scripts in Orsyip Dollar Universe.
- Engaged in workflow development on Apache Hive for a business intelligence reporting web system.

Education

York University

Toronto, Canada

Master of Science in Computer Science

2013 - 2015

- Admitted as one of the only two fully funded master's international students.
- Research in the field of Data Mining with Prof. Aijun An.
- Thesis: Approximate Parallel High Utility Itemset Mining

Simon Fraser University

Vancouver, Canada

Exchange Program in Computer Science

2011 - 2012

- Only 2 students in the major were selected for this course-based exchange program.
- Coursework includes: Computer Architecture, Software Engineering, Networking, Operating Systems, Web-Based Information System, Numerical Analysis, etc.

Zhejiang University

Hangzhou, China

Bachelor of Engineering in Software Engineering

2009 - 2013

- Graduated top 3 in the major from the HE-Zhijun Honored Class.
- Coursework includes: Data Structures, Object-Oriented Programming, Database Systems, Computer Organization, Discrete Mathematics, etc.

Honors & Awards

2016	EIM Q4 Royal Performance Award	Royal Bank of Canada
2016	EIM Q4 Royal Performance Team Award	Royal Bank of Canada
2013	International Tuition Fee Scholarship	York University
2013	York Graduate Scholarship	York University
2013	Honor of Distinguished Engineering Talent	Zhejiang University
2013	Honor of Outstanding Graduates of Zhejiang University	Zhejiang University
2013	Honor of Outstanding Graduates of Higher Education in Zhejiang Province	Department of Education of Zhejiang Province
2012	Google Excellence Scholarship	Google Inc.
2011	Star-net Scholarship	Star-net Communication Co., Ltd.
2010	Sumitomo Mitsui Banking Corporation Scholarship	Sumitomo Mitsui Banking Corporation

Selected Projects

Semantic Analysis of Movie Reviews using Character N-gram

2014

- Implemented out-of-place distance measure and Naive Bayes classifier for semantic analysis of movie reviews by using character n-gram.
- Conducted leave-one-out cross validation to test the accuracy of the two classifiers.
- Open-sourced: github.com/nrthyrik/n-gram

MiniDB: Mini Database Engine in C++

2014

- Implemented a simplified database engine in C++.
- Functionality includes most of the basic SQL operations as well as an index on a B+ tree.
- Open-sourced: github.com/nrthyrik/minidb

Super Mario

2010

- Implemented a remake of the classic video game Super Mario on DOSBox.
- Wrote around 2,000 lines of C code without using any graphics engine or game engine. All the graphics were drawn pixel by pixel.
- Open-sourced: github.com/nrthyrik/super-mario

Publications

- **Chen, Y.**, Yann, M. L. J., Davoudi, H., Choi, J., An, A., & Mei, Z. (2017, May). Contrast pattern based collaborative behavior recommendation for life improvement. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 106-118). Springer, Cham.
- Zihayat, M., **Chen, Y.**, & An, A. (2017). Memory-adaptive high utility sequential pattern mining over data streams. *Machine Learning*, 106(6), 799-836.
- **Chen, Y.**, & An, A. (2016). Approximate parallel high utility itemset mining. *Big data research*, 6, 26-42.