Adrian Ng, MSc.

Seeking Junior-Level Data Engineering Opportunities **Email:** contact@adrian.ng **Website:** adrian.ng

Profile

I am a Computer Science graduate passionate about programming and a career in Data Engineering. I seek opportunities that meet my growing experience in *Java* – a language I have used in numerous academic projects ranging from the implementation of financial models to large-scale data processing with *Apache Hadoop* and more.

Prior to postgraduate study, my expertise in *SQL development* focused on the implementation of segmentation processes for a number of clients including: *Virgin Media*, *TUI*, *UPC*, *MSD*, *Volkswagen*, and *KwikFit*.

After graduation, my accomplishments as a Data Analyst at *Manchester City FC* leaned more towards Data Engineering, which leads me now to pursue a career in this field.

Education

- **Royal Holloway – Department of Computer Science** Sept. 2016 – Dec. 2017
  Master of Science in Data Science and Analytics *with Distinction*
- **King's College London – School of Engineering** Sept. 2007 – July 2010
  Bachelor of Engineering in Mechanical Engineering *Upper Second Class with Honours*

Technologies

Languages: • Java 8 • SQL

Software: • IntelliJ IDEA • SQL Server Management Studio • Git • VS Code • Jira • Maven

Java Projects

- **Implementation of Value at Risk (VaR) measure** (https://adrian.ng/java/var/) (https://github.com/Adrian-Ng/VaR)
  I implemented a number of approaches to estimating *VaR*, a measure of risk, for a (hypothetical) investment portfolio (stocks, options, deltas). A Normal Distribution $N(0, \sigma^2)$ was assumed for daily price changes. Therefore, a number of moving average processes were implemented for estimating variance σ^2. In addition, an implementation of the *Levenberg-Marquardt* algorithm was used for optimisation of *GARCH(1,1)* parameters.

  VaR Measures
  - Model Building
  - Historical Simulation
  - Monte Carlo Simulation.

  Moving Averages
  - *Equal Weighted*
  - *Exponentially Weighted Moving Average (EWMA)*
  - *GARCH(1,1)*

  I made use of object-oriented techniques and patterns to accommodate these numerous approaches. In addition, I used `Java's` concurrency APIs to parallelize the 100,000+ random walks generated by *Monte Carlo* when simulating stock prices. Real world financial data was obtained via *Google Finance* and *Yahoo Finance* APIs.

- **Option Pricing** (https://adrian.ng/java/options/) (https://github.com/Adrian-Ng/OptionPricer)
  As part of the postgraduate module, *Methods of Computational Fianance*, I implemented three approaches to estimating option prices: ∘ Monte Carlo Simulation ∘ Black Scholes ∘ Binomial Trees. And where applicable, I computed the *payoff* for American, Asian, and European options. The Black Scholes approach assumes a Normal Distribution, so Apache Commons Math API was used.

- **Data Mining with Hadoop MapReduce** (https://github.com/Adrian-Ng/HadoopEnron)
  During my postgraduate module *Large Scale Data Storage and Processing*, I wrote a number of *MapReduce* applications. These involved mapping the communications network from a large collection of emails from the *Enron Corpus* and the aggregation of *Twitter* data. I ran my applications on a self-hosted, single-node cluster as well as on the university's distributed cluster. To load/extract data in *HDFS*, I used `hdfs dfs` commands.

  - **Apache Spark** (https://adrian.ng/scala/spark/enron1)
    In a self-imposed exercise, I translated some of these *MapReduce* applications to `Scala`. This code was less verbose and ran in an *Apache Spark REPL*, which could still interface with *HDFS* via `sparkcontext` APIs.

- **Java 8 Streams with financial data** (https://adrian.ng/java/yahoofinance/#stream)
  A small exercise involving the use of *Java 8 Stream* APIs. Processing real-world financial data to return *mean* and *variance* of some market asset.

Manchester City Football Club

Data Analyst *Fan Relationship Management* *Jan. - July 2018*

- **New York City FC Project:** I took ownership of this project to integrate *NYCFC's* transactional and demographic data with *City Football Group's* data-warehouse. This six-month project involved many phases including: discovery, engineering, and analysis. Data came from multiple external sources each with differing schema: *NYCFC*, *Ticketmaster Salesforce*, *Major League Soccer*.
    - **Data Pipeline:** I implemented a data pipeline to ingress data from multiple databases. This process was encapsulated in *stored procedures* in which I wrote appropriate DML & DDL (`OPENQUERY`, `MERGE`) for efficient ETL. This pipeline replaced the slower front-end *Informatica* solution.
    - **Data Cubes:** I used an aggregated dataset to compare the distribution of `NULL` values. These analyses were transformed to *Data Cubes* to pre-compute every possible roll-up/drill-down. As such, bandwidth was minimised across our distributed servers and need for real-time computation in *Tableau* front-end was eliminated, resulting in an improved user-experience.
    - **Mentoring:** As part of this project, I dedicated time to mentoring a junior colleague remotely in New York. I organised weekly workshops to teach basic DML and more advanced DDL with a goal toward self-sufficiency in writing database queries and stored procedure implementation/scheduling. Additional material on my website helped supplement these workshops.
- **GDPR Preference Pipeline:** I worked on the integration of a GDPR preference pipeline into our data stores (*SQL*, *Salesforce*) and the subsequent refactoring of numerous processes downstream. I worked with the SQL development team and provided specification and UAT testing. I built an efficient, automated `MERGE` process using relational database design (primary key constraints, clustered indexes, triggers).
- **Customer Churn Model:** In an intra-team project to estimate each customer's (MCFC, NYCFC) propensity to churn in following seasons, I assisted with feature selection, contributing to data extraction, imputation, and normalisation. Model selection landed on *logistic regression* but I also researched alternate models (involving more probabilistic assumptions) such as *Beta-Geometric/Beta-Bernoulli*; for which I looked at corresponding academic papers and implementation in `R Studio`.

Creator (now Inspired Thinking Group)

Senior CRM Campaign Executive *SQL Development* *Dec. 2013 - Sept. 2016*

The majority of my work in this role involved working with `SQL` processes which were used to transform customer data into CRM segmentations. Having been promoted to the senior position, I developed a number of these processes. On occasion, I held responsibility for resourcing and managing the team's workload using *Jira*.

- **Virgin Media Segmentation** (https://adrian.ng/SQL/cte/Recursion/ (https://adrian.ng/SQL/misc/openquery-xml)
  I built an end-to-end segmentation process in `SQL`. This included building a fast, flexible, and bespoke import tool around `BULK INSERT`. Remote server queries (`OPENQUERY`) made use of `XML` to effectively `INNER JOIN` local and remote tables resulting in speed and minimial resource use on a busy live server. Recursive queries were used to implement a solution (similar to `flatMap` in `Java 8` or `Scala`) for efficient regex.
- **Volkswagen Onboarding:** I worked with `.NET` developers and project managers to bring Volkswagen on-board as a new client. This required implementing a new segmentation process for broadcasting email *and* SMS. In addition, I provided specification to developers for their data warehousing/archiving ingress schema.
- **TUI Redesign:** I collaborated closely with the TUI client to redesign the existing *Thomson* and *First Choice* mailings. `TCL` scripts were developed to dynamically merge fields into the `HTML` body. For my efforts on this three-month project I gained recognition with the client and was rewarded.

Seatwave (now Ticketmaster)

Marketing Analyst Intern *Commercial Team* *May 2013 - Dec. 2013*

Using *SQL Server Management Studio* for the first time, I wrote DML capable of querying the transactional/customer databases to return data for warehousing, reporting, and segmentation. I also worked on pricing and spatial analyses, using *QGIS* as a visualisation tool.