**GRACE LIANG**

+1 202-390-8551 | [graceliang93@gmail.com](mailto:graceliang93@gmail.com) | LinkedIn:

<https://www.linkedin.com/in/xiaoxuan-liang-99888ba2/>

**Summary:**

* Highly motivated and data-oriented **Big Data Engineer** and **Data Science Associate** with working experience in the domains including **finance**, **social media** and **government.**
* Worked as a dedicated professional Data Engineer with a solid background in **Apache Hadoop ecosystem** like **HDFS**, **MapReduce**, **Spark**, **Hive**, **HBase, Yarn,** **Kafka**, **Pig**, **Sqoop** and **Zookeeper.**
* Implemented massive **ETL** (Extract, Transform, Load) workflow from different sources to different destinations **(AWS redshift S3, Microsoft SQL server database, Oracle database, Hadoop HDFS, flat files**).
* Competent in building data pipeline by collecting, aggregating and moving large amounts of **streaming data** using **AWS SDK, Amazon CloudWatch Logs** and **Spark Streaming**.
* Experience with data processing/transformation/aggregation using **Data frame** and **Dataset** **API**.
* Proficiency in developing **Scala Spark** program with **Scala** and **Python API** to perform large scale data collection, data mining, data wrangling, data transformation, data integration and data quality.
* Experience in importing and exporting buck of data using **Sqoop** from **HDFS**/**Hive**/**HBase** to **RDBMS.**
* Experience in developing scalable solutions using NoSQL databases including **Cassandra**, **HBase** and **MongoDB**
* **Certified AWS Solution Architect Associate**
* Solid understanding on the working of **EC2, EMR, SNS, SQS, VPC, Lambda, Glue, DynamoDB, RDS, SageMaker** in Amazon Web Services
* Proficiency in writing different **SQL queries** such as **HQL, Impala SQL, SparkSQL, T-SQL, MySQL.**
* Proficient in **Python** programming on **PyCharm**, **Anaconda Navigator** (**Jupyter Notebook**), **Databricks** and **Zeppelin notebook** for **web scraping**, **NLP**, **Image Classification**, optimizing, plotting with packages including **PySpark** **Pandas**, **NumPy**, **SciPy**, **Scikit-learn**, **Matplotlib**, **Seaborn**, etc.
* Proficient in **Statistical Modeling, Data Mining and Machine Learning Algorithms** in Data Science/Forecasting/Predictive Analytics such as **A\B testing, Linear and Logistics Regression, Random Forest, K Means, Neural Network, Decision Trees, SVM, Naive bayes, Feature Engineering and Dimensionality Reduction.**
* Skilled in **Deep Learning** Framework: **TensorFlow, Keras and PyTorch**; Familiar with Deep Learning Models like **CNN, RNN and LSTMs.**
* Created **Tableau 2020.1 dashboards,** visualizations, and performing advanced analytics.
* Experience in **data mining** with **structure and unstructured data** using **SQL, SAS, advanced MS Excel (VLOOLUP, Pivot tables)** for supporting Data Cleaning.
* Familiar with software development and version control tools like **Git** and **JIRA.**
* Experience in using **Docker** as deployment environment and run applications by using **Docker containers**.

**CERTIFICATION:**

* **AWS Solution Architect Associate**
* Spark and Python for Big Data with PySpark | Udemy
* Apache NiFi Complete Master Course- Automation ETL | Udemy
* Snowflake Decoded | Udemy

**TECHNICAL SKILLS:**

**Hadoop/Spark Ecosystem: \ Programming Languages: \**

Hadoop 2.x, MapReduce, Spark 2.x, Pig 0.12, \ Python2.7/3.6, Scala 2.11.X, R, SQL,

Hive 0.14, Sqoop 1.4.6, Flume 1.6.0, Kafka, \ SAS\

Yarn, Zookeeper 3.4.x\

**AWS: \ Data Science: \**

EC2, SNS, SQS, VPC, Lambda, DynamoDB, \ Machine Learning, Deep Learning with \

RDS, EMR \ TensorFlow, Keras and PyTorch\

**Database: \ Version Control: \**

MySQL 5.x, Oracle 11g, Cassandra 2.1.x, \ Git/GitHub\

HBase 0.98, MongoDB 3.2\

**Tools: \ Operating Systems\**

Docker, Microsoft office, Eclipse, Jupyter\ Linux, Mac, Windows \

Spyder, IntelliJ IDEA, Databricks, Snowflake\

Zeppelin, Azure\

**Professional Experience:**

**Marlabs Inc, Piscataway, NJ JUN 2020 – Till Date**

**Client: PRA Group lnc.**

**Role:** **Big Data Developer**

**Project: Big Data Automatic Processing**

**Description:**

PRA Group is a consumer debt buying and collection company**.** The project was to design a new automated system that combined streaming and batch data for nonperforming loans detection and prevent suspicious solution. The team provided data automatic processing solutions including data ingestion, data cleaning and data storage from various data sources like steaming websites, mobile devices and financial transactions for supporting downstream to drive debts collection and recoveries.

**Responsibilities:**

* Collected, aggregated and moved large amounts of **streaming data** using **Flume, AWS Kinesis** and **Spark Streaming.**
* Used **AWS Kinesis Data Firehose** as **ETL tools** extract data from **Kinesis Data Stream** to do datacollection, transformation, and storage.
* Used **AWS DataPipeline** to schedule workflows in **EMR cluster** whichbased on **Spark** and monitor workloads as data comes into **AWS S3.**
* Deploy services on **AWS** and utilized **Lambda function** to trigger the **Data pipeline**.
* Stored the processed data in **private AWS S3** accessible only within the organization.
* Implemented aggregations using **Spark** **DataFrame API** in **Scala/Python** and sorted the output files in **S3** and build **AWS Glue** tables as **Parquet format**.
* Worked with **SQOOP** import and export functionalities to handle large data set transfer between **DB2** database and **S3/Glue.**
* Loaded **AWS Redshift** with processed output data and analyzed for insights using built-in **SQL queries** and built dashboards using **AWS Athena** and **Quicksight.**
* Perform testing and analysis using **Databricks.**
* Used **Git** for version control, **Confluence** for documentation.

**Environment:**

**MapReduce, Flume, Kinesis, Spark Streaming, Sqoop 1.4.5, Linux, Yarn, AWS DataPipeline**, **EMR, AWS Athena, Quicksight, Scala 2.11.8, Python 3.4, Git, Databricks, AWS Glue, AWS Lambda.**

## SRI International, Princeton, NJ SEP 2019 – May 2020

## Role: Machine Learning Engineer

**Description:**

SRI is the leading research institution that has rich experience supporting government and industry.

The project involves in studying and developing algorithms and models for adversarial attacks. Our team in addition to provides data process and storage solution, it also provides deep learning pipelines improvement, such as hyperparameters tuning and deploying models at scale using HDFS and Spark.

**Responsibilities:**

* Stored processed train and test datasets **HDFS** and **Hive tables** in **Binary Format**.
* Wrote **Python** scripts to speed up data collection and feed into TensorFlow from **HDFS** using **Hadoop**.
* Used **Spark** and a cluster of machines to improve deep learning pipelines and single and ensemble models training with **TensorFlow/Keras**.
* Utilized **Spark SQL** with **Data Frames API** and **MLlib** to provide efficiently data preprocessing and to train models with L2 regularization using **L-BFGS** and **stochastic gradient descent (SGD)** optimizer by using Scala.
* Deployed baseline **TensorFlow**, **Keras** and **PyTorch** models with CNN architecture of **Resenet**, **Lenet** and **Mobilenet,** and distributed TensorFlow training using **Amazon SageMaker**.
* **Created APIs** and database by using **MySQL workbench** and packages—**PyMySQL** and **Flask**. API calls for comprehensive evaluation of **adversarial robustness**.
* Provided testing infrastructure for **DARPA** to evaluate the significance of adversarial robustness of invented **quantized linear classification models** and over 10 CNN defend models with different types of attacks.
* Designed **statistical metric functions** from logits and SoftMax layers using **Python** for assessing model performance and features diversity.
* Visualized the insights of hidden information and provided actionable model optimization suggestions by statistical measurements: **Log Loss, Cross-entropy** and **KL Divergence.**
* Used **Git** for version control and **JIRA** for project tracking.
* Analyzed experimental results and created dashboards using **Tableau**

**Environment:**

**Spark, HDFS, Hive, MySQL, AWS EMR, Amazon SageMaker, Python3.6, Scala, PyCharm, MySQL, SQL, GitHub, Tableau, Jupyter Notebook, Scikit-Learn, TensorFlow, Keras, Pytorch.**

**D.C United*,* Washington, D.C. SEP 2018 - AUG 2019**

**Role: Data Analyst**

**Description:**

D.C United is a sport team that plays in Major League Soccer Championship. This project is mainly about implementing a predictive model based on historical Major League Soccer (MLS) data. Our team’s job was to ingest various market and web data into the platform using Flume, process and analyze data with Spark, Hive and HBase, and assist data scientists to present the result in BI/Notebook tools.

**Responsibilities:**

* Involved **ETL** processes including data cleaning, data processing and data storage.
* Developed **Hadoop/Spark** applications under Cloudera environment to automate deployment, configuration and integration data across different systems.
* Used **Flume** to monitor and collect real-time log data from the web and sink log-data into Message Queue of the **Kafka**.
* Stored the processed train and test datasets in **HDFS** and **Hive tables** in **Binary format**.
* Applied **Spark** using **Scala** to do the data batch processing and store the output in **HBase** for scalable storage and fast query.
* Used **Sqoop** to export data from **HDFS** into **MySQL** database to create interactive dashboards using **Tableau.**
* Conducted **statistical analysis** on the possession time and continuous passes and evaluated the performance of D.C United and other MLS teams, using the 2017 and 2018 data from OPTA.
* Studied the correlation between the number of passes and the scoring chances using **R** and **Python** by applying classification tree and other classification models.

**Environment:**

**Linux, UNIX Shell, Hadoop distribution of Cloudera 5.9, Apache Spark 2.X, Kafka, Sqoop, Flume, HDFS, Hive, MySQL, Python3.6, R, Tableau.**

**Xiaoi Robot Technology Co. Ltd, Shanghai, CN May 2015 – Dec 2017**

**Role: Big Data Developer**

## 

**Description:**

Xiaoi Robot Technology is one of the leading intelligent robot technology providers in China.

The project aims to create a new machine learning algorithm for government case allocation and Power Grid Company business in order to improving working efficiency. It involved working on cloud business intelligence. The application loaded large amounts of data from different government systems.

**Responsibilities:**

* Designed, developed, implemented, testing and maintenance of data ingestion and integration **ETL** pipelines including **batch processing,** **Spark** **streaming** and Big Data APIs.
* Developed and optimized **Spark Scala** programs to perform data enrichment, transformation and wrangling and to process real tine data from **Kafka.**
* Handled importing of data from various data sources, performed transformations using **Spark** and loaded data into **Hive Database** and **HDFS** for **analysis.**
* Implemented **Partitioning, Bucketing** in **Hive** for better organization of the data.
* Collected, cleaned, and formatted batch government case data in **relational database**, maintained database queries through **MySQL** for studying predictive models, and generated reports to identify issues in database.
* Created and implemented a **web crawler** with **Python** to scrape business data from the web to facilitate data training. And developed and designed **APIs using Flask** framework**.**
* Cleaned and conducted **exploratory data analysis (EDA)** with **700,000 text** **descriptions** of feedback calls records to produce insights for decision making and support **feature engineering of Natural Language Processing (NLP)**
* Worked in an interdisciplinary team to build and optimize the **machine learning algorithms** (**Random Forest, SVM**, etc.) of government cases allocation called Autoflow, which accelerated case **classification accuracy to over 90%, reduced government processing time by 22 hours, and raised project revenue by approximately 8%**
* Analyzed the data and created dashboards using **Tableau** and **R Shiny**.
* Collaborate and tracking the work with **Git**.

**Environment:**

**Linux, MySQL 5.x, Python 3.6, Scala 2.11.8, Flume, Spark, Sqoop 1.4.6, Hive 0.14, HDFS, Kafka, EDA, NLP, Feature Engineering, Oracle, Flask, R Shiny, Tableau, Git,** **Random Forest, SVM.**

**EDUCATION:**

* **Master of Science in Data Science**
* **Bachelor of Art in International Finance**