Runhan Yu

(781)786-1586 • runhanyu90@gmail.com
runhanyu.com •linkedin.com/in/runhanyu • github.com/ziyunch

SKILLS

Programming languages	Python, SQL, Java, R, Matlab, Lisp, Javascript, Bash
AWS technologies	EC2, Redshift, Glue, Athena, EMR, Lambda, S3, RDS, DynamoDB
Distributed computing tools	Spark, Airflow, Kafka, Hive, Hadoop, MapReduce
Others	Linux, git, Django, Emacs, Tableau, Vim, UNIX, CSS, HPC

EXPERIENCE

Amazon Seattle, WA
Data Engineer II *Feb. 2020 - Present*

- Design and build an end-to-end platform to measure product adoption metrics and serve historical and real-time reporting and predictive analytics needs of the org.
- Employ a request collection system and write automated data checks to reduce data quality issues.

Rescale San Francisco, CA
Data Engineer *Oct. 2019 – Feb. 2020*

- Developed ETL pipelines and built a data metrics collection system which brought cross-platform analysis from 1 day to 10 seconds (Redshift, Glue, Spark).
- Built a bot to ingest, aggregate and detect error events and provided developers a view for error triage (Python).
- Employed new features for Rescale platform to explicitly track platform activity (Python, Django, Java).
- Wrote complex SQL queries for multiple teams to aggregate platform usage data for performance monitoring, reporting and decision making (SQL, Athena).

Insight Data Science New York, NY
Data Engineer Fellow *Jan. 2019 – Sep. 2019*

- Developed an ETL pipeline to extract, integrate and transform prescription data from multiple providers in AWS cloud computing to enable nation-wide queries on prescription drug usage (Python, AWS, Airflow).
- Validated and combined public available Medicaid and Medicare datasets with NIH, FDA and NPPES sources into a SQL queryable databases in Redshift, visualized in website (SQL, JavaScript, CSS).
- Implemented custom connector to Redshift/PostgreSQL with 20 times more efficiency (Python).
- Built a real-time monitoring pipeline of IoT sensor data and latencies for data center management to handle 10,000+ events per second (Spark streaming, Kafka).

Brandeis University Waltham, MA
Graduate Research Assistant *Oct. 2012 – Dec. 2018*

- Developed programs to parse massive experiment data into structured analysis and visualization (R).
- Built models to tackle metrics and simulate enzyme kinetic mechanisms (Matlab).
- Developed a python-based pipeline that extracted 100,000 gene sequences encoding protein of interest from 200 million gene sequences in the 114 GB GenBank database from RESTful API (Python).
- Managed 24 students and collaborated with 3 teams with multi-disciplines and multi-cultures.

PROJECTS

Multithread Web Crawler

- Write a class to handle multithreading website crawling inside the given domain.
- Feature a breath-first search algorithm and a multithread pool to visit all links asynchronously.
- Handle various status code, time-out and exceptions in a structured manner (Python).

Stomach Cancer Gene Variation - bit.ly/2sBSHTW

- Extracted cancer research data from the Cancer Genome Atlas Network (R).
- Mapped gene list with biological annotations with functional annotation tool DAVID (Python).
- Summarized and visualized stomach cancer-related gene candidates with R package MAFtools (R).

EDUCATION

Brandeis University Waltham, MA
Ph.D. in Chemistry, with specialization in Quantitative Biology *May. 2019*

Nankai University Tianjin, China
B.S. in Materials Chemistry *Jun. 2012*
B.S. in Finance *Jun. 2012*