Jordan Tang

Floral Park, NY | 516-761-7623 | jtang0819@gmail.com | LinkedIn/in/jtang0819 | Github/jtang0819

SUMMARY

Data Engineer with 4+ years of experience in constructing robust data pipelines using Amazon Web Services (AWS), SQL, and Python. Skilled in ETL processes and optimizing data workflows, consistently enhancing system efficiency and data reliability. Proven track record of improving system efficiency and data integrity, ready to leverage expertise in a dynamic new role.

TECH STACK

Languages: Python, SQL, Bash
AWS: Lambda, Step Function, S3, SNS, Eventbridge, EC2, Load Balancer, Route 53, RDS, Glue, Secrets Manager, Eventbridge, Security Groups, VPC, Redshift, DMS
Azure: App registration, Blob Storage, DevOps, Pipelines
Infrastructure: Terraform, Docker
Databases: Microsoft SQL Server, Snowflake, AWS Redshift, Postgres
Applications/Tooling: Informatica 10.5.4, HVR, Snaplogic, SAP WEBI, PowerBI
Misc: Github, GitHub Actions, Jupyter Notebooks, Apache (Spark, Airflow, Kafka), Network Testing, AWS Linux, Gitlabs
Certifications: AWS Cloud Practitioner

PROFESSIONAL EXPERIENCE

Principal Financial Group, Remote
Data Engineer II *Aug 2024 - Present*

- Gained hands-on experience with Informatica PowerCenter by designing and implementing scalable ETL solutions.
    - Developed and updated Informatica PowerCenter mappings, sessions, and workflows for efficient data processing
    - Built parameterized mappings and reusable transformations to enhance maintainability and reduce redundancy
- Learned best practices for ensuring data quality and governance in ETL processes in a data warehouse environment
    - Implemented error handling and logging mechanisms to track and resolve ETL failures
    - Standardized data formats and transformations to ensure consistency across reporting and analytics systems
- Maintained team Salesforce data extraction application and its resources in AWS
    - Updated and secured github actions deployment workflow with secured parameters
    - Customized and deployed enterprise standard pipeline to fit our team needs
    - Implemented organizational standard structured logging for the data extraction application written in Python
    - Mentored engineers on enhancing their Python coding, performance optimization, and debugging techniques

Mindex Technologies, Remote
Data Engineer Consultant *Oct 2023 - May 2024*

- Delivered technical expertise as a data engineer to a leading calibration and compliance company
    - Led re-architectural efforts for the core data pipeline, addressing operational inefficiencies and data redundancy decreasing monthly aws cloud bill from $20,000 to an estimated $10,000
        - Submitted analysis of existing pipelines, identifying flaws, discrepancies, and potential optimizations
            - Created comprehensive documentation for existing datasets, their sources, and transformations, offering clear insights to all stakeholders
            - Illustrated current architecture in detail using draw.io, offering a detailed visual depiction of all involved components and connections
        - Designed a comprehensive data architecture plan that addressed scalability, data quality, maintainability, and performance issues
        - Developed a successful Proof of Concept (POC) using AWS Glue, demonstrating the feasibility and benefits of the re-architecture proposal
    - Ensured the continued functionality and reliability of existing data pipelines
        - Significantly improved processing efficiency and reduced execution time by migrating existing AWS Glue jobs from version 2 to version 4 and optimizing a 2 hour -> 30 minutes Glue job during the upgrade
        - Identified and resolved data/code issues that arose from the existing data pipelines - ‘keep the lights on’

- Performed optimizations and enhancements for SchoolTool - a student management solution by Mindex
    - Enhanced a dataset with 3rd party data from an external SFTP server
        - Updated an AWS Glue script with functions to extract files from SFTP location with Paramiko and transform/filter with Pandas
    - Optimized critical analytical SQL queries
        - Refined subquery joins to reduce total time from 3 hour to 15 minute run time
        - Investigated a query running for 3 hours in production, but only 15 minutes in dev; ultimately rebuilding or reorganizing indexes with high fragmentation and changing compatibility level improving runtime to 3 minutes

NewRez LLC, Remote
Data & Integration Engineer *Oct 2021 - Oct 2023*

- Implemented a resilient cloud-based enterprise data platform capable of handling near real-time ingestion/transformation and analytics workloads to alleviate the operational burden/costs with on-prem legacy SQL servers allowing us to downgrade one of our production MSSQL server core counts from 128 -> 64 saving roughly $400,000 on licensing fees
    - Utilized an ELT strategy using HVR for CDC replication/table refreshes and Snaplogic for data transformations
    - Wrote Terraform scripts to deploy to AWS and a GitHub Actions CI/CD pipeline
        - AWS Resources: Route 53, Load Balancers, S3, EC2, Security Groups, RDS
        - Wrote bash startup scripts for EC2 deployment for both HVR and Snaplogic
    - Planned & administered Snowflake target environment to store core reporting tables and datasets
        - Organized Snowflake to optimize user interactions within team sandbox environments, ensuring data integrity and quality of the replication processes
            - Replication, Refined, Curated, and individual team sandbox databases
        - Migrated legacy SSIS reporting workloads to Snaplogic for Snowflake
        - Reconfigured and simplified IAM roles and users
        - Deployed masking for PII information on legacy SQL servers and Snowflake for end-to-end protection
        - Stored database objects in GitHub for version control
    - Configured LogicMonitor for EC2 instance health with alerting focused on memory and HDD space
- Designed a new workflow for the Control M orchestration tool to enable a programmatic approach to job deployment
    - Wrote a Python script to capture the job data from the development environment to store in Github
    - Utilized a CI/CD pipeline with Azure Pipelines that executes a custom-written Python script to utilize the Control M API Endpoint for production deployment
- Recognized as the subject matter expert(SME) for HVR, led initiatives for enhancements, problem resolution, and maintenance, and provided guidance and solutions related to its use within the data platform
    - Wrote bash scripts for HVR table refresh for Control M to orchestrate
        - Used GitHub Actions to create a CI/CD pipeline for deployment
    - Enabled HVR error alerting with AWS SNS scripted in Terraform
    - Functioned as the initial point of contact for troubleshooting issues
    - Served on-call and documented troubleshooting solutions on team Azure DevOps Wiki
    - Proposed a patching strategy to keep HVR available by leveraging AWS resources
        - Wrote the bash script for patch deployment on cloned EC2 instance
- Developed, maintained, and then migrated AWS data pipelines using the new deployed tooling to support 20+ sponsored datasets and PowerBI Dashboards centered around engagement analytics:
    - Created new data pipelines within an already established AWS-based environment: Glue, S3, Redshift
        - Performed API testing with Postman and Bruno to validate data integrity before productionalizing endpoints
        - Developed Glue jobs to extract and transform data from internal/external APIs and databases, leveraging Pandas for processing and AWS Wrangler to store Parquet files in S3, before copying them into Redshift
        - Partnered with ServiceNow administrators to troubleshoot API issues and resolve missing data fields
        - Integrated diverse data sources including SharePoint, Power BI, Dynatrace, ServiceNow, and SQL Server
        - Automated production deployment via Terraform and GitLab
    - Successfully demonstrated how to access PowerBI dataflows using the Execute Queries API, enabling the team to upgrade multiple jobs from a previously unreliable architecture to the current team AWS solution

    - Created data profiles in PowerBI to enable proactive data quality monitoring and troubleshooting
    - Migrated all jobs using the new tooling: Control M orchestrated HVR table refreshes to Snowflake and Snaplogic jobs to transform the data to its final state
    - Prototyped serverless pipelines to move on-prem SQL tables to Snowflake as part of the Customer 360 initiative
        - Utilized Control M for orchestration with on-prem SSIS packages
        - Used Lambdas to run Python scripts and Step Functions to orchestrate the lambdas
        - Implemented logging and notifications using AWS lambda and SNS for improved visibility
- Developed and deployed a Flask app on AWS EC2 that enabled internal restricted-access users to search Snowflake and download Ringcentral client call recordings from a S3 bucket, saving ~50 minutes monthly

ChyronHego, Melville, NY
Business Operations Associate *June 2019 - April 2020*

- Coordinated the development of an interdepartmental workflow that reduced the monthly ticket count from 400 to 100
- Led training, documented workflows, and managed software environment to support a successful rollout of SF app Kimble

PSL Group, Manhattan, NY
Financial Systems Analyst *Nov 2016 - June 2019*

- Streamlined payroll workflow with tax compliance team by developing an automated process utilizing Kettle, MSSQL database/tables for storage, SAP Universe/WEBI for reporting and delivery to reduce a 7-day process to on-demand
- Wrote a Node.js workaround to a limitation in Google Script in the workflow of a high-visibility Google Sheet Sales Report

PROJECTS

Recipe Ingestion *May 2024 - Present*

- Created an end-to-end pipeline to extract and process recipe data from online cooking websites; goal is to not use Glue
    - Wrote Python to leverage REST API and store data as json in S3
    - Used Spark via Pyspark to capture json files onto an EC2 as parquet
    - Wrote a Github actions pipeline for CI/CD onto the EC2.

Data Pipeline Testing/ Learning *Jan 2021 - Sept 2021*

- Created an end-to-end pipeline to ingest and process daily stock market data from multiple stock exchanges
    - Wrote Python to leverage REST API and store Spark dataframes in memory
    - Used Spark vi Pyspark to clean and transform the data
    - Loaded to Azure blob storage for final storage
- Prototyped a high throughput data pipeline to ingest and process a set of random website forum posts
    - Configured Apache Kafka for streaming and collection of data
    - Utilized Spark via Pyspark to clean and transform the data before loading it into a MYSQL database
    - Designed a simple Metabase dashboard to monitor pipeline speed by calculating rows per minute
    - Achieved a throughput of 5000 rows per second for a test set of 1.5 million JSON files

EDUCATION

State University of New York (SUNY), Purchase, NY
Bachelor of the Arts in *Mathematics and Computer Science* *May 2015*

INTERESTS & HOBBIES

Coffee; BBQ - Brisket, Ribs, Chicken Thighs; First[::-1] at Trivia; Board Games - Blood on the Clocktower, Gloomhaven, Frosthaven