Henry A.
Senior Python engineer with automation, data quality and scientist skills

SUMMARY

- 9 years experience with various data disciplines: Data Engineer, Data Quality Engineer, Data Analyst, Data Management, ETL Engineer
- Built and optimized production-grade PySpark pipelines handling millions of records, including complex transformations, joins, aggregations, and backfills, with a strong focus on performance and data quality.
- Automated Web scraping (Beautiful Soup and Scrapy, CAPTCHAs and User agent management)
- Data QA, SQL, Pipelines, ETL
- Data Analytics/Engineering with Cloud Service Providers (AWS, GCP)
- Extensive experience with Spark and Hadoop, Databricks (hands-on with Spark-based pipelines deployed in cloud environments and have used Databricks as the execution and collaboration layer: jobs, notebooks, cluster configs).
- 7 years of experience working with MySQL, SQL, and PostgreSQL;
- 5 years of experience with Amazon Web Services (AWS)
- Google Cloud Platform (GCP): BigQuery, GCS, App Engine, data pipelines, and Azure (basic knowledge)
- Data Analytics/Engineering services, Kubernetes (K8s)
- 5 years of experience with PowerBI
- 4 years of experience with Tableau and other visualization tools like Spotfire and Sisense;
- 3+ years of experience with AI/ML projects, background with TensorFlow, Scikit-learn and PyTorch;
- Extensive hands-on expertise with Reltio MDM, including configuration, workflows, match rules, survivorship rules, troubleshooting, and integration using APIs and connectors (Databricks, Reltio Integration Hub), Data Modeling, Data Integration, Data Analyses, Data Validation, and Data Cleansing)
- Upper-intermediate to advanced English

TECHNICAL SKILLS

Main Technical Skills	Databricks, Python (9 yr.), SQL (6 yr.), PySpark
Programming Languages	JavaScript, Python (9 yr.), R (2 yr.)
Java Frameworks	Apache Spark
Scala Frameworks	Apache Spark

Python Libraries and Tools	Beautiful Soup, Dask, Django Channels, Pandas, PySpark, Python Pickle, PyTorch, Scrapy, TensorFlow
AI & Machine Learning	Machine Learning, PyTorch, Spacy, TensorFlow
R Frameworks	Shiny (2 yr.)
Data Analysis and Visualization Technologies	Apache Airflow, Apache Spark, Databricks, Data Mining, Data Modelling, Data Scraping, Data Testing (3 yr.), ETL, Pandas, Power BI (5 yr.), Reltio, Reltio Data Loader, Reltio Integration Hub (RIH), Sisense, Tableau (5 yr.)
Databases & Management Systems / ORM	Apache Spark, Aurora, AWS DynamoDB, AWS ElasticSearch, Microsoft SQL Server, MySQL, NoSQL (5 yr.), PostgreSQL, RDBMS, SQL (6 yr.), SQLAlchemy
Cloud Platforms, Services & Computing	AWS (3 yr.), GCP (4 yr.)
Amazon Web Services	AWS Bedrock, AWS CloudWatch, AWS DynamoDB, AWS ElasticSearch, AWS Fargate, AWS Lambda, AWS S3, AWS SQS
Azure Cloud Services	Databricks
SDK / API and Integrations	API, GraphQL, RESTful API
Deployment, CI/CD & Administration	CI-CD Pipeline
QA, Test Automation, Security	Data Testing (3 yr.), Selenium, Unit Testing
Version Control	Git
Operating Systems	Linux
iOS Libraries and Tools	MDM
Platforms	Mendix, RPA
Methodologies, Paradigms and Patterns	REST (5 yr.)
Third Party Tools / IDEs / SDK / Services	RStudio
Other Technical Skills	BIGData, Cronjob, Parallelization, Reltio APIs, Reltio match rules, Reltio survivorship rules, Reltio workflows, Spotfire (1 yr.), Vaex

WORK EXPERIENCE

Senior QA Automation Engineer, Streaming Media

(October 2024 – Present)

Overview: Responsible for end-to-end QA automation and data integrity validation across multiple services, while building intelligent web scraping solutions for test data generation, API monitoring, and content verification.

Key Responsibilities:

- Designed and implemented automated test frameworks for API and UI validation using Python, Pytest, and Selenium.
- Developed robust web scraping tools with BeautifulSoup, Scrapy, and Selenium to automate extraction of structured and unstructured data for QA validation and regression testing.
- Integrated anti-scraping evasion techniques, including user-agent rotation, and CAPTCHA handling (using 2Captcha and headless browsers).
- Created reusable scraping modules to dynamically pull and validate data from third-party platforms, supporting system integration testing.
- Built scraping-based test suites for monitoring content drift, checking for broken data links, and verifying real-time pricing or listing updates.
- Employed CI/CD pipelines with GitLab and Docker to schedule and run automated scraping & QA jobs.
- Collaborated closely with developers and product owners in an agile team to ensure test coverage, scalability, and data reliability across sprints.
- Maintained QA dashboards using tools like Allure, and contributed to performance benchmarking using scraped metrics.

Tech Stack**:** Python, Pytest, Selenium, Scrapy, BeautifulSoup, 2Captcha API, Docker, GitLab CI/CD, AWS (EC2/S3), REST APIs, PostgreSQL, Allure

Data Engineer, MDM

May 2023 - October 2024 (Industry: IT Services, Digital Marketing)

Overview: AI-driven digital marketing that assists companies with creating high customer engagement by providing marketers with the ability to offer real-time, targeted, personalized customer experience in the context of the moment.

Reltio MDM Expertise:

- Configured and optimized key Reltio MDM features, including Data Loader, workflows, Integration Hub (RIH), APIs, match rules, and survivorship rules.
- Designed and managed data structures and workflows within Reltio MDM, ensuring high data integrity and performance.
- Developed and implemented connectors, such as the Databricks connector, to facilitate seamless data integration across platforms.
- Troubleshot and resolved complex MDM issues, delivering reliable and scalable solutions tailored to business needs.
- Data: Modeling, Integration, Analyses, Validation, Transcoding, Cleansing, Unification, Workflows

BI Tableau Experience Highlights for Real Estate Project

- Successfully completed a full end-to-end Reltio MDM implementation to ensure consistent and accurate master data for property datasets, significantly improving the quality of predictive analytics.
- Developed and maintained predictive algorithms for US house prices using machine learning techniques, enriched with data governed by Reltio MDM.
- Created interactive data visualizations for real estate agents and investors using Tableau, leveraging clean and reliable master data.

- Analyzed various data points on comparables for single-family homes and condos, including location, property age, and amenities, using MDM-optimized workflows.

Market Research Project

- Led another full Reltio MDM implementation to integrate and harmonize customer demographics and sales data across multiple sources.
- Utilized Tableau for comprehensive daily and historical data reporting and visualization, underpinned by MDM-enriched datasets.
- Configured match rules and survivorship rules in Reltio to ensure the integrity and accuracy of customer and product data used for analytics.
- Configured and optimized key Reltio MDM features, including Data Loader, workflows, Integration Hub (RIH), APIs, match rules, and survivorship rules.
- Designed and managed data structures and workflows within Reltio MDM, ensuring high data integrity and performance.
- Developed and implemented connectors, such as the Databricks connector, to facilitate seamless data integration across platforms.
- Implement AI image and text generation features from open-source and 3rd party APIs;
- Own and manage data generated and consumed on the front end with Database management best practices in Postgres and Databricks.
- Implement security features using, JWT, OAuth2 Authentication, SSO (Okta provider), Roles, and Permissions (RBAC)
- Perform code reviews for compliance with the best engineering practices, coding standards, and quality criteria set for the projects;
- Provide suggestions to improve the architecture, and coding practices, build/ verification toolset, and solve other technical challenges.

Technologies:

- Reltio MDM, ElasticSearch AI, LangChain, LLM, ChatGPT, Dall-E 3, HuggingFace, Stable Diffusion,
- Python, FastAPI, Asyncio, PostgreSQL, Databricks
- AWS, Lamda, GCP, Pandas, PySpark
- Postman, Docker, Git, Load Test, Locust

AI/ML, Scraping Engineer, Talents HR Platform

February 2023 - May 2023

Overview:

- **Web Scraper:** This project involved scraping job listings and resumes from external platforms. Stack: Requests, BeautifulSoup, Selenium, Scrapy. Used a pool of common user-agents and rotate them using middleware when working with Scrapy or requests-based scrapers. For simple CAPTCHAs: OCR techniques (e.g., Tesseract), for complex challenges like Google reCAPTCHA - integrated services like 2Captcha and Anti-Captcha, especially in Selenium-based scraping pipelines.
- **Enhanced Customer Engagement**: Leveraged AI technologies such as ChatGPT, Dall-E 3, Stable Diffusion, and AWS Bedrock to create personalized marketing campaigns, resulting in a 25% increase in customer engagement. Optimized backend performance for scalability and integrated robust security features (JWT, OAuth2, RBAC), improving security compliance by 30%.
- **Microservice for job resumes (profile) and Job description parser functionality** that includes integration with LinkedIn, a popular workable, glassdoor-like platform, Google Docs, PDF & Word parsers. Used R, Shiny (RStudio) with Python for predictive

analytics. Integration with LinkedIn, PDF & Word parsers. Machine Learning and text/content recognition OCR.

- **AI-Driven Face Matching Model**: Developed a model using Deep Face and Face Recognition Dlib to recognize lost children, achieving an 85% match rate. Applied predictive algorithms like KNN, SVM, Linear Regression, and G-Boost to various datasets, handling large-scale data from Kaggle for diverse applications.
- **Efficient Data Management**: Managed extensive data on the front end with PostgreSQL and Databricks. Wrote efficient Python code using FastAPI and Asyncio, ensuring seamless integration of AI models, and reduced bug rates by 15% through regular code reviews and best engineering practices.
- **System Reliability and Scalability**: Improved system architecture and coding practices, leading to a 20% increase in system reliability. Developed scalable AI solutions with AWS Lambda and GCP, enabling the digital marketing platform to handle varying workloads efficiently.
- **Real-Time Personalized Experiences**: Designed and developed backend infrastructure for AI-driven digital marketing, empowering marketers to create highly targeted and dynamic campaigns, significantly enhancing customer engagement and the overall performance of the digital marketing platform.
- **Use Case**: Develop and train deep learning models with PyTorch to forecast future sales trends and customer demand.

Technologies: AWS, Restful API, Python, Pytest, Allure, R, Shiny, JavaScript, Docker, Kubernetes, ChatGPT, Dall-E 3, Stable Diffusion, AWS Bedrock, Face Recognition Dlib, KNN (K-Nearest Neighbors), SVM (Support Vector Machine), Linear Regression, G-Boost (Gradient Boosting), JWT (JSON Web Token), OAuth2 Authentication, PyTorch, RBAC (Role-Based Access Control), Scikit-learn, FastAPI, Asyncio, PostgreSQL, Databricks, LinkedIn API, Glassdoor-like platform integration, Google Docs API, PDF parser.

Data Engineer and Data Analyst, Power BI - home equity investments

January 2023 - February 2023

Overview: Startup that revolutionizes the home equity market in the US. Our team is working on providing outstanding BI services with accessible data to decision-makers as well as streamlining the current services and their effectiveness.

Responsibilities

- Design and develop Tableau dashboards;
- Utilized Spotfire to design and implement interactive dashboards that provided real-time insights into key business metrics;
- Produce well-designed, efficient code by using the best software development practices;
- Perform code reviews for compliance with the best engineering practices, coding standards, and quality criteria set for the projects;
- Use TensorFlow to automate and optimize code reviews, ensuring compliance with best engineering practices through AI-driven code quality assessments;
- Provide suggestions to improve the architecture, and coding practices, build/verification toolset, and solve customer problems.

Technologies: Tableau, SQL, Snowflake, TensorFlow.

Data Engineer, Data Quality in Data management platform / Amazon E-Commerce Aggregator

2020-August 2022

Overview: Next-generation consumer goods company reimagining how the world's most-loved products become accessible to everyone. We use a deep understanding of rankings, ratings, and reviews to identify and acquire quality brands and use world-class expertise and data science to make their products better or create new ones to meet changing customer demand.

Responsibilities:

- Use Sisense to build dashboards for tracking updates to selected Amazon store brands for determined time periods. I used the interactive SQL palette for querying the tables to filter the needed information (columns) to be displayed in the dashboard. This dashboard provides the data engineering manager with the necessary information to make decisions on store brands.
- Create and support ELT data pipelines built on Snowflake and DBT while ensuring high-quality data
- Develop and deploy data warehousing models, and support existing processes/ETLs (extract/transform/load), and functions (in Python/SQL/DBT) in a cloud data warehouse environment using Snowflake, AWS services
- SQL statements and developing in Python
- Design and develop data pipelines (DAGs). Automation tests.

Technologies: Sisense BI, AirFlow, ETL, ElasticSearch, Snowflake, Python, SQL, DBT, Pandas, AWS S3, Medallion Architecture, MySQL, Hadoop, Spark, GitLab CI/CD, Kubernetes, LDAP, Automation Test, Pytest, Snowflake Schema, Dimensional Modeling, ER Diagrams.

AI Engineer, AI Project

September 2020 – July 2021

Responsibilities:

- Develop automation workflows with RPA (UiPath).
- Set up and manage web-based cloud services on AWS EC2.
- Utilize TensorFlow to build and deploy models that predict key business metric

Python Developer, IoT-leveraged agricultural tech company

A project on monitoring and reporting sample data from agricultural plants on a field of land.

May 2020 – August 2020

Responsibilities:

- Hands-on setting up, maintaining, and deploying services to AWS EC2.
- Automated web scraping of data from webpages (Selenium) to scrape data from sensor-related data sources, public weather sites, and agricultural platform database.
- Carried out multi-processing and parallelizing of code with PySpark.
- Used Spark for 2 cases of data processing in an ELT phase:

1. Data was collected from drones and other specialized bots were used to physically survey the land area and take samples from the soil and air for properties such as soil pH, moisture content, specific gravity, etc for different types of crops planted on the field. This data was received gotten in real-time, and placed on a queue to be loaded into AWS DynamoDB. The transformation involves converting some data properties from the queue such as temperature from degree celsius to the kelvin scale, moisture content from cubic centimeters to cubic meters, etc. The transformed data is then loaded into AWS s3.

2. Process large batch data averaging 10 million rows with spark: There were cases where I had to transform data on a different database containing historical data to consolidate the currently maintained tables in another database. The historical data contains millions of rows of IoT-generated values. To optimize speed and memory usage for transformation, I used python's implementation of Spark (Pyspark) to carry out the same transformation technique on the batch data to backfill the current table in the database.

IT Analyst, FieldworkAfrica

July 2016 – 2019

Responsibilities:

- Developed data visualizations on PowerBI and Tableau to track areas of high and low drink consumption to establish which areas are potentially viable to push a new drink to.
- Provided daily and historical data reports and visualizations to the technical director. Daily and historical reports included tracking the coverage of data collection in geographical areas, and providing updates on data quality checks and target data samples.
- Developed and maintained cloud services on the Google cloud platform.
- Developed questionnaire scripts on ODK for market research.
- Led a data collection team of 10 people.
- Performed data analysis using data tools, visualizations, and dashboards.
- Used PowerBI and Tableau to provide daily and historical data reports and
- visualizations to the technical director. Daily and historical reports included tracking the
- coverage of data collection in geographical areas, providing updates on data quality checks and
- target data samples.

Python Developer, NDA

Jan 2019 – April 2019

Responsibilities:

- Working on websites back-end with flask and Django.
- Maintaining SQL databases for proper scaling.
- Ensuring proper test units are integrated to promote clean codes.

Data Science Trainee, DATA SCIENCE

2017

Responsibilities:

- Implemented optimization algorithms.
- Carried out analytics with Microsoft Azure for prediction models.
- Generated various visualization models for data analytics with Power Bi and Seaborn.

Campus Ambassador, NDA

July 2016 - December 2016

Responsibilities:

- Promoted the ScholarX mobile App on designated campuses and social platforms for the company achieving 1000 downloads on the Google Play Store.

Engineering Intern, NDA

April 2015 - July 2015

Responsibilities:

- Assisted in a supervisory management role and design engineering in various structural steel processes.

BI Tableau Experience Highlights:

1. Real Estate Project

- Developed and maintained predictive algorithms for US house prices using machine learning techniques such as regression and classification
- Created interactive data visualizations for real estate agents and investors using Tableau
- Analyzed a variety of data points on comparables for single-family homes and condos, including location, property age, and amenities
- Assessed factors like ARV (After Repair Value), square footage, year built, number of beds and baths, garages, and local market conditions
- Developed user-friendly dashboards to display real-time market trends and property values, enabling investors to make informed decisions quickly
- Collaborated with a team of data scientists and engineers to continuously improve algorithms and visualizations

2. Tableau Specialist in Market Research Project:

- Utilized Tableau for comprehensive daily and historical data reporting and visualization to support decision-making processes
- Provided data insights and visualizations to the technical director, enabling a better understanding of market dynamics and trends
- Created a range of custom dashboards for daily and historical reports that covered various aspects such as sales, customer demographics, and product performance
- Monitored and analyzed data collection coverage in target geographical areas to ensure accurate representation of the market
- Conducted regular data quality checks, including data validation and cleaning, to maintain high data accuracy and reliability
- Collaborated with data engineers and analysts to optimize data collection methods and improve overall data quality

EDUCATION

- **College of Technology,** 2018 - 2019, Higher National Diploma (HND), Mechanical Engineering
- **College of Technology,** 2013 - 2016, National Diploma (ND), Mechanical Engineering

CERTIFICATIONS

- **Python Developer Certificate** (Sensegrass), 2020
- **Google Scholarship Android,** Basics - 2018

- **Certificate of Completion** (DSN 2nd Data Science Boot Camp), 2017
- **Certificate of Proficiency in Human Resources and Skill Acquisition,** 2014
- **Certificate of Participation ACM** (Association for Computing Machinery), 2017
- **Big Data Foundations** (Level 1), 2017
- **Data Science Foundations** (Level 1), 2016