Andrew Reinke

Data Scientist / Data Engineer / Data Analyst

Address Vancouver, Washington 98684

Phone 1-564-203-8768

E-mail ajreinke@gmail.com

LinkedIn https://www.linkedin.com/in/andrew-j-reinke

WWW https://bold.pro/my/andrew-reinke/115

WWW https://calendly.com/ajreinke

WWW https://bit.ly/4eosMET

WWW https://www.andrewreinke.com

Innovative Data Scientist and Data Engineer with a keen interest in AI and its potential to automate even the most challenging tasks. Passionate about leveraging Python to enhance data-driven processes through generative AI.

- United States Citizen

Work History

Feb 2024 - Current

Senior Lead Data Engineer

Pfizer (Commercial Analytics And Engineering), New York, NY

- Using Python I rewrote DSS Dataiku Zones which originally contained multiple recipes into a single Python script. Successfully, integrated Github Version Control into the DSS Project.
- Built and automated, using Python, several SnowPark and SnowFlake scripts for data transmation projects. Created Airflow DAGS to run the script including scheduling and debugging using Airflow's tools.
- Worked with other data engineers to successfully refactor a SAS script into Python.
- Successfully, ran SAS jobs on a RedHat server including variable modifications and SnowPark configuration changes.
- Helped optimize Marketing Return on Investment MROI projects including realigning marketing optimization efforts for Pharmacological and Vaccine brands.
- IQVIA LAADS data ingestion and analytics
- Snowflake Streamlit development.
- Tableau development, connected with Snowflake

- Usage and Revenue predictions using Meta's Prophet Python library. Visuals and predictions output to Snowflake Streamlit along with options to filter data and dynamically rebuild visuals. Later on improved model accuracy substantially by adding in features into the model.
- Wrote Python code to connect to Confluence's API and auto downloading corpus on top 100 results from a search.
- Wrote Python code to download all fields,tables,schemas,databases Meta Data from Snowflake and filtered on a keyword. Piped the results of the Confluence call and Snowflake data to ChatGPT where questions could be asked. This allowed Project Managers and Analysts a quick way to summarize Confluence and all relevant fields,tables,schemas,databases in Snowflake, thereby finding answers quickly without searching through thousands of schema and a hundred Confluence search results.

Nov 2023 - Feb 2024

Senior Data Scientist / Engineer

Nordic Consulting Group, Madison, WI

- AWS Glue Scripting, AWS Glue Notebooks, AWS Athena Notebooks: created ETL using Pyspark, SQL, Athena, Hive, Glue Catalogs, Hudi Tables
- Using Engineering Requirement Diagrams, audited maps against scripts building as needed.
- Worked with HIPAA medical data moving data from EPIC Caboodle to Hudi Tables.

Oct 2022 - Nov 2023

Senior Data Engineer

Datum Consulting Group, Washington, D.C.

- Created a LLM using RunPod.io GPU instance, with Oogabooga (webUI) and Meta's Llama2, did basic training on dataset and interacted with the LLM.
- Created Python scripts to interact with ChapGPT's API. This allowed for automatic content creation based on a dirty dataset, including prompt engineering to parse unparsable strings.
- Created AWS Kinesis Firehose and AWS Kinesis Data Streams which triggered AWS Lambda jobs to process incoming data from websites/front end.
- Developed Python scripts for AWS Lambda jobs using Docker. Created Python functions to process incoming JSON and pass the data to AWS Aurora Postgres RDS. Uploaded to AWS Elastic Container Registry (ECR) and attached to AWS Lambda Functions. Modified VPC/Security Groups/Subnets to allow for Lambda to access RDS/S3/OpenSearch.
- Wrote scripts to push data from Postgres RDS to AWS Open Search (AWS Elastic Search) using JSON payloads.
- Wrote Postgres Procedures and Functions, called from Python, which allowed multiple CRUD operations to act as a single transaction.
- Developed Python scripts on DataBricks for processing Terrabyte sized tables including developing matching algorithms to find similar records.
- Developed Pyspark scripts for file ingestion, cleaning and analytics problems. Later scaled those Pyspark scripts by creating AWS Elastic Map Reduce (AWS EMR) clusters.

- Developed EMR Workbooks (Jupyter Lab Notebooks) attached to EMR Clusters.
- Collaborated on ETL (Extract, Transform, Load) tasks, maintaining data integrity and verifying pipeline stability.
- Used Bash/Linux shell scripting and Python to design and update databases.
- Employed data cleansing methods, significantly Enhanced data quality.
- Contributed to internal activities for overall process improvements, efficiencies and innovation.
- Loaded data sets into RDS Aurora and RDS Postgres using psql called from an EC2 w/ primary key creation.
- Benchmarked performance of RDS Aurora vs RDS Postgres containing exact datasets using Postgres pgbench utility.
- Built and configured Athena Lambda Functions to connect to RDS Aurora and RDS Postgres allowing for cross database joins in SQL.
- Configured RDS Postgres and RDS Aurora to connect to EC2 instances.
- Built Python in DataGrip using SSH Python Interpreter to connect to remote EC2 instance, developing locally, saving to GitLab for CI/CD.
- Connected to AWS Athena, AWS RDS Aurora, AWS RDS Postgres using Python Libraries: PySpark, PyAthena, Psycopg2 to allow for ETL processes.
- Created IceBerg datalake in AWS Athena using Python calling Athena as well as created IceBerg Tables in Athena using Athena SparkEngine Workgroup environment.
- Converted RDS Postgres to RDS Aurora using Postgres to Aurora Read Replica migration route.
- Created VPC groups, IAM Policies.
- Created database connections in Data Grip, to RDS Aurora and RDS Postgres from DataGrip and Pycharm using SSH tunnuel on an EC2.
- Automated FTP file ingestion using Python Paramiko and private keys passing files and directories to Postgres and later dataframes which would trigger Python functions based on file name and type, later saving data to AWS EFS Elastic File System allowing for cross availability zone and mount point on any EC2 Linux in any AZ. This essentially created a unlimited NAT drive from any EC2.
- Helped designed Engineering Requirement Diagrams (ERD) using Lucid Chart.

Apr 2019 - Oct 2022

Senior Data Engineer / Lead Data Scientist

Bill.Com, San Francisco, CA

- Built API pipelines to allow for automated industry classification codes to be applied for sales leads as well as customers and vendors. This allowed for targeted sales followups based on momentum within certain industries resulting in lower marketing costs and better customer acquisition metrics.
- Extracted data from SalesForce API, and D&B Dun and Bradstreet, Zoom Info for opportunity and customer matching.
- Extracted LDA data from customers and imported into Neo4j allowing for customer and lead clustering analysis. The clustering analysis allowed for

tailored and higher precision Machine Learning models to be applied to the cluster for better customer and lead classification.

- Built API's to ingest socisal media data from Reddit and Meltwater pushing results to AWS S3 parquet files. From there, Athena tables were built from the S3 location and an AWS QuickSight Dashboard was created to filter results including using of Quicksight Calculated Fields, Filters, and Parameters to allow for ad-hoc searches on results.
- Built Python API to read from F.R.E.D. (Economic Research, Federal Reserve Bank of St. Louis Missouri) then pushed data to AWS Quicksight Dashboards for CPI comparisons, industry metrics, etc.
- Converted Python Pandas scripts to Python PySpark, reducing processing time from 6 weeks to 2 hours. This involved moving from Python Record Linkage to Pyspark's Ceja Jaro Winkler algorithm including conversion of Pandas transformations to Pyspark equivalents.
- Built Net Present Value (NPV) calculations, using Python, on each customer's revenue stream allowing us to find those industries most valuable.
- Built and lead Invoice2Go's customer to Bill.com's customer matching program. The python code (using recordlinkage library) and pipeline were automated and scaled on Sagemaker. This allowed us to de-duplicate same customers across multiple companies.
- Built and maintained an internal search engine which scanned millions of websites using Python, saving the results into Athena for quickly finding cohorts matching certain criteria.
- Built financial prediction models using StatsModel's Sarimax / (seasonal Arima) time series forecasting in Python including walk forward analysis allowing for predictions 12 months ahead with 84% accuracy.
- SageMaker Jupyter Lab notebook development including bash terminal scripting for automated library installs.
- Python script automation using AWS EC2 instance resulting in 24/7 availability of Google Sheets custom calculations integrated with Athena.
- ARIMA Forecasting Model in Python with Seasonality Smoothing and Seasonality Parameter Optimization.
- Created Mixpanel API scripts in Python to push events from datalake to Mixpanel. This allowed value added metrics to be displayed alongside and within standard and custom Mixpanel reports.
- Created Twitter API scripts in Python to pull in Twitter data for certain keywords and network diagrams.
- Created tSNE visual based on multidimensional data to visualize multiple features and their correlations. This allowed for marketing and optimization of certain processes.
- Created matching system in Python to match same customers from dissimilar datasets. After purchasing another company, customers from two separate companies needed to be matched to each other. Using Python FuzzyMatcher and RecordLinkage a pipeline was built to match same customers using a probability of match.

- Created Latent Dirichlet Allocation (LDA) and Non-Negative Matrix Factorization (NMF) for NLP Topic Modeling on 80,000 comments.
- Pushed NLP data from Athena to Slack using Slack's API Webhooks written in Python scheduled in Cron. This allowed for near Real Time Customer comment distribution to affected stake holders.
- Website reviews extracted using Python Requests Google Big Query Machine Learning Create Model: BoostedTreeClassifier, Auto_ML including using of Predict for realtime pipelining of predictions.
- GitLab use for pushing Python code to repositories for CI/CD jobs Built NLP extraction processes in Python using TextBlob to extract Noun Phrases, Parts of Speech, Sentiment, Polarity, Subjectivity, N Grams Created Machine Learning Text Classification Model in SKLearn (Multi Class Naive Bayes).
- Created Machine Learning Models in H2O.ai Driverless including Python H2O.ai run locally on laptop. Models used: Auto_ML Leaderboard, XGBoost Created Machine Learning Models using SKLearn including the use of the Pipeline which allows for transformations and splitting of data to be streamlined.
- Built pipelines from Athena to Google Sheets and back to Athena allowing Sheets to automatically update.
- Applied Machine Learning models (Logistics Regression, Random Forest Classifier) for prediction of user behavior in Python. This involved AWS Command Line (aws-cli) calls to pull datasets from AWS Athena into Python Pandas Data Frames. Predictions from models were uploaded using Python to AWS S3 allowing Athena tables to self update.
- AWS Quick Sight Dashboard development including use of Calculated Fields driven from Spice pulled from Athena Views.
- AWS Athena SQL query development / analytics including using AWS S3 command line scripting and transformation queries in Python.
- Extensive Presto SQL development.
- Heat Maps, Tree Maps, Word Clouds Driven by Data Pulled From AWS Athena.
- Converted Big Query SQL to Amazon Red Shift SQL developing extensively in Metabase with AWS Red Shift as Big Data Engine GCP Command Line Scripting Big Query Using BQ Load, GS Util including parameterized Big Query SQL and bash "code that makes code" Big Query Data Studio Development AWS QuickSight Big Data Dashboard Development For International Payments Foreign Exchange FX Dashboard.
- Created SQL scripts using Google's Advanced Window Analytic Functions Ran Sentiment Analysis and Natural Language Processing NLP on user comments' using Python w/ libraries: Pandas, Spacy, Scipy (Spatial), NLTK / Vader (SentimentIntensityAnalyzer) to reduce manual comment processing resulting in 80% reduction of labor for faster Net Promoter Score reporting.
- Created Google Data Studio and Google Sheets graphs such as GeoMaps, Scatter Plots and Bubble Charts to display and analyze data extracted.
- Foreign Exchange FX Data Analysis on an ongoing basis including Revenue Calculations, dashboarding, and ongoing spread discrepancies.
- Extensive R plotting using ggPlot and Seaborne.

- Extracted data from Google Big Query and ran Linear and Polynomial Regression analysis using R.
- Extracted raw JSON data from MixPanel and converted to Google Big Query tables using Linux Debian commands such as JQ, BQ, CURL, SED, BASH Automated ETL for A/B Web Experiments so Analysts could see results of experiments automatically without needing to preprocess their data.
- Created ETL scripts that extracted data from MixPanel (MP) using MP's API scripted in Google Cloud Shell and converted to Google Big Query tables for the Data Analytics teams.
- Developed Google Big Query SQL scripts to extract and transform data for Bill.Com's Product Growth Unit Department.
- Google Big Query API calls and Google Cloud Storage API Calls, data from Mixpanel was transferred and stored in GCP Platform.
- SQL / Linux Scripting / ETL / AWS Athena Python R/pp/p Mixpanel Rest API calls were created to transfer hundreds of GB's of user experience data from Mixpanel into Google Big Query for analysis.
- Used Inferential Boostrapping to successfully predict Profit Loss in very near term months.

Skills

Machine Learning	Very Good
Software Development	Very Good
System Architecture	Very Good
Network Security	Very Good
Server Administration	Very Good
Big Data Analytics	Very Good
System Engineering	Very Good
Data Storage and Retrieval	Very Good
Amazon Web Services	Very Good
Software Component Libraries	Very Good
RDMS Development	Very Good
API Design and Development	Very Good

Education

MBA: Business Management (1 Year)

Jun 1997 - Jun 1997 *University of Alaska Anchorage (1 Year) - Anchorage, AK*
Udemy - Python for Financial Analysis and Algorithmic Trading Python Libraries: numpy , pandas , matplotlib , quantopian for algorithmic trading with Python. Udemy - Python for Data Science and Machine Learning Udemy - SQL For Data Science With Google Big Query

Jun 1996 - Jun 1996 **Bachelor of Arts: Finance**
University of Alaska Anchorage - Anchorage, AK

Jul 2021 - Jul 2021 **Data Science: Inferential Thinking - Simulation**
University of California, Berkeley (edX) - Berkeley, CA

Aug 2021 - Aug 2021 **Data Science: Computational Thinking With Python**
University of California, Berkeley (edX) - Berkeley, CA

Sep 2021 - Sep 2021 **Data Science: Machine Learning And Predictions**
University of California, Berkeley (edX) - Berkeley, CA

Jun 2019 - Jun 2021 **NLP - Natural Language Processing With Python**
Udemy.com

Oct 2020 - Oct 2021 **Python For Data Science And Machine Learning**
Udemy.com

Jan 2021 - Jan 2021 **Data Science And Machine Learning Bootcamp With R**
Udemy.com

May 2020 - May 2021 **Python For Financial Analysis & Algorithmic Trades**
Udemy.com