# SHAWN BECKER

Lehi, UT • (857) 891-0896 • sbecker@alum.mit.edu

## MACHINE LEARNING / DATA ENGINEERING / DATA ARCHITECTURE

As a senior data engineer with growing machine learning expertise, I aim to apply my extensive data management skills to the rising field of ML and MLOps. My background includes creating scalable, secure data pipelines and robust infrastructures across industries like entertainment, healthcare, finance, and manufacturing. Leveraging AWS-based solutions, I excel in problem-solving and cross-functional collaboration, using Agile methods to deliver working increments of well-architected solutions that satisfy product stakeholders. My goal is to contribute to innovative projects combining traditional data engineering with advanced machine learning operations, to share the excitement of working with cutting-edge technologies that provide real-world value.

## WORK EXPERIENCE

**Fannie Mae / Risk Works Analysis Data Lake; Remote** **(Mar 2024 – Jun 2024)**
**Senior Data Engineer**

- Documented processes to build, test, and deploy data pipeline components for in-house ETL framework.
- Extensive work with SQL, AWS Redshift, Glue, S3, IAM, Lambda, REST, Postman, SNS, and dbt.
- Utilized Agile practices with Jira, including backlog refinements, sprint planning, daily scrums, bi-weekly sprint reviews, and end-of-sprint retrospectives. Enabled the product owner to review each shipped product increment, allowing for potential revision or re-prioritization of backlog items.

**The Cigna Group / Data Cybersecurity; Remote** **(Jun 2023 - Dec 2023)**
**Senior Data Engineer**

- Modernized apps via Jenkins CI/CD pipeline upgrade, integrating SetupTools, Artifactory/PyPI, SonarQube, and Xray.
- Investigated and implemented preparation of legacy ETL data pipeline components. Migration from on-prem Unity IoC apps to the AWS cloud using CDC.
- Ensured privacy and encryption standards for user data.
- Engineered Python REST API integration enabling credential retrieval from CyberArk's identity management platform using mutual TLS/SSL authentication via AWS API Gateway
- Initiated CyberArk service updates to extract credentials at runtime, avoiding the need to access locally encrypted files and eliminating engineering efforts to satisfy cybersecurity requirements. The cost was reduced by 95% of the original for each password rollover event.

**Warner Brothers Interactive Entertainment; Remote** **(Sep 2022 – Apr 2023)**
**Senior Data Engineer**

- Utilized PySpark for ETL processes and Python for third-party integrations and dev-ops collaborations with Jenkins, DataDog, and ZenDesk. I leveraged Google BigQuery and AWS services, including Lambda, Aurora PostgreSQL, SalesForce, Sno, and AWS Glue.
- Developed high-volume pipeline ingress and RESTful API integrations, enabling efficient game telemetry and user PII data transfer between WB-distributed games and leading marketing platforms using Segment CDP, Kafka, Redshift, Glue, and Airflow for orchestration.
- Conducted exploration and statistical analysis of marketing data, including principal component analysis, eigenvector decomposition, dimensionality reduction, vectorization, Bayesian clustering, and collaborative filtering. I used Jupyter, Python, Pandas, NumPy, Sci-kit Learn, and Keras for analytical processing and Seaborn, Plotly, and Matplotlib for data visualization.

- Practiced Agile SDLC in Jira with spring planning, daily scrums, and bi-weekly sprint reviews.

**Angel Studios; Provo, Utah** **(Dec 2021 - Aug 2022)**
**Senior Data Engineer**

- Developed a production-ready CNN using AWS SageMaker, Python, PyTorch, and Keras for classifying movie frames, enabling a new revenue opportunity through automated content tagging.
- Earned certifications in Advanced Learning Algorithms, Advanced SQL for Data Scientists, and Supervised Machine Learning: Regression & Classification, boosting professional skills.
- Conducted data exploration with machine learning algorithms using Jupyter, Python, Pandas, NumPy, Sci-kit Learn, and Keras for processing, and Seaborn, Plotly, and Matplotlib for data visualization to enhance analytical capabilities significantly.
- Created RESTful APIs to exchange data with external e-commerce and advertising partners.
- Created Snowflake ingestion scripts to support Looker OLAP and business intelligence reporting.
- Developed effective Business Intelligence strategies using Looker and Tableau with Snowflake and Redshift for comprehensive sales and finance reporting.

**Greenseed Data Laboratory; Orem, Utah** **(Nov 2020 - Nov 2021)**
**Senior Data Engineer**

- Conducted advanced statistical exploration of real-estate sales data using Python, Pandas, NumPy, SciPy, and Scikit-learn.
- Implemented a CI/CD pipeline using GitHub Actions with Coverage, SonarQube, and Xray
- Designed and built a custom star-schema data warehouse on PostgreSQL, using dimensional modeling, featuring SCD type-2 tables sharing a common streaming facts table.
- Managed RESTful APIs used to exchange data with real-estate data teams and customers.
- Enhanced machine learning skills with tutorials for TensorFlow, PyTorch, and Keras using Kaggle datasets.

**NuSkin; Provo, Utah** **(Nov 2019 - Nov 2020)**
**Senior Full Stack Developer**

- Enhanced site registration and login pages by designing workflow and wireframes.
- Innovated Vue Vuetify components with NodeJS SCSS for improved functionality.
- Documented and launched new packages for company-wide use, enhancing efficiency.
- Internationalized content using Adobe Experience Cloud.

**SeniorLink / Vela; Boston, MA** **(Mar 2017 - Nov 2019)**
**Senior Data Engineer**

- Designed and deployed an AWS data pipeline for the Vela platform, which provided messaging, communication, and collaboration tools for the healthcare industry.
- Defined RESTful APIs used by client web applications for posting daily questionnaire forms.
- Managed data ingress by queueing API Gateway-delivered message payloads into Amazon Kinesis Data Stream shards. Utilized SNS-triggered Python jobs running on serverless Lambdas to aggregate shard data into date-partitioned Parquet files in an S3 data lake.
- Performed ETL processes, extracting, transforming, and loading Parquet data from the data lake to Redshift via scheduled Databricks PySpark batch jobs on an EMR cluster orchestrated by Data Pipeline.
- Ensured privacy and encryption standards for PII, PHI, PCI, Patient Data, FHIR, and HL7, as well as HIPAA and GDPR compliance.
- Generated daily business intelligence reports using Tableau with Redshift OLAP.

**ClipFile; Newton Center, MA** **(Feb 2011 – Mar 2017)**
**Technical Lead, Co-Founder**

- Led a team to develop and launch a pioneering SaaS on the AWS platform, empowering individuals and content creators to search and share curated mindsets.
- Designed and implemented patented technology, creating a consumer-facing CMS that facilitated fuzzy matching among user-curated quotes and text fragments.
- Applied machine learning algorithms, including principal component analysis, eigenvector decomposition, dimensionality reduction, vectorization, cosine similarities, K-means, Bayesian clustering, and collaborative filtering to implement fuzzy word matching and word clustering.
- Defined and developed a RESTful API interface for the business logic layer used by the in-app presentation layer and as a service interface for external client apps.

**Sierra Vista Group; Boston, MA** **(Nov 2002 - Feb 2011)**
**Technical Lead, Co-Founder**

- Identified profitable opportunities in product development, software engineering, and data modeling and successfully negotiated budgets and project milestones with C-level management.
- Recruited high-value independent consultants specializing in DevOps, full-stack development, database administration, graphic design, user experience, and quality assurance.
- Developed and aligned comprehensive project schedules and detailed technical specifications with specific business requirements within strict budgets using the Waterfall SDLC process.
- Ensured privacy and encryption standards for PII, PHI, PCI, Patient Data, FHIR, and HL7, as well as HIPAA and GDPR compliance.
- Mitigated schedule and budget issues with client management when required.
- Architected solutions using full-internet and low-bandwidth cache-and-sync IoT techniques for intermittent network environments in enterprise systems with mobile devices.
- Managed IT strategies customized for clients in the entertainment, medical services, manufacturing, insurance, and cyber security industries, including AMI, Rowe Jukeboxes, Eleven Systems, Coca-Cola Corp. Europe, Medical Services Corp., and Intrusic Cyber Security.

## EDUCATION

**Massachusetts Institute of Technology, Cambridge, Massachusetts,**
**PhD, Media Arts & Sciences,** Machine Vision/Video Coding

**Brigham Young University, Provo, Utah,**
**MS, Computer Science,** Medical Imaging/Computer Graphics

**Brigham Young University, Provo, Utah,**
**BS, Design Engineering Technology,** CAD/CAE/CAM

## CERTIFICATIONS, PUBLICATIONS, PATENTS, WEBSITES

Certifications: https://www.linkedin.com/in/shawnbecker/details/certifications/
Publications: https://independent.academia.edu/shawnbecker
Patents: https://patents.justia.com/inventor/shawn-c-becker
LinkedIn profile: https://www.linkedin.com/in/shawnbecker
GitHub: https://github.com/sbecker11

## CURRENT WORK / WORK IN PROGRESS

AI-Powered Resume Visualization Tool: Developing a novel 2.5D resume visualization application using LangChain and OpenAI's GPT models. This project involves PDF parsing, natural language processing, and full-stack development with NodeJS.

- Project URL: http://spexture.com
- GitHub Repository: https://github.com/sbecker11/flock-of-postcards

## SKILLS and EXPERTISE

AWS Architecture • Amazon S3 • Amazon EC2 • Amazon VPC • Amazon ElastiCache • Amazon Aurora • Amazon CloudFormation • Docker • Kubernetes • Amazon ECR • Amazon ECS • Amazon EKS • Amazon Fargate • AWS Migration Service • DBT • Amazon Glue • Amazon Glue Catalog • Amazon Lambda • Amazon Step Functions • Kinesis Data Streams • Amazon SQS • Amazon SNS • Amazon Data Pipeline • Amazon EMR • Airflow • Databricks Medallion Architecture • Delta Lake • DataDog • New Relic • Amazon CloudWatch • PostgreSQL • Amazon Redshift • Amazon DynamoDB • Amazon SimpleDB • Oracle • SQL Server • MongoDB • SQL • Python • PySpark • Java • Git • REST API • CI/CD • Jenkins • GitHub Actions • GitHub • Bitbucket • Looker • Tableau • Amazon QuickSight • Amazon Sagemaker • Machine Learning • Regression • Classification • CNN • Clustering • Dimensionality Reduction • PCA • RAG • NLP • Encoding • Embedding • Cybersecurity • CyberArk • PII • PHI • PCI • HIPAA • GDPR • Quality Assurance Testing • Patient Data • FHIR • HL7 • Agile Scrum SDLC • MS Project • Visio • Office 365 • Scheduling • Budgets • Milestones • Risk Mitigation • Certified ScrumMaster • Stakeholder Management • Confluence • Jira • Presentation Skills • Communication Skills • Writing Skills • Resource Allocation • Leadership • Tutoring • Team Building • LangChain • NLP • Private GPT