Kevin Thomas Luzbetak

Software Engineer | Data Engineer | Machine Learning Engineer

📧 luzbetak5739@gmail.com | 📞 (747) 221-3264 | 🌐 https://github.com/luzbetak

📍 Agoura Hills, CA | U.S. Citizen | 25+ years in software & data systems

Summary

Distributed Systems
I bring over 25 years of experience in software and data engineering, including extensive work architecting and scaling distributed data platforms using Databricks, Delta Lake, Apache Spark, and Unity Catalog. My experience includes designing global Databricks environments across AWS and Azure, implementing performance optimization and cost-efficiency strategies, and building complex ETL and analytics pipelines leveraging Medallion Architecture principles.

Software Engineering & Machine Learning
Developing machine learning and AI-driven applications, including predictive analytics, classification systems, and retrieval-augmented generation (RAG). Hand on experience Databricks IQ, Snowflake Cortex AI, building visualizations dashboards.

Data Engineering & Modeling
Expert in developing robust data pipelines and architectures to support advanced analytics and business intelligence. Proficient in Physical, Hierarchical, Relational, Dimensional, and Canonical modeling, enabling optimized storage, analytics, and integration across complex systems.

Education

California Lutheran University
Master's Degree, Computer Science (2012 – 2014)
Focus: Machine Learning, Distributed Systems, Computer Vision, Embedded Systems

University of Phoenix
Bachelor of Science, Information Technology (2001 – 2004)
Focus: Software Development, Database Systems, Artificial Intelligence

UCLA Extension
Certificate, C++/Unix Programming (1998 – 2001)

McGill University
Certificate, System Analysis and Design (1995 – 1998)

Core Competencies

Programming Languages:
Python, PySpark, SQL, Bash, Java, C++, GoLang, FastAPI, Node.js

AI & Data Science:
Machine Learning, Retrieval-Augmented Generation (RAG), FAISS Retrieval QA, Databricks IQ, Snowflake Cortex AI, Natural Language Processing (NLP), PyTorch, TensorFlow, MLlib, Pandas, NumPy, Tableau, Data Visualization, Telemetry

Data Platforms & Databases:
Databricks Lakehouse, Delta Lake, Apache Spark, MLflow, Unity Catalog, Databricks Notebooks & Workflows, API Integrations, Medallion Architecture, MySQL, PostgreSQL, Cassandra, ScyllaDB, Apache Iceberg, Kafka, Hadoop, Excel

Data Modeling:
Conceptual, Logical, and Physical Modeling; Kimball Dimensional Modeling; Star and Snowflake Schemas; Canonical Data Models; Entity–Relationship (ER) Modeling; Data Lineage and Metadata Modeling; Master Data Management (MDM); Semantic Layer Design;

Infrastructure, DevOps & Monitoring:
Apache Airflow, Jenkins, Terraform, Docker, GitHub, Kubernetes, Grafana, Datadog, Splunk, Nagios

Professional Experience

Machine Learning & Data Engineer – Pacific-Design.com (Self Employed)
Los Angeles, CA · On-site · Apr 2025 – Present
Designed and prototyped intelligent document understanding systems powered by Retrieval-Augmented Generation (RAG) and large language models to transform unstructured business content into interactive knowledge assets. Built distributed Python pipelines for document ingestion, indexing, retrieval, and ranking, integrating heterogeneous data sources such as PDFs, databases, and internal wikis. Developed custom ranking strategies that combined semantic embeddings with metadata-driven heuristics to optimize contextual relevance. Evaluated multiple vector databases and embedding backends, balancing latency, accuracy, and cost for production-ready performance. Implemented end-to-end workflows connecting retrieval components with LLMs through REST APIs and orchestration frameworks, enabling real-time query-to-answer pipelines. Established analytics and monitoring frameworks to measure precision, recall, and latency, iterating on ranking and retrieval methods to improve overall system quality. Ensured data privacy and security by architecting deployments that maintained strict boundaries around client information. **Technologies:** Python, PySpark, C++, LangChain, FAISS, MongoDB, MariaDB, NVIDIA

Senior Data Engineer – Databricks Inc. (Contract)
San Francisco, CA · Dec 2024 – Apr 2025
Developed data-driven solutions within the Business Applications, People Systems team by integrating and consuming the Greenhouse and Workday external systems APIs to streamline hiring analytics and enhance recruitment and HR data processing. Designed automated pipelines to extract, clean, and transform recruitment and employee lifecycle data, ensuring privacy, accuracy, and consistency. Conducted anomaly detection by writing custom reports to identify irregularities in candidate applications, interview processes, and employee records. Built workflows to correct and clean data anomalies, improving data reliability for internal analytics and decision-making across People Systems and Business Applications functions.
Technologies: PySpark, SQL, Databricks Delta Lake, Medallion Architecture.

Senior Data Engineer – Disney Streaming (Contract)
Santa Monica, CA · May 2023 – Mar 2024
Developed data engineering solutions to support lifecycle marketing analytics, engagement metrics, audience insights, and campaign performance analysis. Built a Streamlit-based chat application integrated with Snowflake using Cortex AI Agent and Snowflake LLM, enabling data analysts to query the YAML-formatted data dictionary and generate dynamic SQL simply by asking questions in plain English. Utilized Snowpark and Cortex function calls to enable LLM-driven SQL generation and interactive analytics. **Technologies:** Python, Snowflake Cortex AI, Streamlit, Snowpark, PySpark, Databricks.

Senior Data Engineer – Nike Inc. (Contract)
Beaverton, OR Feb 2022 – May 2023
Developed enterprise data pipelines with Databricks Delta Lake and Medallion Architecture with notebooks, visualizations, dashboards, and data science. Worked on data migration from Snowflake to Databricks, developing dashboards to support business intelligence and analytics, and an internal data catalog featuring full-text search on schema tables, columns, and stored procedures. **Technologies**: PySpark, Databricks, Snowflake, AWS, Snowflake Data Warehousing, Analytics, SQL, Airflow DAGs, Lambda, Jenkins, Terraform, Tableau, Databricks Dashboards.

Senior Data Engineer – Brighthouse Financial (Contract)
Charlotte, NC Aug 2021 – Dec 2021
Developed solutions leveraging Azure DevOps and Ambari cluster scaling to optimize data processing efficiency. My responsibilities included debugging Spark jobs, tuning query performance, and managing Time Series backups on Azure Storage. As part of my role, I collaborated closely with Microsoft to resolve cluster issues and played a key part in the successful migration of data from on-premises systems to Azure Databricks. I utilized Azure Data Factory to design and automate data pipelines that ensured seamless integration and processing across various platforms. **Technologies:** Python, PySpark, Cloud and Cluster management: Microsoft Azure, Ambar, Spark UI.

Senior Data Engineer – Apple Inc. (Contract)
Cupertino, CA Jan 2021 – Apr 2021
Development critical data pipelines, including the Apple Pay Wallet data pipeline and the Rio CI/CD pipeline. Designing and implementing data ingestion processes, developing and testing unit tests for pipelines, deploying and provisioning infrastructure components, overseeing workload optimization and entitlement management. **Technologies:** Scala, Spark, AWS, Apache Iceberg, Apache Kafka.

Senior Data Engineer – DISQO Inc.
Glendale, CA Jul 2020 – Nov 2020
Developed and implemented predictive analytics projects that drove data-driven decisions. Key contributions: Designed real-time behavioral and attitudinal data collection using Flask APIs. Leveraged AWS services (EMR, Athena, S3) for efficient data processing. Conducted large-scale analysis with Spark and SQL. Integrated third-party APIs (Google Docs, Facebook) to enhance data flows. Managed deployment and infrastructure. **Technologies:** PySpark, Google Docs API, Facebook API, AWS Data Mesh, Glue, Athena, Redshift, EMR, GitLab CI/CD, Airflow, Terraform, Tableau.

Database Engineer III – Amazon Web Services (AWS)
East Palo Alto, CA Aug 2019 – May 2020
Designed and implement scalable, big data solutions using Redshift distributed databases. Migrating data sets from Teradata and Netezza to Amazon Redshift, ensuring data integrity, optimized performance, and minimal downtime during the transition. Conducting a proof of concept to scale Redshift clusters to handle petabytes of data, measuring performance, designing optimal data distribution, and calculating costs. petabytes of data, measuring performance, designing optimal data distribution, and calculating costs. **Technologies:** Python, PySpark, Redshift, AWS

Senior Software Engineer – Nexstar Media Group, Inc.
Playa Vista, CA Apr 2018 – Jul 2019
Software development for Lakana, a digital media and content solution platform serving the television broadcasting industry. Built web applications for TV station websites and national weather services. Designed and developed back-end microservices using RESTful API architecture. Developed predictive analytics models to analyze user engagement across streaming platforms, incorporating data such as user interactions, session logs, and subscription details. Delivered predictive analytics to forecast user behavior and trends, enabling data-driven marketing strategies and informed content strategy decision **Technologies:** Python, Java, PostgreSQL

Senior Data Engineer – Movies Anywhere (Contract)
Burbank, CA Jul 2017 – Apr 2018
Developed scalable Spark jobs on EMR clusters to efficiently manage data streams leveraging data stream-processing technologies. Designed and maintained a data lake for optimal storage and access, ensuring efficient data retrieval and processing.
Cleaned and computed data to guarantee its quality and accuracy, addressing potential inconsistencies and errors. Conducted in-depth analysis of event data to identify behavioral patterns and trends, drawing insights from this information. Applied geospatial and sessionization techniques to derive actionable insights and inform business decisions. Developed a robust data pipeline for Apple Music royalty payments. **Technologies:** Python, PySpark, Hadoop, Hive.

Senior Data Engineer – ZEFR, Inc.
Venice, CA Jun 2016 – Jul 2017
Designed and developed a high-performance search engine for YouTube's extensive video library. The search engine supported searching through billions of videos via a REST API, featuring natural language processing (NLP), part-of-speech tagging (which assigns grammatical categories such as nouns, verbs, and adjectives to words, aiding algorithms in understanding the sentence's structure and meaning). It also included BM25 ranking, search filters for likes and views, geospatial analysis, brand safety measures, content detection for copyrighted material, and ranking classifications. **Technologies:** Scala, C++, Xapian probabilistic search libraries, Cassandra, Kinesis, PySpark, Flask, SQL.

Solution Architect – DataStax, Inc.
Santa Clara, CA Oct 2015 – Jun 2016
Designed and implement solutions for Apache Cassandra integrating it with Databricks on both AWS and Microsoft Azure. Delivered proof-of-concept solutions that addressed clients' specific needs. Successfully installed and tested data ingestion and reporting systems across multiple data centers on both coasts, managing workloads, optimizing cluster scaling, and measuring performance and network latency. During peak tax season, I worked in the Intuit war room, where I played a critical role in supporting high-volume data ingestion processes to handle sudden surges in load. **Technologies:** Scala, Python, PySpark, Cassandra, Kafka, SQL, Databricks, AWS, and Microsoft Azure.

Senior Platform Engineer – MD Insider, Inc.
Santa Monica, CA Nov 2014 – Oct 2015
Developed a machine learning platform to analyze and aggregate billions of medical records, performing multi-class classification on structured patient data for churn prediction, medical diagnosis, and fraud detection using a Feedforward Neural Network. The platform provides data-driven insights into physician performance, focusing on clinical experience, quality, and cost metrics. It supports healthcare organizations and employers by optimizing provider networks, facilitating better decision-making, and ensuring transparency in matching patients with high-performing doctors based on comprehensive, risk-adjusted data. **Technologies:** Scala, Python, Machine Learning, Databricks, AWS.

Staff Software Engineer – Skyworks Solutions, Inc.
Moorpark, CA Aug 2013 – Nov 2014
Developed and maintained automated quality measurement systems for semiconductor manufacturing process control, specifically for RF chips used in Apple iPhones. I implemented a Six Sigma-driven analytics system to monitor processes and ensure compliance with Process Capability (CPK) standards, incorporating data science and statistical process control (SPC) techniques into the workflow. This enabled detection of process variations, elimination of defects, and ensured high product quality through real-time analysis. I also developed systems to collect and analyze production data, generating real-time quality reports. **Technologies:** Python, Java, Oracle for data modeling, and data warehousing.

Senior Software Engineer – AT&T Interactive Yellow Pages
Glendale, CA 2010 – 2013
As part of the AI team, I contributed to developing classification models for business listings in Yellow Pages using natural language processing (NLP) and. Specifically, I focused on building models that could categorize businesses based on textual data, which involved processing large datasets and automating workflows by leveraging decision trees for structured decision-making. To further enhance the model's performance, I utilized Random Forests - a technique that combines multiple decision trees to improve generalization and accuracy. Additionally, I applied general feedforward neural networks for feature processing and category prediction. T
Technologies: Java, Perl, Hadoop

Lead Search Engine Developer – Spark Networks
Beverly Hills, CA 2007 – 2010
Designed and developed a semantic search engine for Spark Networks' dating platforms, including JDate.com, using natural language processing (NLP) to classify and match profiles based on attributes such as eye color, hair, family status, location, income, interests, and other relevant factors. The engine incorporated syntactic parsing to analyze adjacent words and phrases, greatly improving search relevance. Developed system to detect fraudulent profiles, spam messages. **Technologies:** Xapian open-source probabilistic search library, BM25 algorithm ranking, decision trees classification, Naive Bayes.

Lead Search Engine Developer – MySpace.com
Beverly Hills, CA 2005 – 2007
Developed the original MySpace Video search engine on Linux using the Xapian open-source probabilistic search library during a period of rapid data growth and heavy user demand. Scaled the system across multiple servers with distributed indexing and load balancing, enabling queries on over 500 million records in under 200 milliseconds. Optimized query execution, caching, and indexing to support real-time ingestion of new videos. Designed and implemented spam filtering and anti-spamdexing systems to protect search quality from keyword stuffing and duplicate content. Applied NLP and SVM-based models for content classification and adversarial behavior detection. Built automated monitoring and anomaly detection workflows to identify abnormal query patterns and spam campaigns. **Technologies:** C++, Linux, Xapian, SVM, NLP, distributed systems, load balancing, spam detection.

Software Developer III at Countrywide Financial
Agoura Hills, CA 2003 – 2005
Designed and developed artificial intelligence software to ensure regulatory compliance and detect issues in loan underwriting, with a focus on preventing fraudulent activities and accurately verifying income reports. Leveraging automated rule-based systems, AI-driven engines enforced pre-defined compliance rules and regulations, thereby guaranteeing that loans met all legal requirements, proper documentation was submitted, and income calculations conformed to specific guidelines. **Technologies:** Java, C#, MSSQL

Computing Analyst at California Institute of Technology
Pasadena, CA 2001 – 2003
Worked at the Andersen Lab for medical research, collaborating with postdoctoral researchers to understand how behaviors are influenced by intentions. Developed software to record the electrical activity of nerve cells in the posterior parietal cortex of paralyzed patients. This brain region, located between the sensory and motor areas of the cerebral cortex, acts as a bridge from sensation to action and is crucial for planning both eye and arm movements. Created algorithms to interpret neural signals, translating patients' intentions into electrical control signals to operate external devices like robotic arms and computers. **Technologies:** C++, Linux Inter-Process Communication (IPC), shared memory, Matlab. LabVIEW.

Senior Software Engineer at Digital Insight
Calabasas, CA 1999 - 2001
Programming Internet Banking infrastructure and core components including real-time data transformations, synchronization with financial institutions via OFX (Open Financial Exchange) and Symitar interfaces, as well as online analytical processing for credit card transactions. **Technologies:** C++ on Informix IBM RS/6000 AIX Unix, and Java, leveraging Oracle databases through PL/SQL, remote procedures, and triggers.

Programmer C at SCI Systems, Inc. (SCI)
Huntsville, AL 1995 - 1999
Developed statistical process control (SPC) and statistical quality control (SQC) software to analyze and minimize defects in circuit board manufacturing. This work enabled the production of high-end electronics, including Apple circuit boards, Silicon Graphics systems, Matrox video cards, and other critical components. During my tenure, SCI Systems Corporation emerged as the largest electronics manufacturer, earning a reputation for exceptional quality and timely delivery. **Technologies:** C programming language, UNIX, IBM AIX.