Oleksandr Kazakov

Montreal, Canada
+1 514 569 5789
me.oleksandr.kazakov@gmail.com
https://www.linkedin.com/in/oleksandrkazakov/

Summary

AI Engineer with 7+ years of experience building large-scale data and machine learning systems, with recent specialization in **GenAI, RAG pipelines, and agentic AI systems**. Proven track record of deploying **production LLM applications** on AWS, designing **multi-agent workflows**, and processing **petabyte-scale datasets**. Strong focus on **evaluation, reliability, and real-world AI deployment**.

Core Skills

Agentic AI: Multi-agent orchestration (LangChain/LangGraph, CrewAI, Pydantic AI), Bedrock AgentCore, function calling, tool-use design, failure-mode handling, human-in-the-loop integrations

LLMs & RAG: Anthropic, OpenAI, Llama, Mistral (fine-tuning, inference optimization); prompt engineering, hybrid retrieval (OpenSearch/pgvector), grounding,ranking, guardrails (Bedrock); Ollama, llama.cpp, vLLM, SLM's tuning

Evals & Reliability: LangSmith, LLM-as-judge, golden datasets, CI/CD eval gating

MCP: Designed and deployed internal Model Context Protocol servers with auth, scoped tools, streaming, and schema validation

Platform: Python (FastAPI, async), AWS (Bedrock, HealthOmics, Lambda, Step Functions, Redshift, S3, DynamoDB), Airflow, Docker, Kubernetes, GCP

Data & ML Foundations: Large-scale ETL, data modeling, Spark, Hadoop, PyTorch, scikit-learn, transformers

Work Experience

Provectus — **Remote**
Solution Architect / Senior Data Engineer — **2022/09 - present**

- Led design and deployment of **production-grade GenAI systems**, including **multi-agent workflows** (LangChain, CrewAI, AWS Bedrock) for enterprise applications
- Built **large-scale RAG pipelines** processing **12M+ medical records/day**, integrating OCR, classification, and retrieval for clinical data access
- Designed **LLM-powered analytics systems** for performance tracking and cost optimization, reducing infrastructure costs by **45%**
- Developed **Model Context Protocol (MCP) servers** with authentication, streaming, and structured tool interfaces for text2sql multi-step processing flows
- Developed an applet for large volume file search on DNAnexus and their transition to AWS S3
- Orchestrated Omics pipeline execution on AWS with Apache Airflow (metastore in DynamoDB, SQS/SNS queues, Lambda triggers) with extensive batching logic
- Data Quality for empty VCF schema, files ingestion, and general quality control
- Developed a Service for BAM/VCF files discovery within the company data assets for sales and delivery teams
- Developed QuickSight and Datadog dashboards for Data Quality monitoring of the pipeline runs and lineage tracking

Kinner Inc (Contract) — Remote
Data Engineer / Data Scientist — 2020/05 - 2022/09

- ETL development - scripting in Python/R, setting up workflows in Airflow (Python), aggregation to/from MySQL, NoSQL, Spark, and Hadoop in AdTech
- Developing reporting dashboards and alerting systems integrated in different channels (Tableau, Python, Slack)
- Analytics of partners KPI's, fraud detection, user retention, and ROI forecasting (MySQL, Hive, Hadoop, R, Python, scikit-learn, AWS). Automated Data Scraping using Selenium on Python

Gazelle AI On-Site (Montreal)
Machine Learning Engineer 2020/01 - 2020/05

- Developed text preprocessing frameworks for supervised classification of NAICS codes with NLP models based of BERT transformers (PyTorch, Dask, MongoDB, S3)
- Porting and optimizing proprietary classification ML algorithms from MATLAB into Python
- Data Lake and pipeline development with AWS Athena/Glue (Python, MongoDB, S3)
- Migration of the services from MongoDB to DynamoDB and introducing Airflow as orchestration

SSENSE On-Site (Montreal)
Senior Data Engineer 2018/03 - 2018/08

- Batch ETL pipelines development in Apache Airflow (Python, AWS EMR (Spark, Hive), S3, Redshift)
- Developed GDPR pipeline service to anonymize/remove user data from AWS S3 buckets and various databases (Python, Athena, DynamoDB, Postgres, Redshift)
- Integrating AWS Glue/Athena to replace the existing ETL pipeline

Bandsintown Inc. On-Site (Montreal)
Back-End Developer 2017/10- 2018/03

- Tuning the Hadoop cluster. Adding and configuring new services such as Apache Spark, Hive on Spark. Adding cluster security levels (Bash, Hive, Spark).
- Custom MapReduce writing and integration with the pipeline (Java, Oozie).
- R&D for AWS Glue and Athena usage with current infrastructure (PySpark, Zeppelin).

Concordia University On-Site (Montreal)
Course Instructor (part-time) 2018/12 - 2020/05

- CEBD 1160 "Introduction to Big Data Technology Course" class material development to cover concepts of Big Data (Hadoop, Spark), basic data wrangling in R and Python, workflow automation, databases and designs, and cloud solutions
- Preparing and grading students' homework
- Assisting with final project development and homework/class materials
- Lead the change to the course structure to split it into separate modules

Wajam On-Site (Montreal)
Big Data & BI Developer (BI Team) 2015/10- 2017/09

- Development of the data collection, processing, and visualisation pipelines for the current and new products of the company (Scribe, Scala, Python, BASH, Apache Pig, Hive, Sqoop, MySQL, Tableau)
- Administering 60-node Hadoop cluster via Cloudera CDH 5.x suite, receiving more than 100 GB of raw data hourly
- Improving and developing Apache Pig scripts for data preprocessing of the existing pipeline
- Developed Business KPI's forecasting and alerting system integrated with Slack (Hadoop, MySQL, R, Slack)
- Prepared Tableau Dashboards and ad hoc reports for the upper management
- Integrated data "streaming" and visualization from HDFS to Tableau using Hive
- Optimized Hive performance and query execution time. Developed Java UDFs and UDAF's

Education

University at Albany, State University of New York Sept 2010 - Dec 2017
PhD Physics
Dissertation: *Application of Stochastic Liouville Equation for Nuclear Magnetic Resonance Line Shape Calculation (SpinAl software). Application of Supervised Machine Learning Techniques to Raman Spectroscopic Data.*

University at Albany, State University of New York Sept 2010 - Oct 2015
M.S. Physics
Computational Methods, Quantum Mechanics, Machine Learning, Bayesian Inference, Statistical Modelling, Probability Theory

V.N. Karazin Kharkiv National University Sept 2006 - Jun 2010
BSc. Physics Overall GPA 3.6/4.0
Theoretical Physics, Discrete Mathematics, Linear Algebra, Mathematical Analysis, Algorithms, Distributed Computing

Selected Publications

- (2020) Keith Earl, Oleksandr Kazakov. The Spin Echo, Entropy, and Experimental Design https://doi.org/10.3390/proceedings2019033034
- (2016) Elena Ryzhikova, Vitali Sikirzhytski, Oleksandr Kazakov, Lenka Halamkova, Joseph Quinn, Earl A.Zimmerman and Igor K. Lednev. **Raman spectroscopy of Cerebrospinal Fluid for Alzheimer's disease diagnosis** Journal of Biophotonics
- (2014) Elena Ryzhikova, Oleksandr Kazakov, Lenka Halamkova, Dzintra Celmins, Paula Malone, Eric Molho, Earl A. Zimmerman and Igor K. Lednev. **Raman spectroscopy of blood serum for Alzheimer's disease diagnostics: specificity relative to other types of dementia.** Journal of Biophotonics, 1864-0648
- (2014) Igor K. Lednev , Elena Ryzhikova, Oleksandr Kazakov, Lenka Halamkova, Dzintra Celmins, Paula Malone, Eric Molho, Earl A. Zimmerman. **Raman Spectroscopy of Blood for Alzheimer's Disease Diagnostics**. In Proceedings of Annals of Neurology, 76, 94