Open to B2B engagements · Lisbon, EU

Oleksii (Alex) Kotliar

Staff AI Infrastructure Engineer

I build AI infrastructure — **retrieval, agent memory and evals** — on top of distributed data systems (Python, Spark, Kafka, JVM). 13 years of experience, independent contractor, overlapping with US and EU hours.

me@oleksiikotliar.dev · linkedin.com/in/oleksii-kotliar · github.com/zfrgt

HIGHLIGHTS

1B/day ad events through a Scala/Spark/Python/Go data platform I led — in production for 5+ years.	**2.5M req/s** throughput on a high-performance Scala/Spark data platform I contributed to.	**−50%** storage cost on several of Lyft's petabyte-scale datasets, with integrity and granularity preserved.
−75% resource usage after rebuilding a low-latency processing service on Akka/Pekko.	**−80% MTTR** after redesigning incident response while managing a team of 7 engineers.	

EXPERIENCE

Staff Software Engineer · Confidential client — May 2025 - Present

Remote · full-time May–Aug 2025, B2B contractor since Aug 2025

- Designed and built retrieval (RAG) infrastructure for production AI agents, covering ingestion, indexing, retrieval and reranking.
- Built agent memory components, including relevance modelling that strengthens and decays over time.
- Built evaluation pipelines for retrieval quality using synthetic golden datasets and LLM-as-a-judge.
- Contributed to security reviews, mentored teammates and drove cross-team architectural alignment.
- Earlier in the engagement: data pipelines and AI-driven data enrichment.

Engineering Manager · Nexla — Apr 2024 - May 2025

Lisbon, Portugal (remote)

- Managed a team of 7 engineers building and maintaining a high-performance data ingestion/delivery system.
- Established a culture of transparency and trust that measurably improved team morale and productivity.
- Designed and rolled out a new incident-response process that cut response time by over 80%.

Senior Software Engineer · Nexla — Dec 2022 - Apr 2024

Lisbon, Portugal (remote)

- Designed and built a highly available, fault-tolerant data processing pipeline on Apache Spark using Nexla's core platform and modern cloud infrastructure.
- Built a low-latency data processing service on Pekko, cutting resource usage by at least 75%.

Data Engineer, L6 (Senior/Staff) · Lyft — Apr 2022 - Dec 2022

Lisbon, Portugal (remote)

- Led a team of engineers delivering datasets used by roughly half of the company's data science teams.

- Reduced storage cost and cut several datasets' size by 50% while preserving data integrity, relevance and granularity across the petabyte-scale storage layer.
- Stack: Python, SQL, Apache Airflow, Boto3, AWS.

Senior Software Engineer · Sigma Software

May 2018 - Apr 2022

Odesa, Ukraine · remote from the EU since early 2022

- Contributed to high-throughput data platforms in Scala/Spark (~2.5M req/s) and an AdTech ETL pipeline on Hadoop (Spark, Kafka, MemSQL).
- Led the Clean Data Platform (Oct 2019 – Nov 2021): Scala/Python/Go on Spark, processing up to 1B ad events/day with ELK as a low-cost data lake for ad-hoc analysis.

Software Engineer · Sigma Software

Apr 2013 - May 2018

Odesa, Ukraine

- Migrated a major customer to the next product version, owning migration issues end-to-end.
- Contributed to "Gaming Platform", a framework handling game-client communication, transactions, logging, sessions, game history and portal/widget delivery, letting game studios focus on game logic (Java, OSGi/Apache Felix, Apache CXF, Spring, Hibernate, Akka; BDB, H2, Oracle).

WRITING

Nexla Spark ETL On Databricks · Nexla blog — article

How Nexla's Spark ETL integration with Databricks works: sources and transformations, cluster lifecycle, and Delta table output.

SKILLS

AI / LLM infrastructure	RAG, Reranking, Agentic memory, LLM evals, Vector search, Embedding models, Image processing & OCR
Languages	Python, SQL, Java, Scala, Groovy
Data & streaming	Apache Spark, Kafka, Flink, Airflow, Presto/Trino, Akka/Pekko, Beam
Storage & databases	PostgreSQL, DynamoDB, ClickHouse, Vertica, MySQL, BDB/H2/Oracle
Cloud & infra	AWS, GCP, Hetzner, Kubernetes, ECS/EKS, ArgoCD, Docker
Tooling & delivery	Git, GitHub Actions, Jenkins, Poetry, Boto3, Pandas, NumPy, Shell, Claude Code
Leadership	Engineering management, Mentorship, Code & design review, Process improvement, Cross-team technical leadership

EDUCATION

MSc in Computer Science, Odesa National Polytechnic University

2009 – 2014 · Odesa, Ukraine

Thesis: OCR with image pre-processing, segmentation and stroke-width transform.

Degree recognised in Portugal (DGES, 2024).

LANGUAGES

English	C2
Ukrainian	Native
Russian	Native
Portuguese	B1