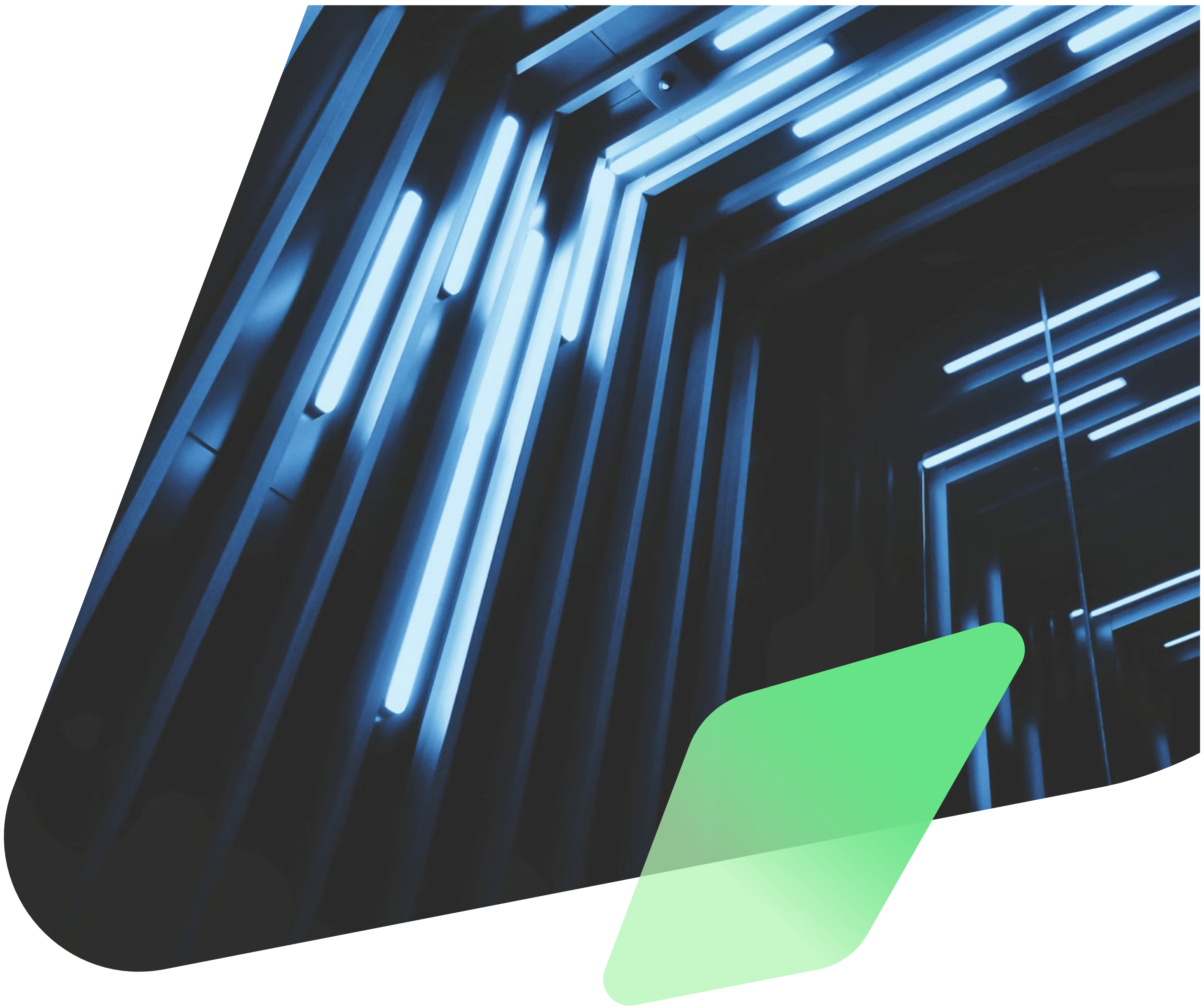

Success story

Entity resolution ML pipeline

Our client, a leading provider of B2B intelligence platforms, equips its customers with comprehensive company and contact information. This enables sales, marketing, and human resources teams to engage effectively with their targeted leads.

They faced challenges in managing large datasets efficiently and called on VirtusLab's expertise to optimize and streamline their data processing. This collaboration significantly improved the accuracy of their data operations.

The Challenge

The leading provider of business intelligence encountered substantial difficulties in effectively handling its vast datasets.

- **Entity matching at scale:** The organization processes vast volumes of data from diverse sources daily. Discrepancies in data representations for the same entity often result in redundant or mismatched information. Traditional each-to-each matching methods became computationally unfeasible.
- **Limited in-house ML knowledge:** The client's team initially had limited experience with machine learning engineering, specifically in the productionization and management of custom AI-driven solutions. This gap created a barrier to developing and maintaining tools to process data and deliver said solutions.

At this point, our client reached out to VirtusLab Group to implement a scalable and precise entity resolution system, aimed at decreasing processing loads and enhancing matching accuracy.

The solution

VirtusLab Group provided a customized solution to meet the entity resolution requirements, improving operations for the client's customers.

VirtusLab engineered a two-step entity resolution pipeline to optimize data accuracy while reducing computational overhead.

- **Two-step matching with embedding vectorization and locality-sensitive hashing (LSH):** By employing multi-level embeddings, the system vectorizes data for high-density encoding, while LSH-based partitioning segments the data into manageable clusters. This approach allows the system to evaluate only the most likely matches within each partition, preserving accuracy while reducing computational demands.
- **Scalable Spark architecture:** Built on Apache Spark, the entity resolution pipeline scales easily to accommodate new data sources. This flexible setup enables the client to streamline entity matching, quickly trace data origins, and integrate new data channels with minimal configuration.

The results

With VirtusLab's guidance, the client achieved transformative improvements in data management.

- Introduced scalable, precise entity embedding and matching with LSH
- Developed a Spark pipeline that enables effortless scalability.
- Introduced coarse-grained partition calculations to optimize performance.
- Performed high-precision entity similarity scoring only on candidates with a higher probability of generating a match.
- Significantly reduced the time required to perform a full search for potential matches in incoming data without compromising result quality.
- Enabled quick configuration of new sources for incoming data.
- Facilitated the tracing of sources for incoming entity candidates.
- Simplified the process of reproducing entity-matching workflows.

The tech stack

/ **Platforms:**	amazon EMR	amazon S3	Apache Airflow
/ **Tools & Frameworks:**	Apache Spark	Pinecone	
/ **Libraries:**	Spark MLlib	Apache orc	Hugging Face
/ **Embeddings:**	● BPEmb ● Flair	● Sentence Transformers ● Snowflake Arctic	

About VirtusLab

At VirtusLab, we aim to lead in software technology, working consistently to enhance efficiency. Our profound commitment to research and development and a dedicated focus on emerging trends and inspirations fuels an innovative culture. This ethos precisely guides advancing our cutting-edge solutions, inviting collaboration to expand the boundaries of software technology collectively. We welcome you to be a part of this transformative journey.

Let’s connect

Contact Details

info@virtuslab.com

/ POLAND

Kraków Headquarters

Virtus Lab Sp. z o.o.
ul. Szlak 49
31-153 Kraków

/ GERMANY

Berlin Office

+49 30 52014256

VirtusLab GmbH
Potsdamer Platz 10
10785 Berlin

/ UNITED KINGDOM

London Office

+44 (0)20 4577 1051

Virtuslab Ltd.
40 Bank Street HQ3
London E14 5NR

WWW.VIRTUSLAB.COM

VIRTUSLAB