From web scraping to intelligent apps in Azure

Sergei Häyrynen, COO at Veracell, friend of Azure
21.8.2024

We develop data-driven services.

We develop applications that enable full utilization of data. We know how to avoid pitfalls along the way and master the technologies that guarantee an efficient and secure end result.

Our strength is in untangling data and solving challenging problems with strategically designed artificial intelligence.

Some of our clients

About me

Role

Founder and COO of Veracell. I design and lead new AI and data projects focusing on Azure services. Stack includes Azure, full-stack development, data engineering, AI and predictive models.

Background

Studied computational biology at TUT and TUNI. Acamedic background in cancer genomics and computational biology.
Started consultant career at a CRO with research broups and medical companies as customers.
Switched to IT, the "industry", 7 years ago with founding of Veracell

veracell

Why Azure?

Full-fledged data science proofs-of-concept

Data science PoCs are in risk of being proofs of ability and missing a concept. We aim to design and implement complete concepts with cloud-first approach.

Efficiency through uniformity and best practices

Azure compatible tooling and approaches, DevOps practices, version control help us being efficient and speed up transition to production.

Ready-to-use AI services

Azure provides OpenAI and other AI services and simplifies billing and controlling resources.

veracell

Case examples

AI-driven insights from scraped data

We used generative AI to process large volumes of unstructured data from Internet pages for sentiment analysis.

Data enrichment and intelligent recommendations

We are creating an enriched company service catalog and building a recommendation engine in Azure for Tampere city.

Document structuring

GenAI provides a shortcut for automated processing of documents and other unstructured data.

veracell

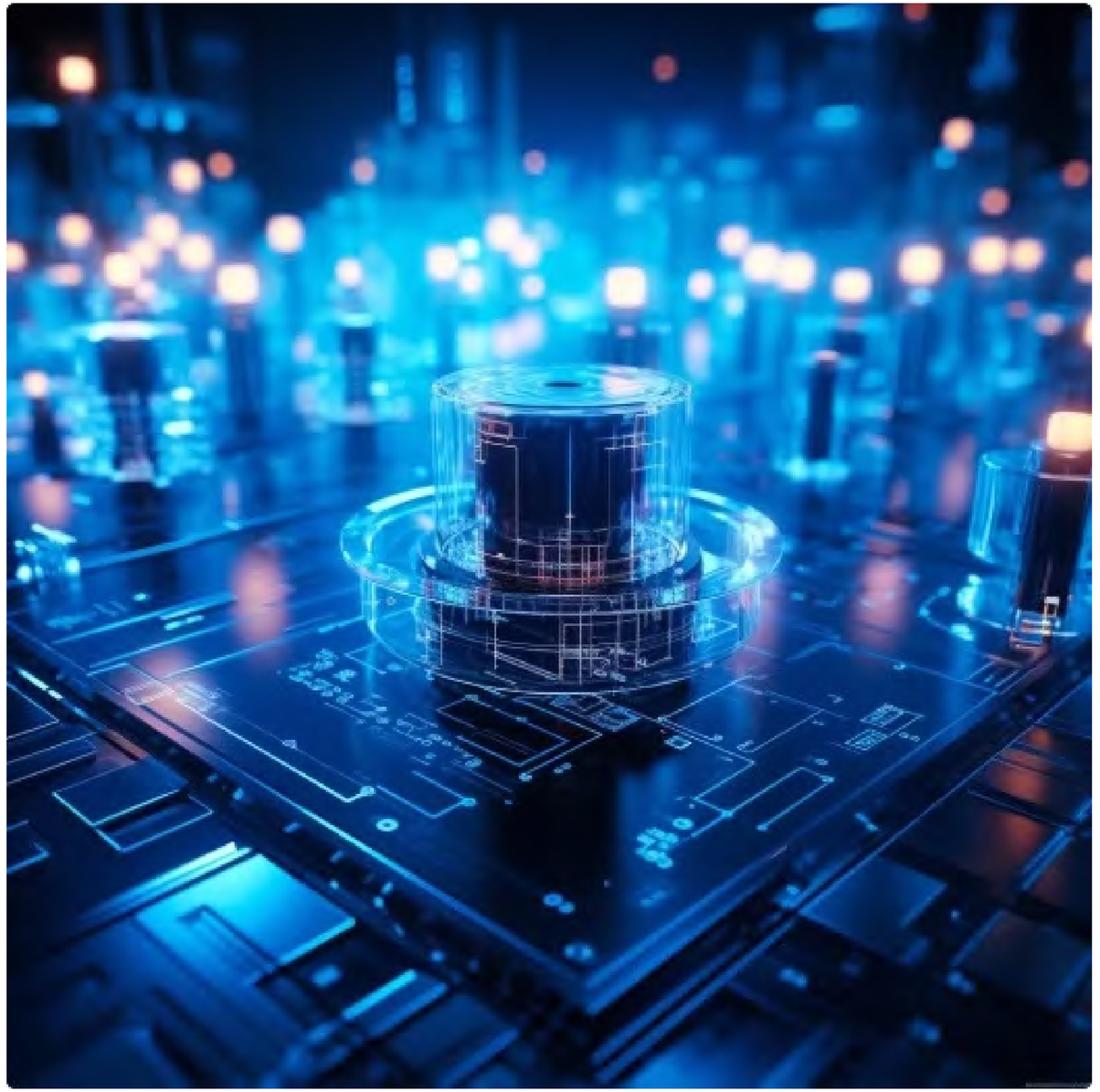
AI-driven insights from scraped data

Customer need

Customer was interested in expanding market research into online reviews and discussions not available through APIs. They wanted to capture discussed subjects and sentiment in an unobserved setting.

Hypothesis

Combining traditional web scraping tool kit with generative AI allows shortcuts in processing large volumes of data.
Generative AI is suitable for quantifying qualitative data for downstream analyses.
Building solution in Azure enables repeatability and integration in customer's processes.



veracell

Core components

Gen AI experiments and pipelines were done directly in Azure ML.

We used different models deployed in Azure AI for tasks of different difficulty - starting always from the best to see what's the best possible result.

Production use cases will be moved to prompt flow for communication and evaluation purposes. Prompt flow allows non-technical users to contribute to development and evaluation.

veracell

Results

We processed ~6000 text documents - review, discussion forum messages, articles - and generated over 15 000 data points for different attributes and their sentiments.

Azure implementation allows extending, repeating and parametrizing data collection and analyses. Solution is applicable to building datasets to different questions and analyses.

veracell

Data enrichment and intelligent recommendations

Customer need

City of Tampere wants to streamline matching companies and available services. In addition to existing data assets, they need to create a service catalog with enriched metadata. Application should recommend services based on company's information and user's input.

Hypothesis

Gen AI allows creating metadata and summarizing scraped web pages.
Integrations to Tampere city's CRM and company clustering results help to identify service need patterns.
Gen AI created metadata together with textual descriptions can be used to recommend services based on company metadata and user's textual input.

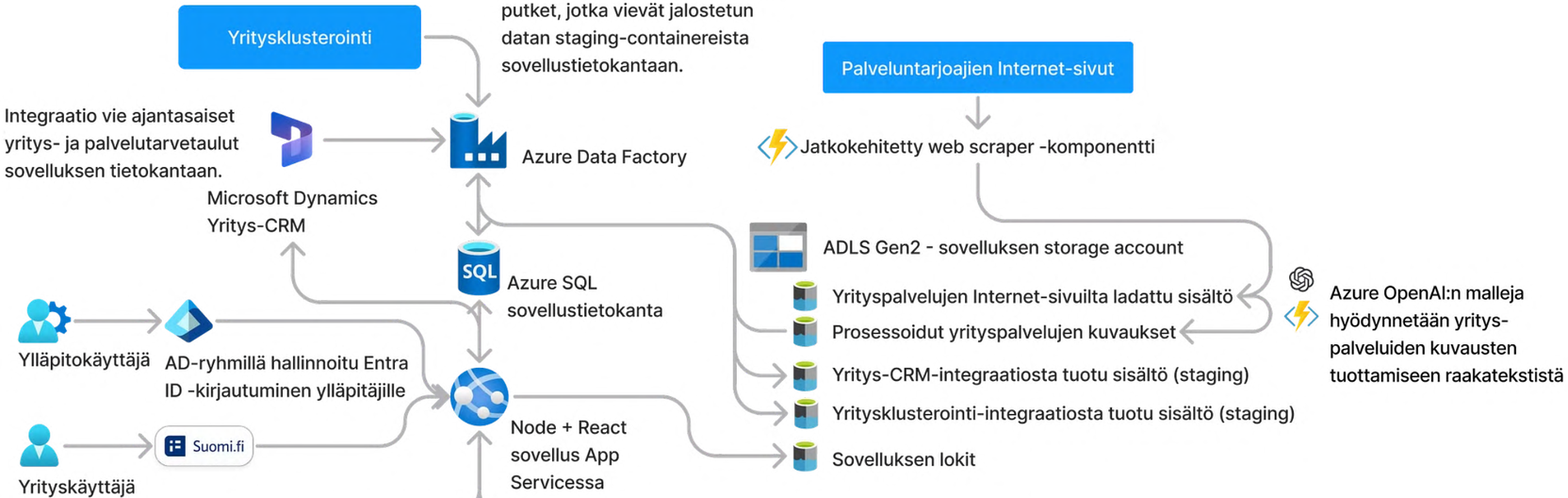
TAMPERE veracell

Architecture

Kokonaisarkkitehtuuri

Data Factoryn integraatioputket vievät integraation raakadatan ADLS Gen2 -storageen staging-containereihin.

Data Factoryyn toteutetaan myös putket, jotka vievät jalostetun datan staging-containereista sovellustietokantaan.



Sovelluksen rahoituksen AI-assistenttia laajennetaan projektissa tuotetuilla yrityspalvelukuvauksilla ja palveluiden tiedoilla sekä lisäämällä keskustelun ja lisäkysymykset mahdollistava chat-ominaisuus.

AI-assistentin jatkokehitys mahdollistaa pilotin tulosten hyödyntämisen osana yrityspalvelua ja sopivien yrityspalvelujen hakemisen keskustellen. Suurin osa laajentamiseen vaadittavasta työstä tehdään osana yrityspalvelusuottelijan kehitystä.

Kahdensuuntainen integraatio lähettää elinvoimatestin vastaukset elinvoimabarometriin ja palauttaa sovellukseen tulosten analyysin tulokset.

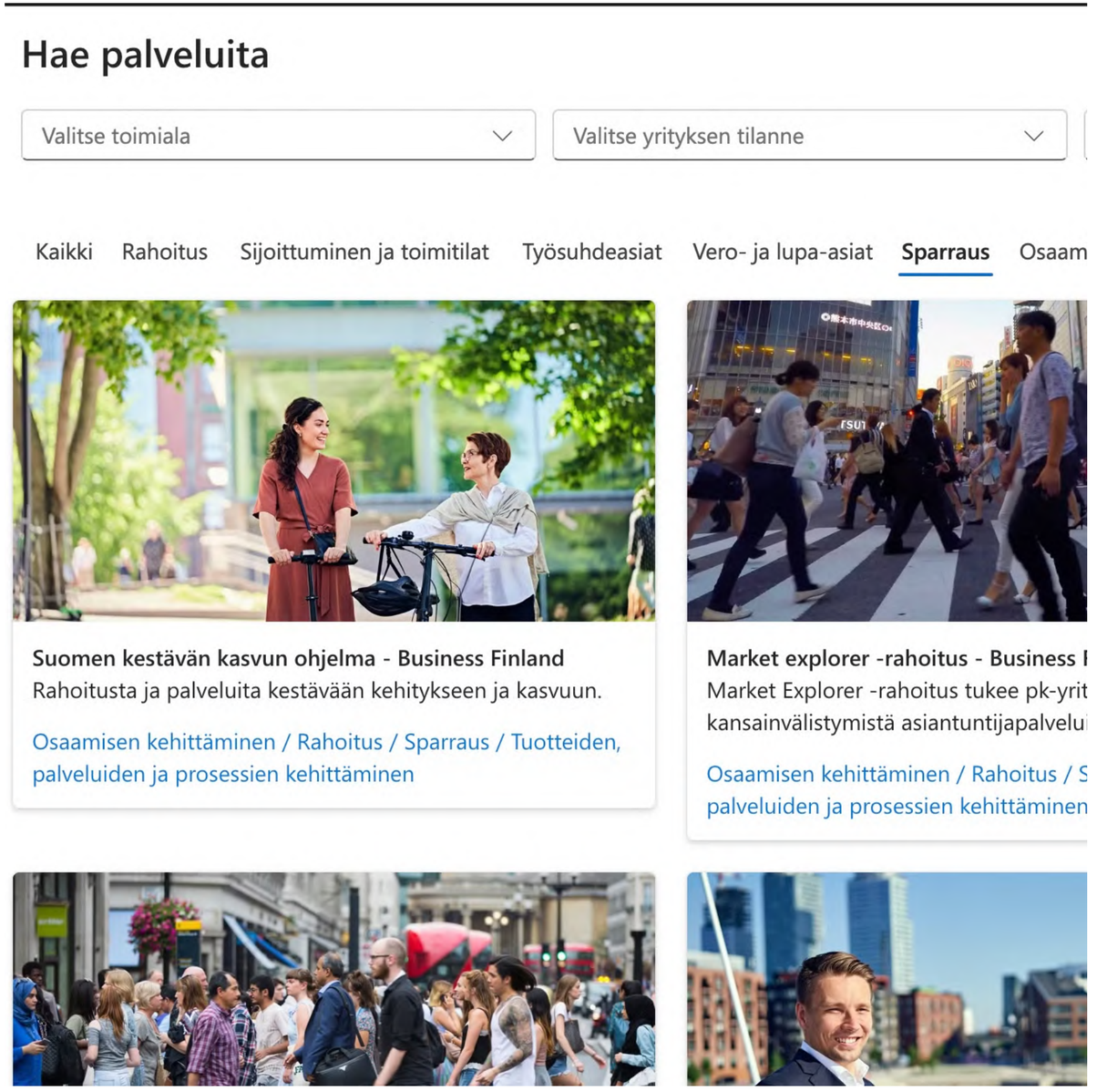
Yrityspalvelusuosittelija toteutetaan erillisenä mikropalveluna. Toteutus on Azure Function -pohjainen rajapinta.

veracell

Results this far

We built the infrastructure and data processing framework for building company service catalog with enriched metadata.

We are moving to integrating existing data assets and experimenting with recommendation approaches based on structured metadata and company description embeddings.



veracell

Document structuring

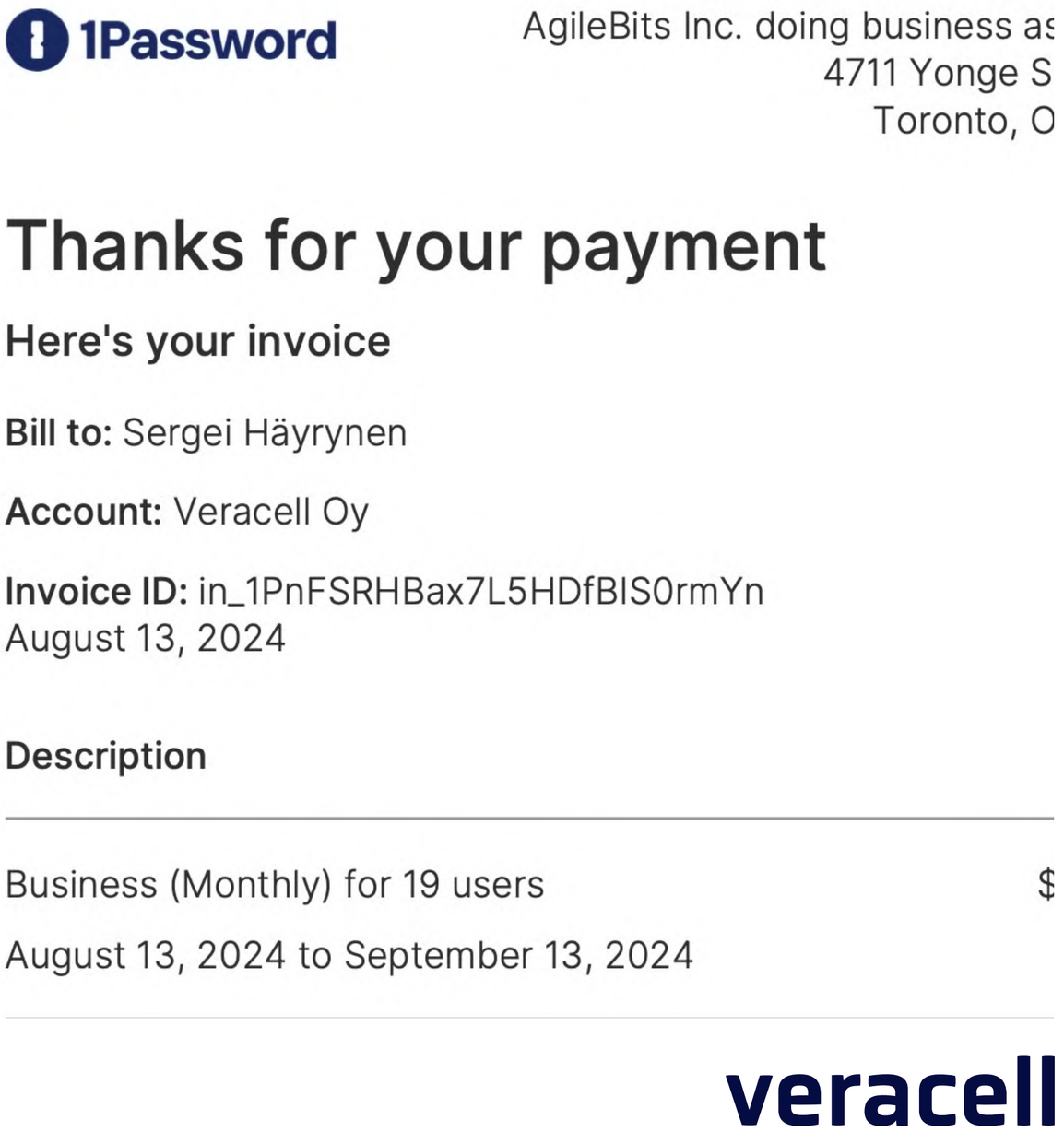
Customer need

Companies process large volumes of documents manually - either themselves or using outsourcing. They are looking to automate the processing of documents. Available solutions are either not accurate enough, suitable for integrating to their processes or too expensive to implement.

Hypothesis

Document volumes of individual Finnish companies are too small to train dedicated neural network models for structuring.

OCR combined with latest GPT models, post-processing heuristics and handling uncertainty provide a shortcut to cost-efficient automation.


1Password

AgileBits Inc. doing business as
4711 Yonge S
Toronto, O

Thanks for your payment

Here's your invoice

Bill to: Sergei Häyrynen

Account: Veracell Oy

Invoice ID: in_1PnFSRHBax7L5HDfBIS0rmYn
August 13, 2024

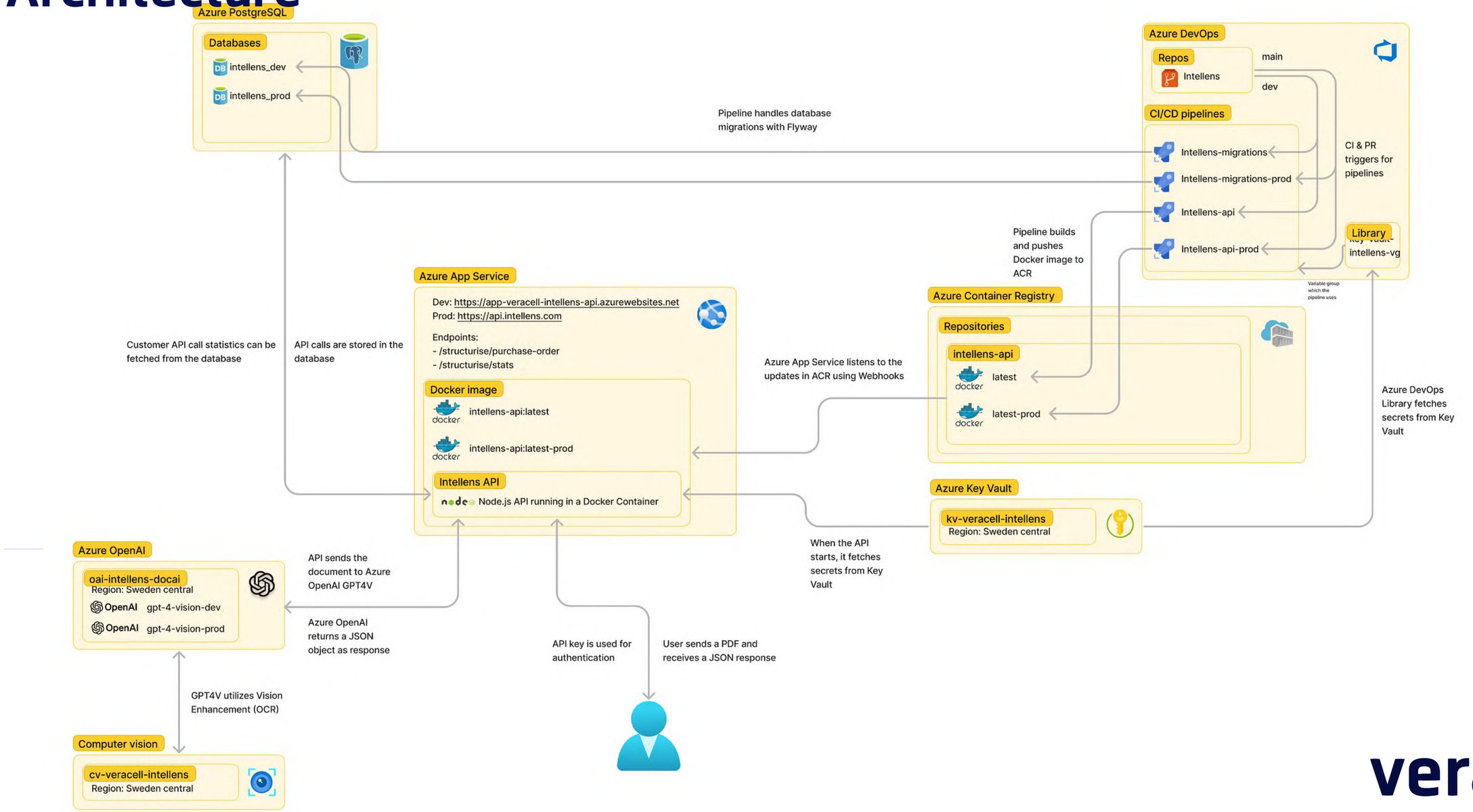
Description

Business (Monthly) for 19 users $

August 13, 2024 to September 13, 2024

veracell

Architecture



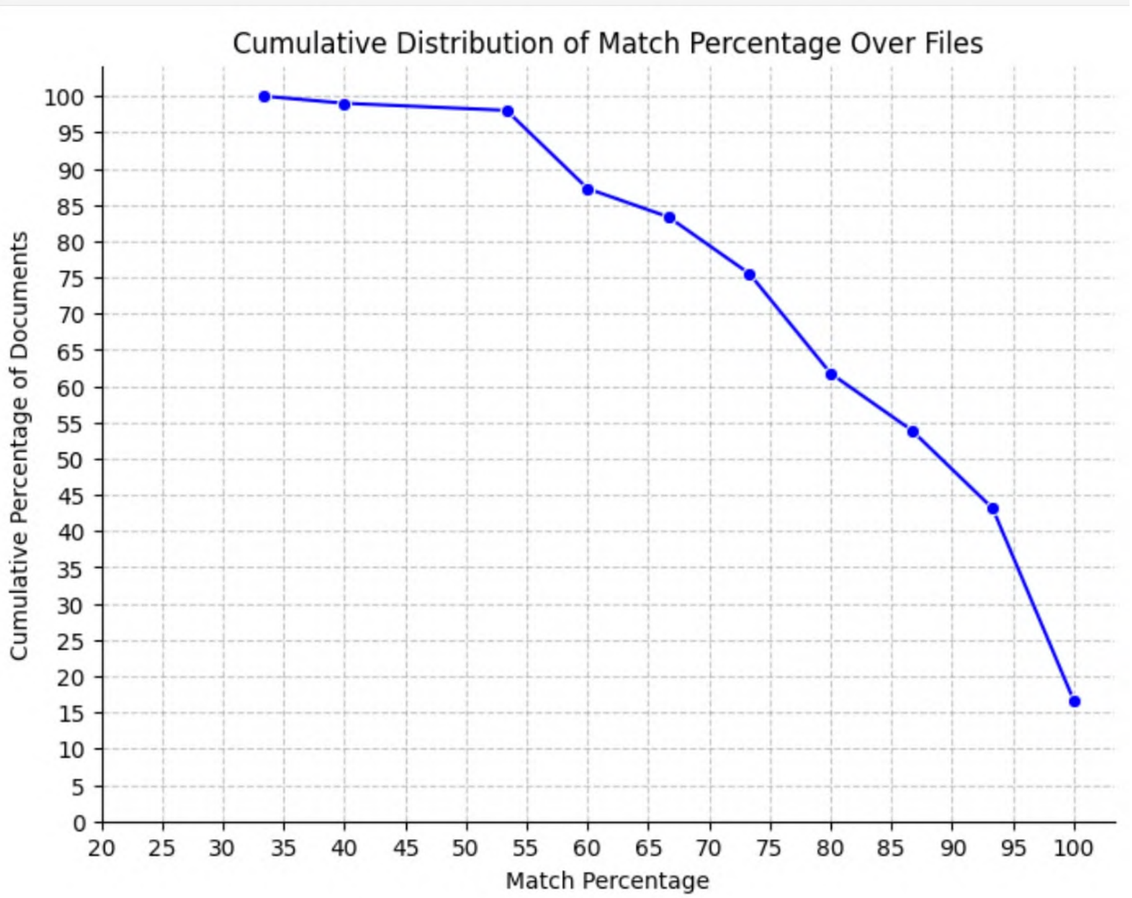
veracell

Results this far

We built a document structuring service, Intellens that we are piloting this fall with ERP providers, accounting and manufacturing companies.

We have devised ways to handle uncertainty and integrating structuring to fully automated processes.

Focusing on Azure based development from the start, we were able to proceed to pilots quicker than with ad-hoc local data science development.



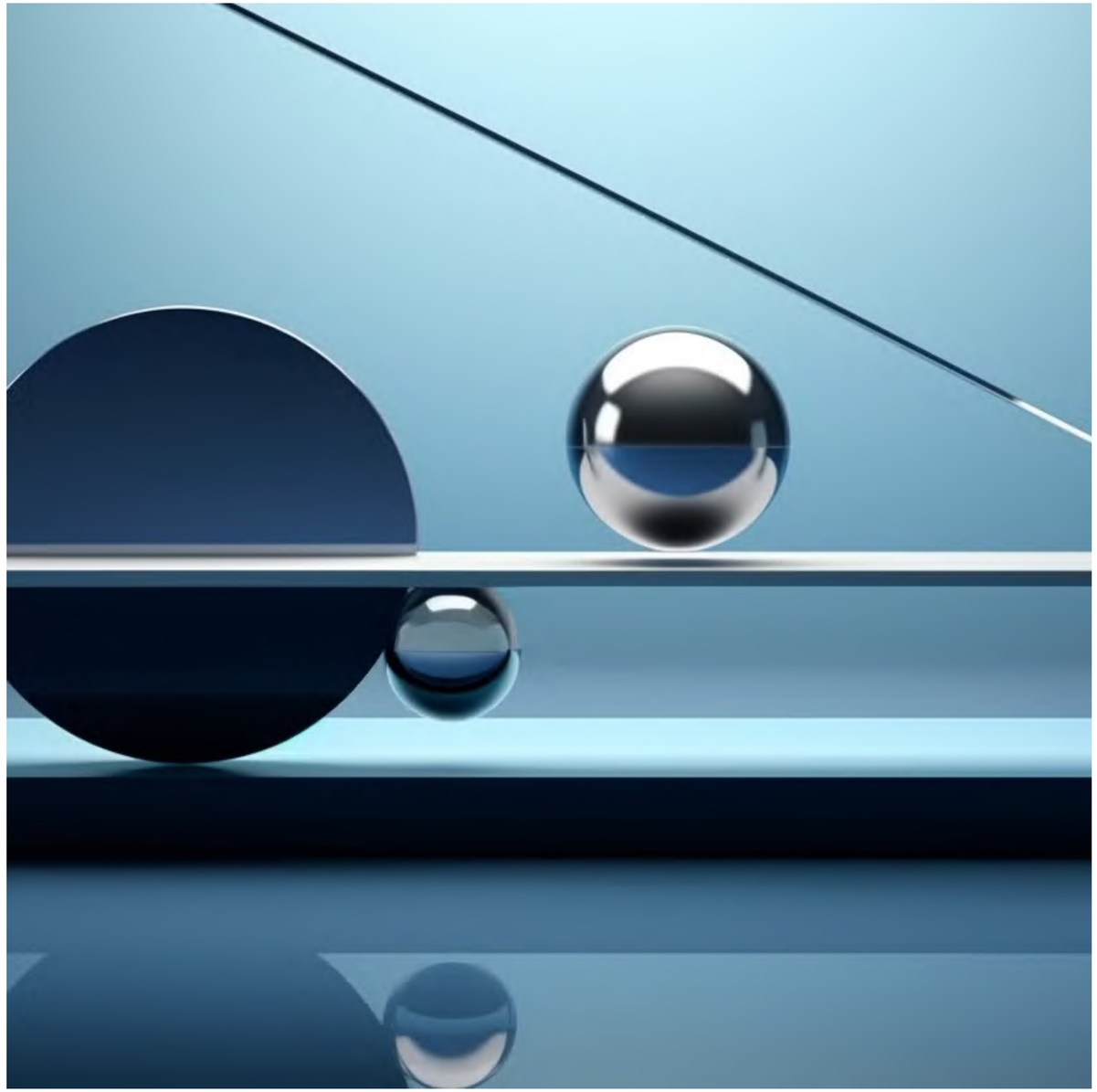
veracell

Evaluations and uncertainty

Model evaluation and handling uncertainty is key to building ML and AI application for production use. Often it is more important to get 40% of cases 100% right and knowing which they are than getting overall accuracy to 80-90%.

Solutions built on generative AI are often undeterministic. You need to measure the quality when model behaves as it should and detect when the issue is in the behavior itself.

Azure Prompt Flow has recently provided tooling for evaluation pipelines combining determnistic evaluators and tests and using Gen AI models themselves as evaluators.



veracell

What have we learned from working with Gen AI in Azure

Investing in cloud skills and knowledge brings benefit to all AI and data science projects by standardizing and streamlining practices.

We need to detect common solution patterns and use infrastructure automation, e.g. Terraform, to jumpstart new projects.

Evaluation must be automated and its results must be actionable: they must help to understand uncertainty and doing cost analysis.

Demand fluctuates and affects latency - model type or size may affect processing times less than region or time of day of year.

veracell