Cognitivo

The UI of AI

AI Factory Product Information

February 2024

'Artificial Intelligence is the new electricity.'

- Andrew Ng

CONTENTS

PART 1: ABOUT COGNITIVO

OUR COMPANY

Cognitivo a New South Wales based AI Software company with a mission to help our customers scale AI faster and more responsibly.

Our flagship product, AI Factory, helps our clients mass-produce in the AI era by allowing them to build composable business workflows that take advantage of pre-built ML models that can be fine-tuned in a user-friendly manner.

We are also experts in the fields of Technology Architecture, Data Science, Data Engineering, AI and Data Risk Management, advising both digital native (startup) and Enterprise clients.

Cognitivo collaborates with Australian Research Institutions, UNSW and CSIRO Data61 to develop new Software Engineering methods in building and managing novel risks posed by AI. Cognitivo is an authorised reseller of Data61's privacy re-identification tool (R4) and has led the founding of an industry-led research group called the Fintech AI Innovation Consortium (FAIC).

In 2023, Cognitivo was selected to be part of Austrade and InvestmentNSW's Fintech & AI Going Global export programs, participating in trade delegations to Singapore and New York.

Singapore Fintech Festival 2023

FAIC (Fintech AI Innovation Consortium)

PURPOSE

AI promises to deliver tremendous economic value for humanity, but at with all disruptive technologies has the potential cause greater inequality and even harm.

As digital dependence grows, Australia faces the risk that AI competencies may be concentrated in foreign owned entities (as with internet/social media).

We believe that AI is a technology that should be accessible and adoptable by all organisations. AI is the next major development paradigm in software engineering and therefore organisations cannot outsource or be dependent on external consultants.

However AI is new so requires new engineering and risk management practices to deliver projects successfully. Organisations need help, we want to do this in a way that helps build capabilities within our client organisations.

MISSION

Our mission is to drive successful AI adoption in our customers in a way that creates social, economic and environmental sustainability.

To achieve this we are:

1. Building software engineering methods and products that accelerate the adoption of AI-powered systems. Good design principles and architecture not only results in better business outcomes, but also more computationally efficient AI systems.
2. Building an developer ecosystem and marketplace to ensure the best algorithms and talent have a pathway to customers that is open and accessible.
3. Furthering the develop of responsible AI and privacy guardrails and tools.

We dedicated to building and empowering ecosystem of developers and customers to accelerate the responsible adoption of AI with novel tools and modes of collaboration.

R&D

AI is a new technology concept to most organisations, therefor we need new knowledge, practices and skills to drive successful commercial adoption.

Cognitivo has engaged in research with UNSW since 2018. Our first project was a study of the efficacy of machine learning methods for credit risk scoring versus traditional logistic regression methods and the effects of bias within training data.

In 2019, Cognitivo successfully obtaining a NSW TechVoucher grant to conduct research into the area of purchase receipt categorisation using natural language processing and semantic modelling.

Cognitivo has had a close relationship with CSIRO's Data61 since 2019 in the area of research translation. Cognitivo has been a commercial reseller of Data61's R4 privacy tool since 2019.

In 2022, Cognitivo, BrewAI, Westpac, Data61, AWS, Databricks and CSIRO's Data61 formed an industry-lead research consortium called the Fintech AI Innovation Consortium (FAIC). The FAIC is founded under the umbrella of UNSW's AI Institute.

Though 2023, the FAIC has onboarded 10 new international PHD students engaged in projects relating to MLOps, autoML and AI applications for ESG.

FINTECH AI INNOVATION CONSORTIUM · UNSW SYDNEY · UNSW AI Institute · DATA 61 CSIRO

cognitivo The UI of AI · aws · BrewAI · databricks · Westpac

In 2022, Alan Hsiao (Founder of Cognitivo) was conferred the title of a senior visiting fellow at UNSW and works actively to support the FAIC's research program and industry engagement.

In Sept 2023, Cognitivo acted as the lead organisation in an $8m project, applying for a $3m federal government (DISR) CRC-P grant. The program included pledges from Westpac, Data61 and UNSW to develop new technologies on top of the existing AI Factory framework and software platform.

Selected Experience in Data and AI

Development of a custom application provide automation in compliance activities relating to Ongoing-Customer-Due-Diligence.

ReactJS application front-end with Azure Data Factory for data ingestion, Microsoft Azure functions to process PDF files into image files.

Development Databricks image processing models and ML/Ops pipelines for continuous training.

Advised the Westpac CTO on the revision of the Bank's Data Architecture.

Development of a new technology target state based on Data Mesh.

Charted a transition from on-premise data warehouse platforms to a cloud-based Lakehouse architecture.

Development of a roadmap in response to APRA's 5-year data collection roadmap

Developed a data management reference architecture which unifies privacy, information security, data use/availability and records management.

Develop a Proof of Value (POV) to demonstrate the application of Artificial Intelligence / Machine Learning to 4 automation use cases within RevenenueNSW

POV demonstrated unassisted service through a chatbot, handwriting recognition, text to service request detection and fraud detection capabilities using a knowledge graph.

Performed a privacy reidentification risk assessment on TfNSW data sets prior to publication on the Transport Open Data portal (as part of the Enterprise Analytics platform program). Project utilised Data61's R4 tool with Cognitivo's in-house developed data risk management methodology (based on ISO31000).

Develop a data sharing framework for DRIVES to handle the production, publication, and approval of DRIVES data for internal, inter-agency and public requesters of data. Data sharing framework set the strategy for data usage, publication and governance for TfNSW's Enterprise Analytics Platform.

Partners

We leverage industry leading software within our product.

Databricks is the Data and AI company. We are helping governments across the globe leverage data, analytics and AI to help reduce fraud, make real-time decisions, and provide better experiences to their citizens. Databrick's lakehouse architecture, combines the best elements of data lakes and data warehouses to help you reduce costs and deliver on your data and AI initiatives faster.

Stardog is a knowledge graph platform that helps manage and query graph data, representing information as a graph database with nodes, edges, and properties. It supports RDF, SPARQL, and OWL standards for modeling and analyzing complex relationships in data.

Fivetran is an automated data movement platform that extracts, loads and transforms the world's data. Built for the cloud, Fivetran enables data teams to effortlessly centralize and transform data from hundreds of SaaS and on-prem data sources into high-performance cloud destinations.

Microsoft Azure and Amazon Web Services are Cognitivo's preferred cloud vendors, with AI Factory available within both cloud hosting services.

PART 2:

RESPONDING TO CHALLENGES POSED BY AI

AI PROMISES GREAT THINGS, BUT COMES WITH NEW RISKS

The duality of AI

AI promises to deliver tremendous economic value but needs new software engineering and risk management practices to deliver projects efficiently, responsibly & at-scale.

AI presents significant economic upside but also poses new risks to our society.

- AI is growing into a $13tr USD pa global opportunity by 2030 (McKinsey), projected to be worth $315b pa by 2028 (Australia AI Action Plan 2021).
- “AI is seen by many as an engine of productivity and economic growth but may also have a highly disruptive effect on our economy and society, through the creation of super firms and hubs of wealth and knowledge.” EU Parliament’s report Economic impacts of AI (2019).
- 2017 defeat of the world’s top Go Player by AlphaGo was seen as a Sputnik moment in China, sparking an AI arms race (Kai Fu Lee, AI Superpowers, 2018).
- In 2021, US National Security Commission on AI report called out new threats, including information operations, biotech, data harvesting and targeting of individuals and adversarial AI.
- “Using AI safely and responsibly is a balancing act the whole world is grappling with. The upside is massive. What is needed next to build trust and public confidence in these critical technologies.” (Hon Ed Husic, 2023).

A new paradigm requires new tools

AI differs from traditional programming because it does not rely on prescriptive rules (if-else) rather being programmed through examples. Given this shift, AI development maturity is low with Gartner reporting in 2022 that 85% of all AI projects fail. For example, developers are typically segregated from real data for security reasons.

As digital dependence grows, Australia faces the risk that AI competencies may be concentrated in foreign owned entities (as with internet/social media). The integration of OpenAI’s ChatGPT into Microsoft’s AI offering.

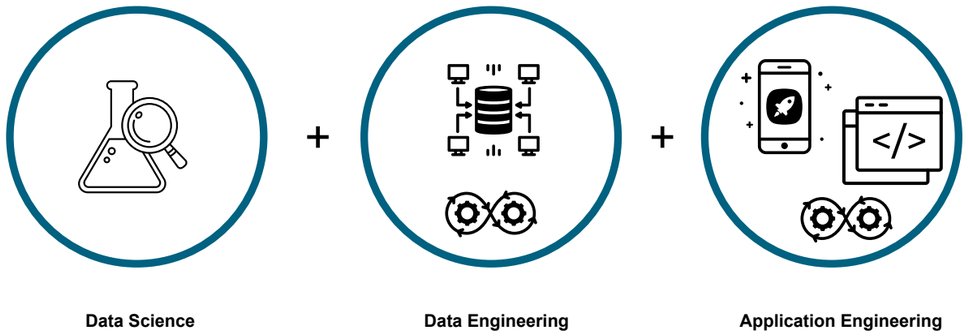
Trust and cyber security such as explainability (how can we ensure systems perform as intended?) and privacy (how do we meet the right-to-be forgotten in trained models?) pose new challenges to be solved.

To mass produce in the AI era, we need new software development methods and security controls supported by tools that enable both data scientists and software engineers to work seamlessly (i.e., a software factory).

Our philosophy and design principles

Cognitivo has productised over 5 years of AI development field engineering experience and with that we have incubated a distinct software design and engineering philosophy. These include;

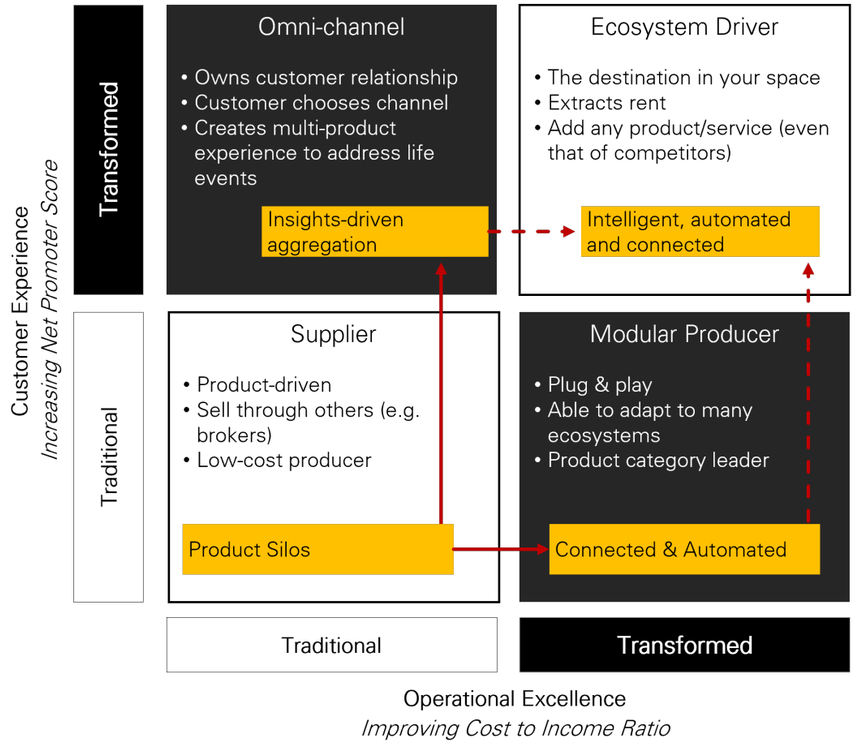
- **Human Centered Design** put into practice in a process we call AI infusion, which is the combination of CX and AI. The hardest part about getting AI to work is getting AI to understand what "right" is. We have developed a process where human input is plugged into machine learning algorithms but we also implement a human in the loop process that results in highly accurate business processes and systems that can reach a very high level of automation over time.
- **Software Engineering for Machine Learning (SE4ML)**, Cognitivo has fine-tuned a set of software development processes and architectures that support the development of software engineering systems that include machine learning (ML) components.



- **Openness** in the way we collaborate and adopt open standards.
  - We are plugged into research organisations (UNSW and Data61) as well as industry partners in driving architectures that drive successful adoption of AI.
  - We take care to adopt open / industry standards e.g., RDF, FIBO.
  - Where possible we leverage open source technologies that are multi-/hybrid cloud compatible. E.g., Apache Spark, Delta Lake format.
  - We believe in federated / mesh architectures as opposed to monolithic architectures.

Cognitivo's Digital Strategy Framework

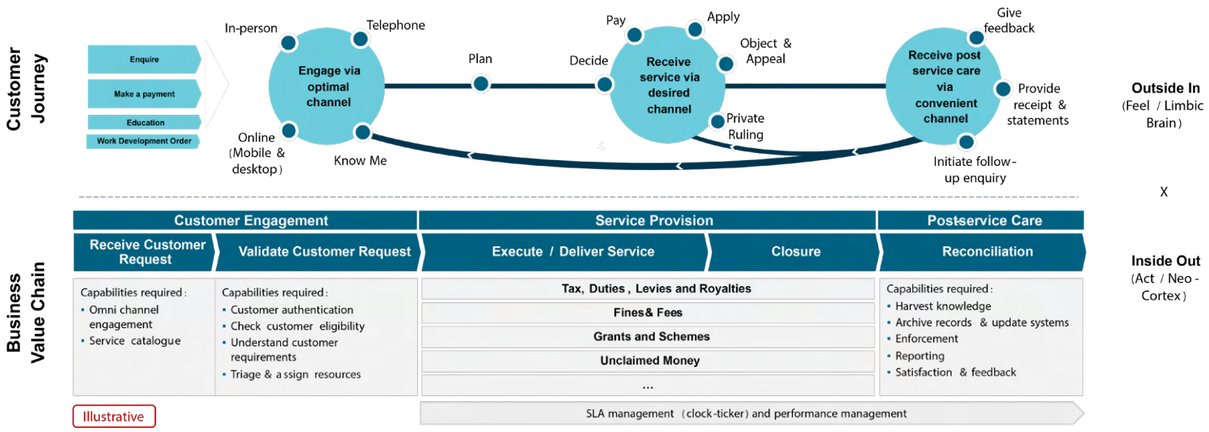
For most large organisations CX and process efficiency are at opposite ends of the see-saw, with one being achieved at the expense of the other. For example, a customer liaison would be able to guide a customer through a maze of complex business processes and organisational silos to achieve the desired customer outcome, however this enhancement of customer satisfaction erodes operational efficiency.



Customers want a seamless and convenient process where they are guided through the process of fulfilling their needs. For example, customers hate being told that they have to be transferred to another team and restart the process of explaining what their needs are or re-authenticating themselves again.

On the flip side, operational efficiency is built on standardisation of processes and the division of labour to simultaneously achieve process quality and lower the level of employee expertise.

Customers see organisations, products and services from the outside in where organisations see capabilities, offerings, processes and teams from the inside-out. Leading digital enterprises need to be able to bridge these opposing forces. Technology and automation can help, for example, a chat-bot that is integrated into an API enabled core system can offer 24x7 service, with a human touch.



To simultaneously enhance customer service and achieve process efficiency objectives, organisations need to develop in foundational (horizontal) capabilities.

For example, these include:

- Product service catalogue
- Customer consent and entitlements register
- Customer authentication
- In-bound & out-bound omni-channel messaging (email/sms/chat)
- Knowledge base of customer interactions
- Customer interaction and risk graph
- Proactive credit risk & fraud detection
- Workflow, queue and resource management

Building an engaging and highly automated organisation

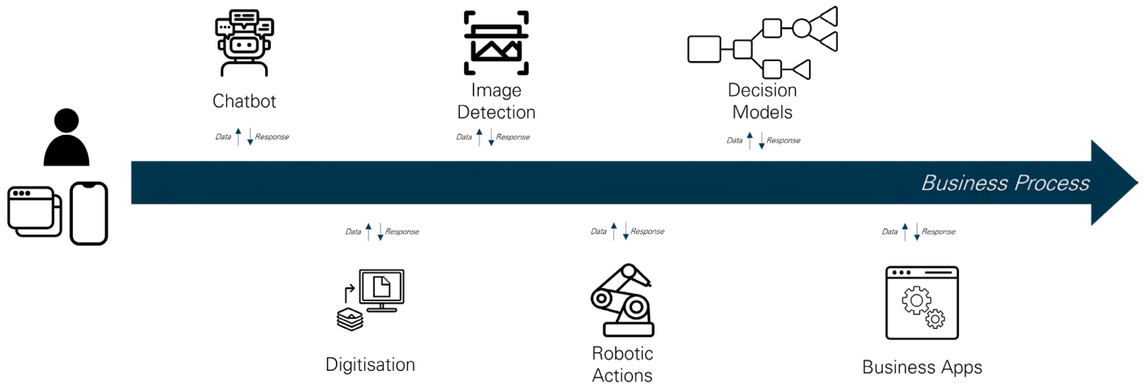
There is a race for vendors to start utilising AI in their service offerings, however as an organisation using AI, you end up spending all your time sending data into multiple black-box AI systems you can't control.

How do you manage trust with so many externally owned AI models? AI models can hallucinate, be biased and cause harm. As you use AI, your vendors get smarter, you don't.

Given so many disparate services are being called, there is a rigidity to business processes and a cap on how much overall automation is possible (~5%).

Organisation's create a lot of data, it's a tremendous challenge to move data between applications. Throughout through message-based interfaces (e.g. API's, web services, MQ) have always struggled to keep pace with data volumes within business applications.

With the growth of AI, every application wants to be smart, so we will witness a growth in business application building their own data storage and analytical capabilities. Salesforce Einstein is an example of how a front-office CRM developed an adjacent data lake and AI offering that completes with an organisation's traditional centralised data warehouse teams.



Well defined (and slow changing) interfaces are important for inter-team or B2B integration, but within a domain where change frequency is high, an approach that supports higher agility is required. From an enterprise architecture perspective, over the past 5-10 years, the emergence of real time streaming and parallel compute / big data technologies such as kafka and spark there has been a gradual convergence of capabilities within OLTP and OLAP applications.

Where there is significant data volume or gravity, it is far more efficient to move the compute or algorithm to where data resides.

To a much higher level of automation (e.g. 20%-80%) an organisation needs to assemble the core ingredients involved in human activity and cognition.

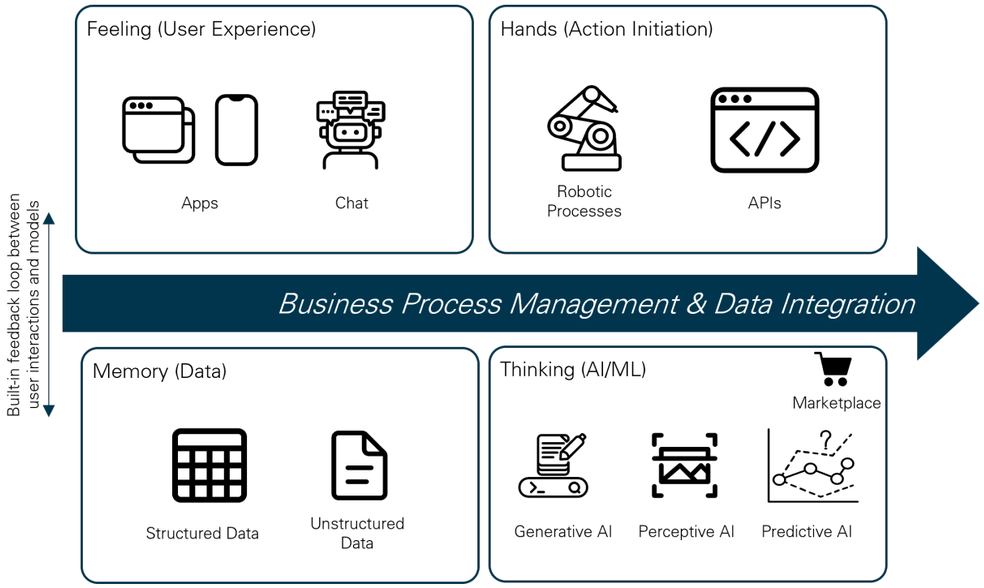
This starts off with the "**data**" layer providing the ability to bring data from multiple sources and formats (e.g. documents, video, audio and structured data. To achieve this, both low cost object storage (S3/ADLS) as well as high performance, specialised storage such as SQL warehouses, graph, document and vector databases.

The "**thinking**" part of the system involves an ensemble of methods that can process data to derive new outputs. This includes deterministic decisions logic (if-else), what we call Good-Old-Fashion AI as well as machine learning methods.

All 3 branches of machine learning is required to replicate what humans see, think and say. These are:

- **Perceptive AI** – ability to read images, interpret audio.
- **Generative AI** – ability to generate human-like responses (text, voice, etc.).
- **Predictive AI** – ability to forecast future events.

Compared to traditional data architectures which send data to functionally siloed analytical applications, we can achieve a much better result executing an ensemble of algorithms over all of the data we have available inplace.



Once decisions are made, actions must be taken, the "**hands**" of the system involve ability to choreograph and invoke internal and external services.

Finally, the "**feeling**" part of the system composes the 3 other capabilities into seamless and convenient experience, available in any channel of choice.

Mesh topology

We want to build "intelligent" applications while not creating a high degree of coupling into a centralised data store (e.g. data lake or data warehouse). To achieve this, the enterprise needs to adopt a domain-driven approach, organising application domains within a mesh or federated topology.

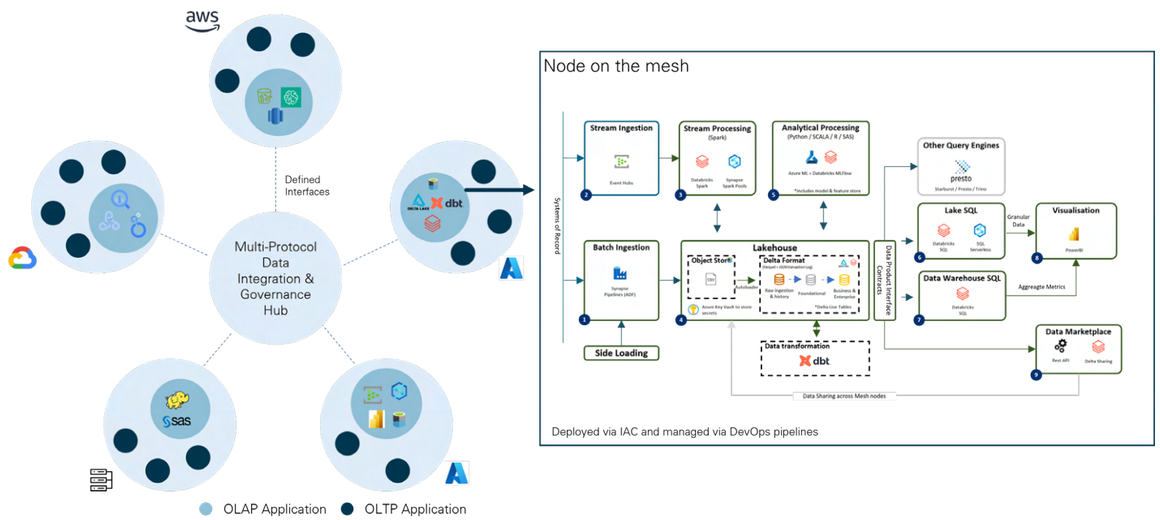
This means within a domain (a collection of related business offerings, processes and customers), more data can be sent to a domain-level data repository to perform department-level data joining, aggregation and analytics. Those domain-level data products can then be exposed as API's to inter-domain consumers. This means business systems within domains have a high degree of change agility while subscribers of those data products are insulated from source system changes given they use domain-aggregate data rather than business-system level interfaces.

To put this into perspective, if AI Factory was to be used across 3 different departments, you would deploy an instance (via code) of the AI Factory into each hosting environment closest to where the source data resides. Computation would be right-sized for each department or use case and data between these departments would be segregated enhancing privacy/security as well as resilience.

There are many "meshes":

- a service mesh handles communication between services within an application (intra-application messaging)
- an event mesh a network of interconnected event brokers that enables the distribution of events information among applications (inter-application eventing)
- a data mesh is an analytical data architecture and operating model where data is treated as a product and owned by teams that most intimately know and consume the data

AI Factory has been architected to respect and adopt these mesh architecture patterns, but importantly is designed to be multi-pattern, multi-protocol capable.



PART 3:

OUR PRODUCT

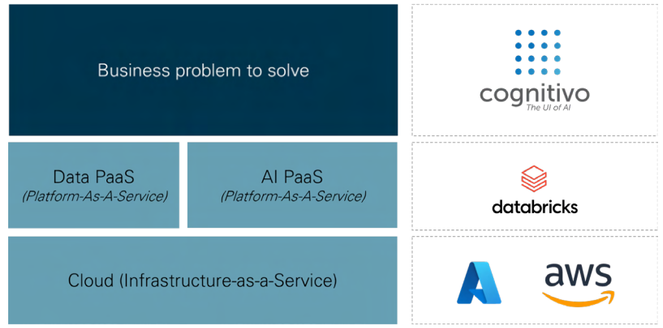
AI FACTORY

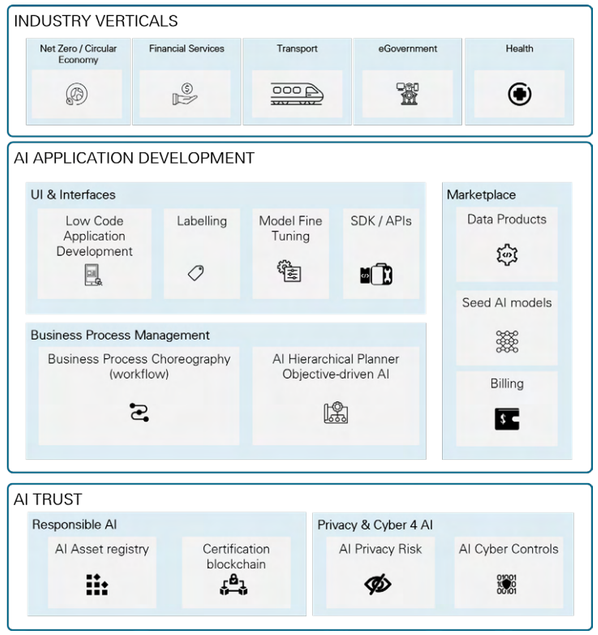
Self-service AI-powered workflow automation

AI Factory is a platform and framework that builds workflow-based applications on top of cloud-based data & AI platforms, powered by a marketplace of fine-tunable machine learning algorithms.

For example, someone wants to build a process to automate workflow invoice processing; they can use our platform to define the UI and workflow and choose an invoice detection model from our marketplace.

Cognitivo's AI Factory is "the UI of AI" allowing business users to utilise highly technical capabilities in a user friendly way.





AI Factory allows users to:

- Develop bespoke user applications to complete workflow based tasks.
- Incorporate structured and unstructured data inputs (including documents, images, audio and video).
- Perform data transformation over incoming data (mapping of data to industry data standards e.g. FPML, FIBO, XBRL, eInvoice PEPPOL/EDIFACT).
- Integrate with 3rd party systems and data services (e.g. Xero, Redbook, Transport Open Data Portal).
- Perform data labelling / annotation efficiently in production environments.
- Fine-tune pre-built machine learning algorithms covering perceptive (object detection, document classification, text extraction, predictive (credit risk scoring, fraud risk), generative (chat, summarisation, recommendation).
- Access an ML developer marketplace.
- AI trust and privacy tooling.

UX

- Low-code user interface to configure front-end screens.
- Software Development Kit (SDK) for 3rd party integration into AI Factory (e.g., Pega, Appian).
- Low-code business process and workflow definition
- Labelling & annotation
  - In-app (production) data labelling & annotation
  - AI assisted labelling

ACTION INITIATION

- Messaging (SMS / Email / Whatsapp)
- Workflow and process orchestration
- API Invocation
- Data preparation for RPA initiation

AI

Generative AI
- Unassisted chatbot
- Text Summarisation and feature extraction
- Retrieval Augmented Generation (RAG)

Perceptive AI
- Handwriting detection
- Object detection (street signs, identification documents, invoices)

Predictive AI
- Fraud detection
- Credit risk assessment

AI Trust & Assurance
- Privacy, re-identification risk assessment
- ML model user rating
- ML model accuracy bias assessment

DATA & INTEGRATION

Lake storage on commodity object store
- Storage of structured / tabular data
- Storage of unstructured / document data

Data processing & serving
- Data transformation, conformance and serving (SQL data warehouse)
- Semantic Graph Database

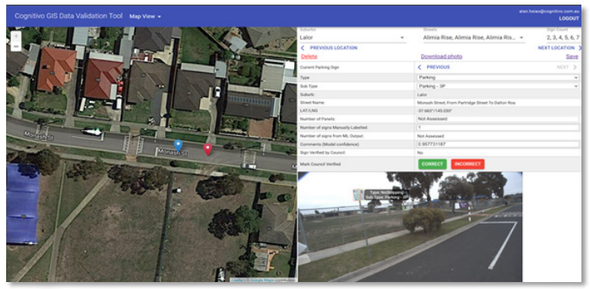
Integration
- Batch & microbatch: 160+ batch integration via FiveTran
- Real time: API, pub/sub, kafka
- 3rd party application integration (e.g. Xero)
- Data sharing (delta share)

Data Governance
- Data catalogue (Unity catalogue)
- Privacy, retention and disposition management

Key features and use cases

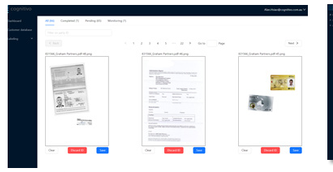
Local council roadside asset detection

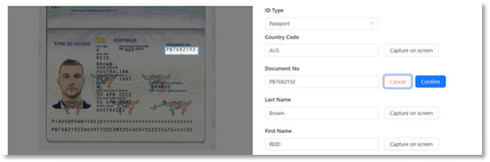
- Computer Vision used to catalogue every street sign within a local council.
- After the data science work, there was no way to review the results, so we built them a custom ReactJS application to interact with AI detected outputs.
- We then added a button to mark which detected outputs had been manually verified.



Customer Due Diligence, Know Your Customer

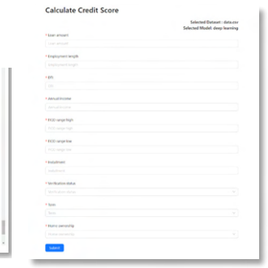
- Bank needed to check if IDs held on file are still current. Estimated effort of > 5000 days effort.
- AI Factory Ingested millions of PDF documents from FileNet and detected 13 types IDs.
- Corrections to detected results (including bounding boxes) used in retraining.
- Deployed to client's Azure tenancy, passing global security requirements.
- Saved ~80% manual effort.

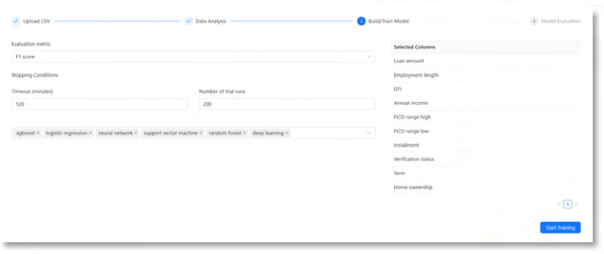




Risk Scoring (Credit Risk & Fraud)

- Credit risk and fraud risk scoring using multiple machine learning algorithms (e.g. xgboost, random forest, support vector machine).
- User can select taget feature for prediction, input attributes are auto-suggested.
- Data preparation and data quality analysis are automated.
- Out ML algorithm served through a serverless API endpoint.

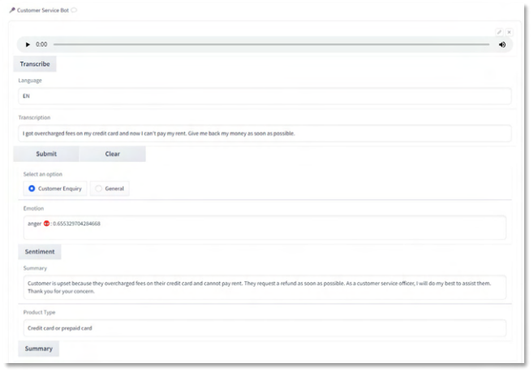




Key features and use cases

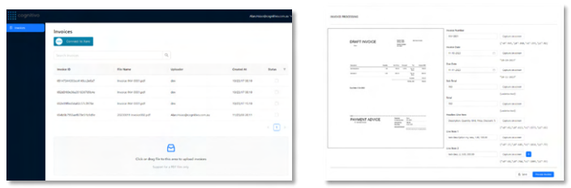
Customer Service Agent LLM

- LLM trained to transcode audio to text.
- Perform sentiment analysis and request urgency.
- Detect product or service type.
- Integrated knowledge base for retrieval augmented generation (RAG) for generating recommended staff responses.



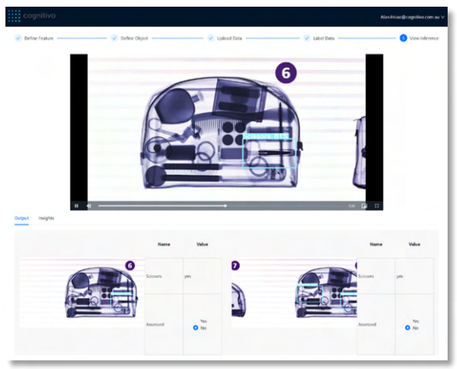
PDF invoice to e-invoice conversion

- Extraction of invoice data from PDFs.
- Mapping of extracted fields to multiple eInvoice schemas (UBL, CII, EDIFACT, FactureX, GS1, Odette).
- Access to eInvoice communication networks (network access points).
- Automated creation of invoice and bills within ERP/accounting systems.



Object detection in x-ray imagery

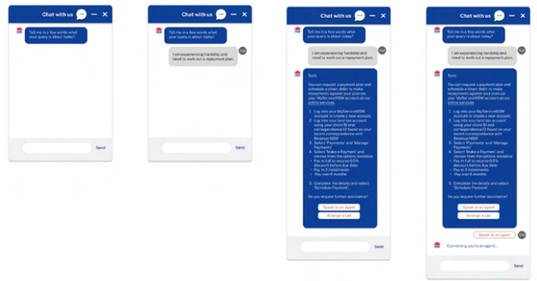
- Real time processing of x-ray video imagery at travel ports to detect banned items.



Key features and use cases

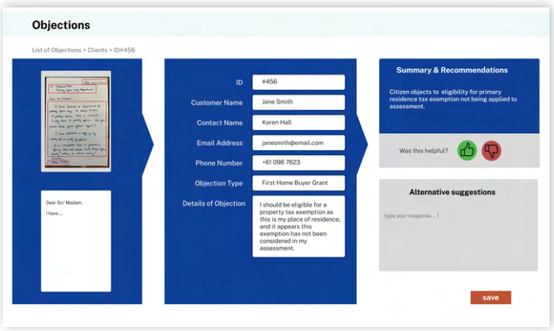
Customer Service Chatbot

- RAG (Retrieval Augmented) enhanced chatbot that can handle frequently asked questions relating to tax collection.
- Integration of 2FA (SMS OTP authentication) within chat.
- Real-time customer details and workflow status lookup within chat.



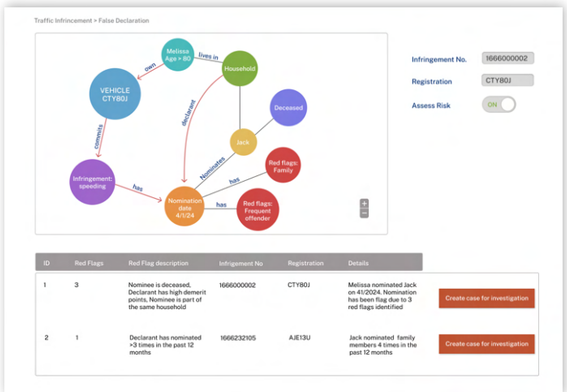
Customer Service Request

- Handwriting recognition to process inbound scanned mail.
- Automation extraction and interpretation of customer details and service request / objection details from transcribed text.
- Automated suggestion of next-best-action.



Detection of fraud relating to traffic fines

- Fraud detection using a relationship-based graph database.
- Automated raising of red-flags through inference over relationships between networked entities.
- Virtualisation of data stored within data lake storage to ensure real-time consistency between knowledge graph and other analytics data sources.



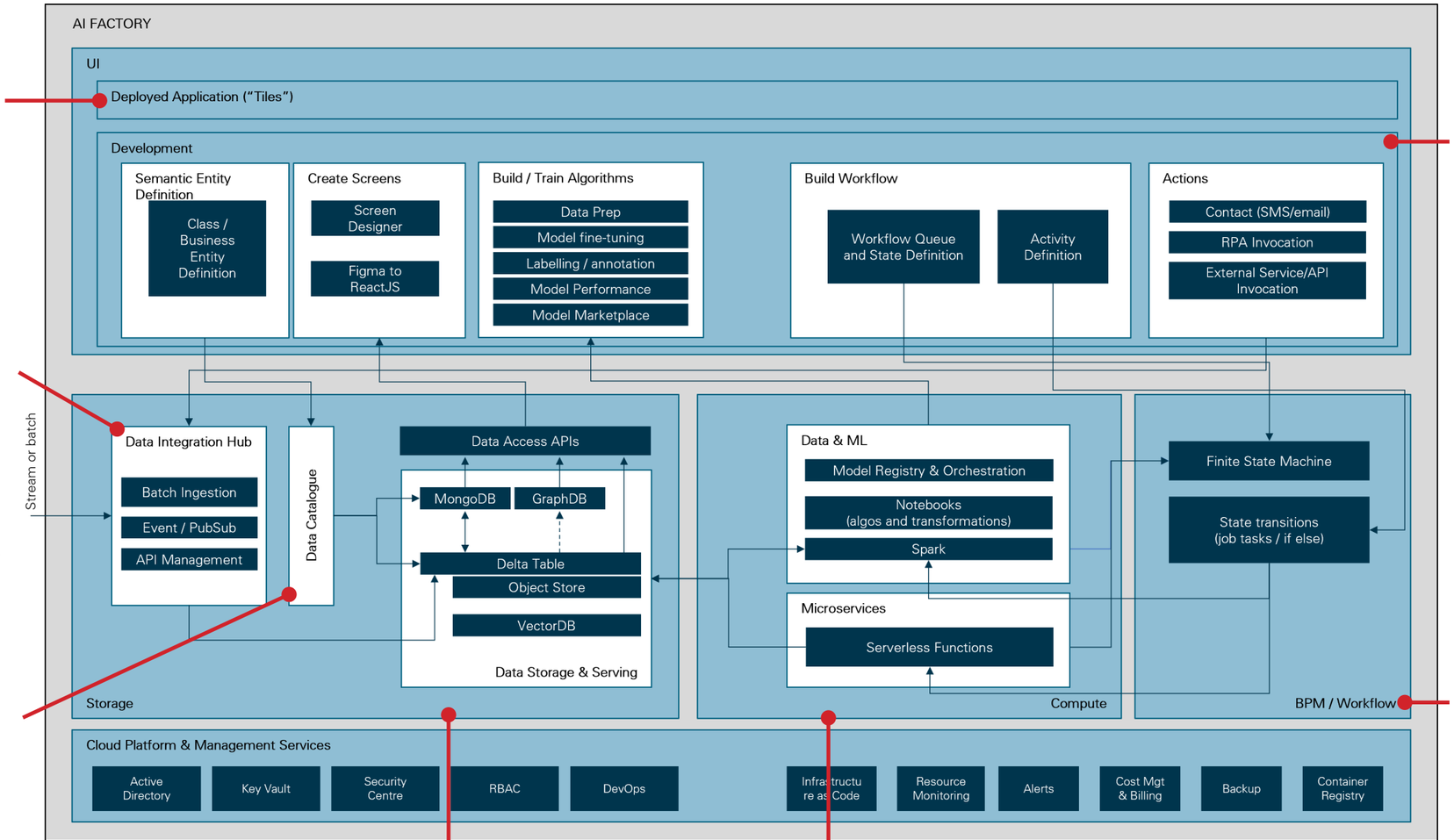
AI FACTORY SOLUTION ARCHITECTURE

Deployed business user applications
- An organisational unit will have multiple “tiles” deployed performance various automation use cases.

Batch integration through FiveTran
- 1-minute latency microbatch connection to over 160 applications.
- Ability to develop custom connectors to proprietary systems.
- Near realtime batch capability through HVR database integration

Event Streaming and pub/sub support for:
- Azure EventHubs, Confluent Kafka, AWS Kinesis and Solace

Data Catalogue & Data Governance using Databricks Unity Catalogue.



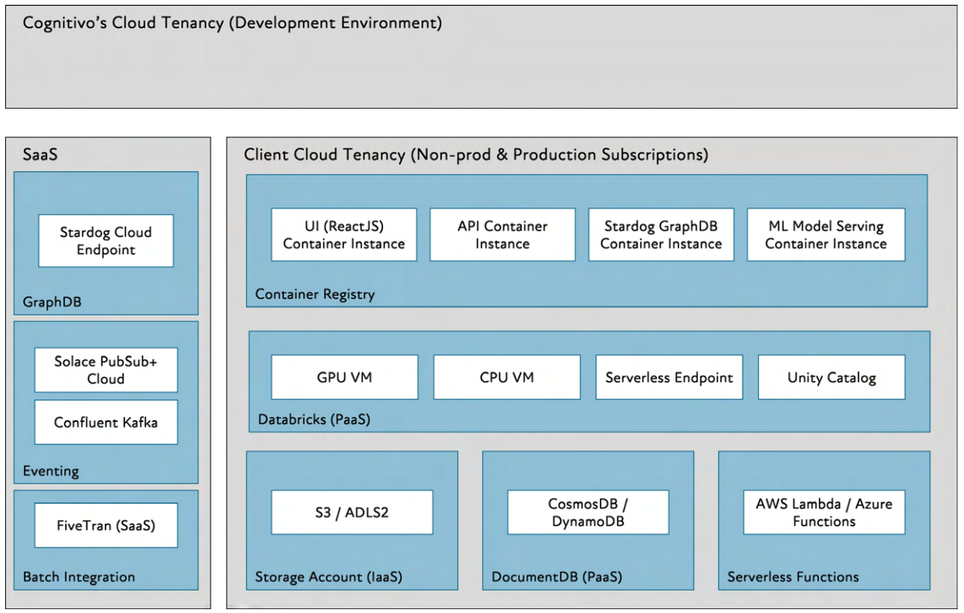
Application Development Process
- Definition of business entities (classes) in a semantic (human understandable) manner.
- Screen design where fields are bound to entity attributes.
- Fine-tuning algorithms entails mapping of the semantic model to model features.
- Labelling data entails mapping of training data to semantic model attributes.
- Workflow queues and states and activities define the navigation of individual screens defined.
- Actions can be invoked through the workflow or bound to on screen buttons.

Business Process Management
- The progression of business entities (objects) through a workflow is dependent on status fields within the finite state machine.
- State transitions are driven by business rules define over case states.

- MongoDB serves transactional ata to the application front-end.
- Ingested CSV files are processed into parquet format within cloud object store (ADLS2/S3).
- Files are further processed into delta-format, providing ACID integrity, zero-copy clomining capabilities.
- Data is virtualised into a GraphDB where relationship tables are transformed into edge/predicates and rules-based relationships inferred.
- The VectorDB is used by LLM RAG pipelines to access real-time information.

- Functional programs (i.e. deterministic logic) are deployed and executed through serverless functions.
- Probabilistic (i.e. ML) or data intensive programs are developed in Databricks notebooks and deployed as containers, executed via Kubernetes on appropriately sized compute instances (cloud VMs) or Databricks serverless endpoints.

DEPLOYMENT MODEL



AI Factory is developed in Cognitivo's development subscription and deployed into the Client's cloud tenancy through DevOps pipelines. Cloud resources are instantiated through Infrastructure-as-code pipelines.

AI Factory utilises a combination of IaaS, PaaS and SaaS components. The deployment model (IaaS/PaaS/SaaS) can be chosen based on each component according to each client's security, performance and cost requirements.

OPERATING MODEL

LICENSING

- AI Factory monthly license cost (increments of 100 users).
- Model invocation fee (based on complexity of ML models).
- Partner services (e.g. Redbook, Twilio, Experian) available through resale. Generally a pass through rate unless additional services are required.

MANAGED SERVICES

- Dedicated support (fail / fix) capacity.
- Dedicated on / offshore capacity for use case delivery / implementation.

KNOWLEDGE EXCHANGE AND TRAINING

- AI field engineering tutorials.
- AI short courses (jointly delivered with UNSW School of Computer Science & Engineering).
- Invitation to Cognitivo's Architecture working groups (topics such as MLOps, AI Software Engineering Architectures, BPM in the age of AI).

INTELLECTUAL PROPERTY

- Cognitivo is open to a range of IP arrangements for R&D purposes with partners and client organisations.


cognitivo
The UI of AI