transACT
TECHNOLOGY SOLUTIONS

Case Study: INQ Data

INQ Data unleash additional customer revenues with AI Chatbot service

Partnership

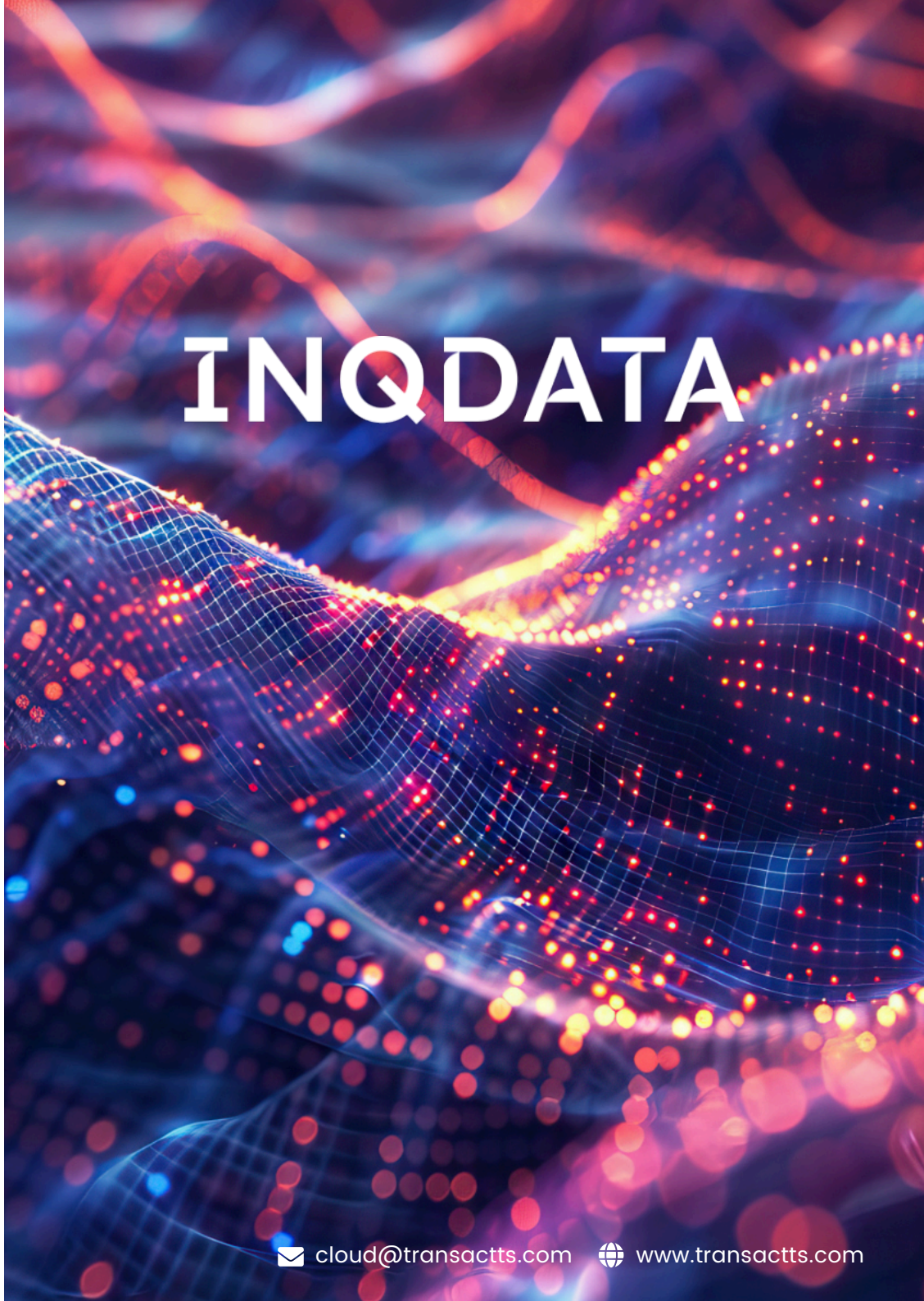
INQ Data have worked closely with transACT throughout their onboarding in addition to a number of projects both completed and ongoing. For this reason they have chosen transACT to help design and deploy a Natural Language chatbot to allow their end users to interrogate data held by INQ as well as allowing it to be enhanced and supplemented by external data sources. Throughout the project transACT worked closely with INQ to design, build and operationally handover with continued ongoing assistance.

Business Challenges

To enhance the avenues for INQ to provide their data to their customers and to extract more value out of it. Making the data available in multiple different ways include through the use of natural language querying, while also enhancing it with external sources and allowing for additional interrogation.

Solution

INQ partnered with transACT to build out a financial chatbot system to enable their customers to access market data hosted by INQ through natural language querying with the added benefit of utilising 3rd party APIs to further enrich the data provided by INQ, offering a one stop shop for financial data. To do this the system has been built utilising AWS Kendra, AWS Sagemaker, AWS Bedrock, AWS Lambda, AWS Lex and AWS DynamoDB. Using AWS Kendra to connect to S3 Buckets hosting financial metadata, Custom data connector to connect to INQs KDB databases (Not all data from the KDB engine was ingested into Kendra due to cost). Kendra allows for fast unified search of enterprise data making it readily available for customers to use.



cloud@transactts.com www.transactts.com

INQDATA

The solution relies on data from Kendra as well as making available to it API calls to KDB to fetch live market data and calling 3rd party tools and extensions to further pull in more data when needed, e.g Yahoo finance articles, financial reporting, SEC reports. AWS Lambda was used to orchestrate the different steps. This has allowed INQ to give access to their data in a novel, easy to use, and easy to understand way while providing additional context. Opening a new business avenue for them and an improved experience for their customers.

The solution used multiple AWS technologies:

1. Amazon Sagemaker and Amazon Bedrock: Chat based LLM models in bedrock additionally models from within Sagemaker and models imported from Huggingface utilising the built-in functionality of these two services to perform additional tuning.

2. Amazon Kendra: Used to connect to all internal data sources and provide fast and efficient searchability of the data.

3. AWS Lambda: Used to orchestrate between Lex, Kendra, the LLM and the third-party tooling, as well as to make appropriate calls to 3rd party tooling when determined by the LLM.

4. AWS Lex: Allowed easy building of the Chat solution and provided initial NLP for the orchestration to pull the data from Kendra to pass to the LLM.

Implementation:

Model Selection and Customisation

- Chat based FM models such as Titan, Claude 3 Sonnet, Claude 3 Haiku from bedrock, Meta Llama, Dolly from sagemaker, Bert and some of its derivatives from Hugging face were tested and were selected based on speed, answer accuracy, consistency, and context size.
- Fine tuning including parameter tuning and prompt engineering were then carried out to ensure the model would answer in the appropriate tone and frame the information correctly.
- Sagemaker models were ported into Bedrock and evaluated.
- Fine tuning the models by training with additional data to teach when to call the APIs (when possible dependant on the model card), changing parameters such as temperature, Top K and/or Top P to find what worked best and also using different multi prompt approaches and testing against single prompt comparing performance to token(price).

Performance Optimisation

- A number of solutions were considered for the data stores, vector DBs, and data indexing. In the end Kendra was chosen for it's speed and scalability, and because of it's ongoing optimisation as it is used.
- Evaluating model response speed, and accuracy and balancing against cost.

System Integration

- Lambda Functions were built to orchestrate between the steps of the solution. Also to handle calling on 3rd party tools to pull in more data and orchestrate that pack to the LLM.
- AWS Lex was used as the frontend chat, allowing for easy integration and customisation and scalability.

Scaling and Load Management

- All AWS services used provide effective scaling. This was further tested with load testing to make sure there were no bottle necks or issues with any of the interactions.
- The system was built with exponential back off and timeouts with the ability to return appropriate responses if their was any issues with the 3rd party tooling.

Production Rollout

- The system was was end to end validated ensuring it is fully operational.
- It was run through a QA process testing for fringe cases as well as a wide range of use cases.

INQDATA

- UAT was carried out with key stakeholders and select friendly end users to ensure the system met all business requirements.
- Meta LLama 7b was chosen for being a light weight cost efficient model that met the correct tone and information accuracy requirements.
- Using a longer multi-shot prompt with examples indicating how to present information and tone to use gave the best results.
- Low temperature 0.40 gave the best information accuracy.

Monitoring and Continuous Improvement

- AWS Cloudwatch was configured with logging and metrics, Alarms set up on various metrics to make sure the system all was working as expected.
- CI/CD pipeline was set up and a dev pipeline, integrated as part of INQs processes and procedures for continued development and improvement of infrastructure and software.
- Pipeline for deploying and testing new LLMs and ability to promote them to production if they beat existing LLM on benchmarks and are user accepted.

Products and Services:

AWS Kendra, AWS Sagemaker, AWS Lambda, AWS Lex and AWS DynamoDB

Outcomes:

- Novel Solution for delivering data to INQs customers.
- Highly Scalable resources scaled up when needed and down when not needed allowing for cost efficiency
- Enhanced Core business of being a data provider by providing new avenue for data.
- User friendly easy to use means of accessing data.