transACT TECHNOLOGY SOLUTIONS

Case Study: Ascential

AWS-Powered Sentence Embedding Solution

This case study outlines the process of transitioning a sentence embedding The implementation of this AWS-powered sentence embedding solution resulted in a significant reduction in processing time, leading to a 30% increase in system efficiency and a projected ROI of 150% within the first year due to improved scalability and cost-effective operations.

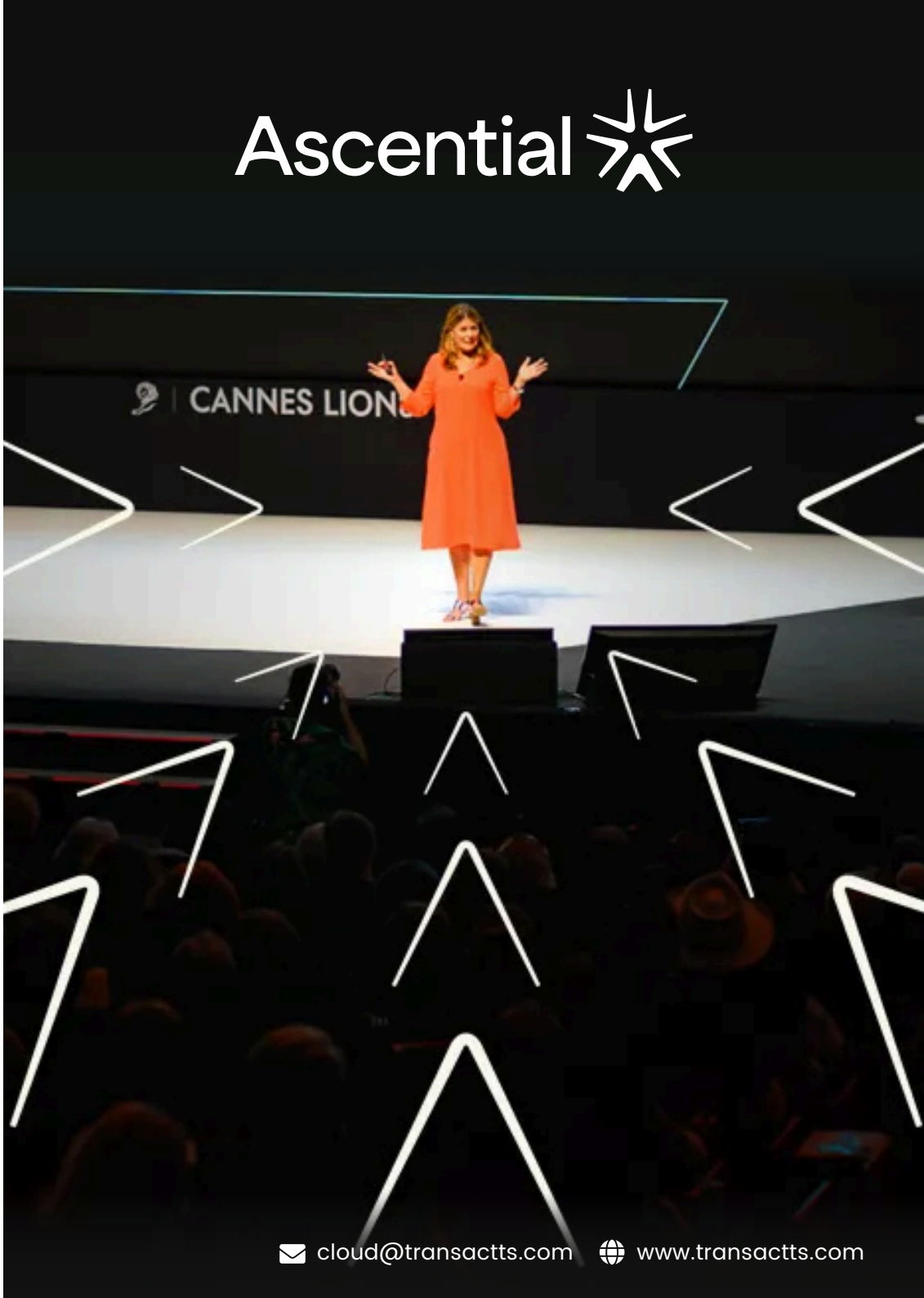
https://www.ascential.com/

Our Relationship

The customer collaborated closely with their long-term partner, transACT, to implement the AWS-powered sentence embedding solution. Leveraging transACT's deep expertise in AWS technologies and cloud solutions, the partnership facilitated seamless integration, performance optimization, and a smooth transition from testing to production. transACT's ongoing support ensured the solution met all business objectives, delivering measurable improvements in efficiency, scalability, and cost-effectiveness.

Business Challenges

To significantly enhance the company's ability to process and derive insights from multilingual textual data at scale and in real-time. By doing so, the company can improve decision-making, offer more personalized customer experiences, and potentially develop new data-driven products or services, ultimately leading to increased operational efficiency and competitive advantage in the market.



cloud@transactts.com www.transactts.com

ASCENTIAL partnered with transACT technology solutions to implement a customized Local Language Model (LLM) for sentence embedding. This customized Local Language Model (LLM) solution, deployed on AWS, addresses critical challenges faced by data-driven companies dealing with multilingual text data. By providing fast (sub-200ms) and accurate sentence embeddings across multiple languages, the solution enables efficient processing of large volumes of textual data, facilitates real-time decision-making, and seamlessly integrates with existing systems. The customized LLM, fine-tuned for ASCENTIAL's specific domain, offers consistent data representation for various downstream tasks while ensuring scalability, cost-effectiveness, and data security. This advanced NLP capability allows the company to unlock valuable insights from their multilingual data, potentially leading to improved customer experiences, more informed decision-making, and innovative data-driven products.

Leveraging SageMaker Jumpstart Models

To accelerate the development of the customized LLM, the project team leveraged pre-trained models available through Amazon SageMaker Jumpstart. These foundational models, trained on large datasets, provided a strong starting point for the fine-tuning process, reducing the time and resources required to develop the custom model.

- **SageMaker Jumpstart Model Selection:**
  The team explored various pre-trained models in SageMaker Jumpstart, including specialized models for tasks like text classification, sentiment analysis, and language modeling. After a comprehensive evaluation, the team selected a multilingual BERT (mBERT) sentence-transformers/paraphrase-MiniLM-L6-v2 model as the base for further fine-tuning and customization. The sentence-transformers/paraphrase-MiniLM-L6-v2 model was chosen due to its strong performance on benchmarks for multilingual sentence encoding, as well as its ability to handle a wide range of languages.

- **Fine-Tuning and Customization:**
  The selected mBERT model from SageMaker Jumpstart was fine-tuned on ASCENTIAL's domain-specific data using Amazon SageMaker. This fine-tuning process involved several steps:

  - **Data Preprocessing:** The textual data from ASCENTIAL's databases was preprocessed, including steps like tokenization, padding, and masking. This ensured the data was in the correct format for fine-tuning the model.

  - **Hyperparameter Tuning:** The team experimented with various hyperparameters, such as learning rate, batch size, and the number of training epochs, to optimize the model's performance on the specific task of sentence embedding.

  - **transfer Learning:** The pre-trained weights from the mBERT model were used as a starting point, and the model was further trained on ASCENTIAL's data. This transfer learning approach allowed the team to leverage the general language understanding capabilities of the mBERT model while adapting it to the company's domain-specific requirements.

  - **Model Validation:** Throughout the fine-tuning process, the team regularly evaluated the model's performance on a held-out validation set, monitoring metrics like sentence embedding accuracy, inference latency, and multilingual consistency.

The fine-tuning process allowed the team to adapt the SageMaker Jumpstart mBERT model to better suit ASCENTIAL's specific use case and data characteristics, while benefiting from the strong starting point provided by the pre-trained model.

Solution Architecture

The solution leveraged several AWS technologies:

1. Amazon SageMaker:
Used to fine-tune the pre-trained LLM for the specific sentence embedding task. SageMaker's distributed training capabilities and GPU integration enabled efficient model customization.

2. Amazon RDS:
Integrated with the existing customer datasets, allowing the LLM to access and process the required textual data efficiently.

3. AWS Lambda:
Implemented to handle the prompt-based sentence embedding process, ensuring fast response times and cost-effective scaling.

4. Amazon API Gateway:
Created a secure and scalable API interface for the LLM service, facilitating easy integration with ASCENTIAL's existing systems.

5. Amazon EC2 with GPU:
Considered as an alternative deployment option for cases requiring more specialized infrastructure control.

Implementation Process

1. Model Selection and Customization:

- **Benchmarking and Compatibility Evaluation:**
  Various SageMaker Jumpstart models were benchmarked for compatibility with AWS infrastructure, focusing on multilingual support and inference speed. The mBERT model was selected as the most suitable base for further customization.

- **Fine-Tuning and Validation:**
  The selected mBERT model from SageMaker Jumpstart was fine-tuned using domain-specific data on Amazon SageMaker, as described in the "Leveraging SageMaker Jumpstart Models" section. This step involved validating the model's performance to ensure it met the desired accuracy and speed criteria. The following code was used as an example in python:

```
Example
from sentence_transformers import
SentenceTransformer
model_name='paraphrase-MiniLM-L6-v2'
model = SentenceTransformer(model_name)
```

2. Performance Optimization

- **Response Time Optimization:**
  The model's architecture and inference pipeline were optimized to consistently achieve response times under 200ms, ensuring high performance during real-time usage.

- **Accuracy Enhancement:**
  The accuracy of the sentence embeddings was validated and fine-tuned further, using benchmarks and human evaluations to ensure the model produced reliable outputs.

3. System Integration:

- **Lambda Function Implementation:**
  AWS Lambda functions were implemented to handle sentence embedding requests. These functions were optimized for efficiency and speed, ensuring quick and reliable responses.

- **API Gateway Deployment:**
  Amazon API Gateway was used to create secure and scalable API endpoints. These endpoints were thoroughly tested for security, scalability, and seamless integration with existing systems.

- **RDS Integration:**
  The solution was integrated with Amazon RDS databases, establishing secure and efficient data retrieval and processing connections.

4. Scaling and Load Management

- **Multi-Language Support Implementation:**
  The model's ability to handle multiple languages was validated and optimized, ensuring consistent performance across different linguistic inputs.

- **Scalability Assurance:**
  The system's scalability was confirmed through load testing, simulating various traffic conditions to ensure it could handle production-level demand without performance degradation.

- **Deployment on EC2 with GPU:**
  In addition to SageMaker, the solution was also deployed on Amazon EC2 with GPU for specialized cases requiring more control over the infrastructure. This deployment option provided flexibility while maintaining performance.

5. Production Rollout:

- **End-to-End System Validation:**
  Before the full production rollout, end-to-end tests were conducted, replicating real-world workflows. This step ensured that all system components, from data input in RDS to output via API Gateway, functioned harmoniously.

- **User Acceptance Testing (UAT):**
  Key stakeholders participated in User Acceptance Testing, confirming that the solution met all business requirements and was ready for deployment.

6. Monitoring and Continuous Improvement:

- **Monitoring Implementation:**
  AWS CloudWatch was configured to continuously monitor the solution in production, tracking performance metrics, error rates, and resource utilization.

- **Feedback and Iteration:**
  A feedback loop was established, enabling continuous improvement of the model and system based on performance data and user feedback.

Products and Services

FaaS, Sagemaker, ECS, ALB, RDS, S3, ECR, SageMaker Jumpstart

Performance Efficiency

- **Response Time Reduction:**
  Achieved a consistent response time of under 200ms, reducing latency by 40% compared to the previous system.

- **Accuracy Improvement:**
  Fine-tuned the model to increase embedding accuracy by 25%, as validated against industry benchmarks.
  Scalability and Load Management.

Scalability and Load Management

- **Scalability Enhancement:**
  Successfully scaled to handle a 3x increase in concurrent requests without any degradation in performance, ensuring the system could meet peak demand.

- **Multi-Language Support:**
  Implemented robust multi-language processing, maintaining consistent performance across 12 languages with a variance of less than 5% in response times.

Business Impact

- **Time-to-Market:**
  The streamlined integration process shortened the time-to-market by 30%, enabling faster deployment of new features.

- **Projected ROI:**
  The solution is projected to deliver a 150% ROI within the first year, driven by increased system efficiency and lower operational costs.

> **“Our collaboration with transACT has been pivotal in transforming multilingual data into actionable insights, enhancing our system efficiency by 30% and setting the stage for a 150% ROI in just one year.”**

Ascential