**Solutions Introduction**

What is your solution, and how does it work? Discuss the main features.

We want to build a credit scoring system that uses both traditional bank data and alternative data sources (social media). Users fill out an online form and get instant credit assessment results. Our main innovation helps individuals with limited banking history (thin-filed customers) by allowing them to connect their social media profiles, which we analyze for credit risk using non-traditional data patterns.

Main features:

• Intuitive UI for customers to fill both traditional data and non-traditional data (via social media usernames).

• Admin panel for administrators to access each customer’s file and see explanations for each credit score in natural language.

• Traditional data is processed in the cloud using an ensemble of conventional machine learning models. (Details in Deep Dive)

• Alternative data processed in the cloud using a hybrid of large language models and traditional machine learning. (Detail in Deep Dive)

• A credit score is returned as a number ranging from 0 to 100. The higher the score, the lower the customer's credit risk.

**Impact of Solution**

How does your solution benefit society / the target audience?

Our solution addresses Vietnam's banking challenge, where most adults have bank accounts but many are new users with no credit history. Traditional banks can't assess these people for loans, so we use social media to evaluate their ability to repay.

This approach enables more people to obtain loans more quickly and affordably, while expanding access to credit cards, home loans, and business funding. Young people, rural communities, and small business owners who were previously blocked from financial services can now participate in the formal banking system.

By bringing more people into banking, our system supports economic growth through increased funding for small businesses. Banks also benefit by serving more customers safely with our instant assessment system, creating a win-win situation for both lenders and borrowers.

Why is your solution a good solution? How does it surpass existing market solutions and competitors? What is your solution’s competitive advantage / unique selling point?

Our solution leverages the abundance of social media data, requiring minimal effort from the customer. By analyzing a person's social media data, we can gain valuable insights into their spending habits, personality, level of responsibility, educational background, and other characteristics.

Our approach provides several main advantages:

• Effortless for customers - most people already have established social media accounts

• Always up to date - we can gather both historical patterns and current behavior data

• Fraud resistant - authentic social media accounts take time to build, making them difficult to fake.

There are many competitors in the alternative data space, including global players like LenddoEFL, RiskSeal, and FinScore, as well as local solutions like FPT AI Credit Scoring. However, our solution offers significant advantages over these existing competitors. Most global competitors are not specifically tailored for the Vietnamese market, with solutions like LenddoEFL focusing on psychometric data and FinScore specializing in telecom data for the Philippines and Indonesia rather than understanding Vietnamese users' social media behaviors and cultural patterns. While FPT AI Credit Scoring targets the Vietnamese market, it is still new and unproven, operating primarily as a black box where the results lack explainability for lending decisions. Our internal solution is superior because it prevents sensitive customer data from being passed to third-party providers, ensuring that we maintain complete control over the evaluation process and allowing for better transparency, customization, and data security.

**Deep Dive into Solution**

Detailed write-up of solution, e.g., level 2 data flow diagram. Include any supporting features if necessary.

Our credit assessment system operates through two application flows depending on the customer's credit history. For customers with sufficient traditional credit data, the regular application flow processes their information through our decision tree algorithm to generate a score between 0 and 1. For thin-file customers with limited or missing traditional data, we implement a specialized thin-file flow that first processes available traditional data and then requests supplemental social media information to provide an enhanced score, accompanied by a bonus for non-traditional data.

**Data Routing**

The system begins by receiving raw input data from consumers and immediately routes this information through our data classification engine. Traditional structured data (credit history, financial records, demographic information) is directed to the Traditional Data Store. In contrast, non-traditional, unstructured data (such as social media profiles, posts, and network connections) is stored in the Non-traditional Data Store. Simultaneously, the system performs a data volume assessment to determine the relative abundance of non-traditional data, which directly influences the final weighting parameters.

**Traditional Flow**

The system processes traditional credit data using an ensemble approach with two tree-based models: LightGBM and decision tree. This approach helps avoid overfitting and provides more stable predictions across different market conditions. We will consider additional models if time permits. The traditional scoring component processes structured credit data, demographic information, and historical financial behavior to generate a baseline traditional score, ranging from 0 to 1.

**Non-Traditional Flow**

Non-traditional social media data is passed through a combination of Large Language Model (LLM) evaluation and decision tree processing. Our non-traditional data sources include basic profile information, such as education and family details, user posts containing both image and text content, and network connections represented by user IDs of friends and connections.

For social media posts, we employ LLM-based rubric scoring that evaluates content across multiple categories, then aggregates these scores using a weighted average with recency decay, where more recent posts receive higher weights and different categories are weighted based on their relevance to creditworthiness. Network analysis follows a recursive approach, scoring connections using the same methodology applied to the primary user, excluding their network to prevent circular references. The final social credit score is calculated by combining three components: Basic Information Score, Posts Score, and Network Score.

**Weight Calculation**

The system calculates weight parameters based on the volume and quality of available non-traditional data. Higher volumes of non-traditional data result in increased weighting toward the LLM-generated score, whereas limited non-traditional data maintain a heavier reliance on traditional scoring methods. This calculation incorporates stochastic elements to prevent overfitting to either approach and encourages the user to provide additional non-traditional data.

**Final Scoring (Stochastic Weighted)**

Our final scoring mechanism combines both traditional and non-traditional scores through a dynamic weighted formula:

Final Score = α × Non-traditional Score + (1-α) × Traditional Score

Where α represents the dynamically calculated weight parameter that increases with non-traditional data volume. This stochastic weighted combination ensures that:

• Customers with rich non-traditional data benefit from enhanced LLM insights

• Customers with limited non-traditional data rely primarily on proven traditional methods

• The system maintains scoring consistency across different data availability scenarios

**Architecture of Solution**

How does your solution make use of AWS infrastructure? Do not just describe the services you are using, but explain how you are using the services in your solution.

How are you integrating and combining the different AWS services to make your solution seamless? Include architecture diagrams. Refer to this link for examples.

**Architectural Flow & Service Integration**

Our proposed system utilizes key Amazon and AWS services for rapid prototyping with options to scale and ensure security. As a cloud-native and serverless solution, it aims to be both cost-effective and efficient. The core service will be integrated through APIs and event-driven triggers. The goal is to provide a seamless, real-time experience for the users from uploading supporting documents to receiving an explainable credit score.

To achieve this, the team will implement the following workflow:

**User Interface & Document Upload**

A simple front-end will be built using React + TypeScript, hosted on Amplify for an AWS-native solution that is also scalable. It has built-in options for CI/CD pipeline and easy integration with GitHub version control. This enables quick feedback and updates, ensuring the UI is highly performant and always available. The user will upload their financial and any supporting documents through this secure interface.

**API Gateway & Lambda Trigger**

Through Amazon API Gateway, the front-end will send secure requests, which will be validated and authorized, then routed to trigger the central orchestrator: AWS Lambda.

**Data Archive & Orchestration**

The API Gateway trigger will invoke the primary AWS Lambda function, which orchestrates the entire credit scoring workflow. Firstly, the function ensures that the user’s raw, unmodified documents are safely stored in Amazon S3—a durable data lake and archive that serves as a single source of truth for future auditing and model retraining.

**Structured Data Analysis**

For quantitative data analysis, the Lambda function extracts numerical features. It compiles them with existing traditional financial data (if any) as feature vectors for a LightGBM model, which the team will fine-tune and further train through an Amazon SageMaker notebook instance. The model itself is hosted on a real-time Amazon SageMaker endpoint, which will return an initial traditional score to the Lambda function.

**Unstructured Data Analysis**

The Lambda function will call Amazon Textract to perform Optical Character Recognition (OCR), extracting raw textual content from uploaded images and PDFs. The function will also compile this with any obtainable social media information through user-provided handles. This formatted text data will then be passed to BedRock and processed with a non-restricting LLM model, such as DeepSeek. With a carefully engineered prompt, the LLM model can analyze the text for qualitative signals of financial stability. The output will be structured JSON, including the analysis and a non-traditional score, as actionable items for the next steps.

**Explanation Generation**

After receiving back the outputs from SageMaker and Bedrock, as well as self-processed weight-by-volume parameters, the Lambda function will aggregate all quantitative data to a final credit score. It then makes a final call to Bedrock with the full context of the final score and qualitative analysis to generate a simple, understandable explanation in natural language of why the score was assigned.

**Storing & Retrieving Final Results**

The complete result, a comprehensive JSON file containing the final score, explanation, output from all models, and links to the S3 raw files, is stored in Amazon DocumentDB, a robust NoSQL data lake solution for storing and retrieving operational data. By storing the complete result, the service can retrieve a user’s credit history without any recalculations or re-analysis.

**Returning Final Result**

As the final step, the Lambda function returns the stochastically calculated credit score and its explanation to the API Gateway, which then passes it on and displays it on the front-end.