Project 1: Prison Management System Modernization

1. Project Overview:

You are stepping into the final year of a $350M, five-year AI modernization project aimed at consolidating seven legacy systems into a single solution for a large-scale prison management system. The system has implemented several AI initiatives, including a recommender system for inmate housing, a security compliance and anomaly detection system, and a predictive staffing model. The client and AI consulting firm are not satisfied with the current progress due to project delays, budget overruns, poor system performance, and issues with incorrect SQL queries and system latency.

Your role is to assess the situation, identify key problem areas, and develop a recovery strategy for the project's final year. You will also need to propose strategies for the upcoming five-year operations and maintenance contract, which will involve continuous improvement of the AI systems.

Current Challenges:

- **SQL Query Errors:** The system frequently returns inconsistent prisoner data, leading to incorrect schedules for court appearances and medical appointments.
- **Poor System Latency:** Delays in judicial updates to prisoner status are causing real-time decision-making issues, particularly for critical events like prisoner transfers or releases.
- **AI Model Underperformance:** The recommender system for assigning inmates to facilities, the anomaly detection model for security compliance, and the staffing model to predict the staffing requirements are not delivering the expected results in terms of accuracy and speed.
- **Budget Overruns:** The project has already exceeded its budget by $20M with one year left, and there is pressure to avoid further overspending in the final year.

Key Project Details:

- **Budget:** $350M (currently over budget by $20M)
- **Timeline:** 5 years (currently in year 5)
- **System Scale:** Manages 60,000-70,000 prisoners and less than 500,000 total records.
- **Technology:**
    - **Front-end:** Pega (for managing prisoner operations)
    - **Back-end:** AWS (cloud infrastructure for compute, storage, and AI models)
    - **Previous Data Solution:** Oracle Data-Mart (legacy system)

AI Initiatives Implemented:

1. **Inmate Recommender System:** Recommends prison facilities for new inmates based on factors such as transportation costs, proximity to courts, and medical needs (e.g., maternity care).
2. **Security Compliance and Anomaly Detection:** Monitors security inspection rates and detects anomalies, such as missed inspections or contraband incidents.
3. **Staffing Prediction System:** Uses AI to predict staffing requirements for different facilities based on inmate population and operational needs.

2. Current System Architecture Diagram

Component	Technology	Description
Front-end	Pega	User interface for prison management operations, data entry, and reporting.
Back-end	AWS Cloud (EC2, S3)	Cloud-based compute and storage handling AI models, databases, and business logic.
Database	AWS RDS (PostgreSQL)	Cloud-based relational database to manage prisoner information.
Previous Database	Oracle Data-Mart	Legacy database that previously stored prisoner records, now replaced by AWS RDS.
Inmate Recommender	AWS SageMaker	AI-driven recommendation system for optimal prisoner placement.
Security Monitoring	Custom AI Model	Monitors and reports security compliance and anomaly detection.
Staffing Predictions	AWS Lambda Functions	Predicts staffing needs for various prison facilities.

3. Original Staffing Breakdown (Year 1 to Year 3)

Role	Number of Staff	Primary Responsibilities
Project Manager	2	Overall project coordination, client communication, and delivery.
AI Architects	3	Design of AI systems (recommender, security monitoring, staffing).
Data Engineers	6	Data integration, SQL query development, and data migration.
DevOps Engineers	4	Cloud infrastructure (AWS), system integration, and deployment.
Front-End Developers	4	Pega development for user interface and workflows.
Back-End Developers	5	AWS service implementation, system logic, API development.
Quality Control (QC) Analysts	5	System testing, bug tracking, data validation, and model accuracy.
Data Scientists	5	AI model training, optimization, and algorithm development.

Role	Number of Staff	Primary Responsibilities
Business Analysts	3	Requirements gathering, process mapping, and client liaison.
Cloud Security Specialists	3	Security protocols, compliance checks, and anomaly detection.
Solution Architects	2	System design, architecture decisions, and client technical liaison.

Summary of Original Team:

- **Total Team Members:** 42
- **Focus Areas:**
  - AI model development for inmate recommendations, security monitoring, and staffing predictions.
  - Data migration from Oracle Data-Mart to AWS.
  - Front-end integration with Pega.
  - Regular quality control (QC) testing to ensure accuracy and system reliability.
  - Strict adherence to budget and timeline, given the project's complexity.

4. Staffing Cuts Due to Budget Overrun (Years 4 to 5)

Role	Number Cut	Impact
Quality Control (QC) Analysts	4	Reduced system testing and oversight, leading to increased SQL query errors and data inconsistencies.
Data Engineers	3	Fewer resources for optimizing SQL queries and integrating data sources, resulting in performance issues.
Cloud Security Specialists	2	Less focus on security monitoring, leaving vulnerabilities in the security compliance system.
Front-End Developers	2	Slower development of Pega front-end features and delay in integrating AI recommendations.
Business Analysts	1	Reduced communication with the client, leading to misaligned project features and priorities.
DevOps Engineers	2	Reduced capacity for maintaining AWS infrastructure, leading to higher latency and AWS inefficiencies.

5. Performance and Budget Reports

Performance Report:

- **Recommender System:**
  - Accuracy: 65% (Target: 85%)

- Processing time per recommendation: 7 seconds (Target: 3 seconds)
- **Security Compliance Monitoring:**
  - Incident Detection Rate: 75% (Target: 95%)
  - Latency in reporting incidents: 12 minutes (Target: 5 minutes)
- **Staffing Prediction Model:**
  - Staffing prediction accuracy: 60% (Target: 80%)
  - Frequency of incorrect predictions: Weekly issues
- **Prisoner Records Update:**
  - Latency in updating prisoner status: 12 hours (Target: 15 minutes)

Budget Report:

- **Initial Budget:** $350M
- **Current Spend:** $370M
- **Estimated Final Year Spend:** $60M
- **Overruns:**
  - **AI Development Costs:** + $20M
  - **Cloud Compute Costs:** + $15M due to inefficiencies in resource use (overuse of AWS GPUs for non-intensive tasks)

6. AI Initiative Summaries

AI Initiative 1: Inmate Recommender System

Goal: Reduce transportation costs by recommending nearby prison facilities for court appearances or specialized needs (e.g., maternity care).

- **Technology:** AWS SageMaker, Pega front-end for viewing recommendations.
- **Challenges:** The model is slow in delivering recommendations, and the accuracy of recommendations is below the required threshold.

AI Initiative 2: Security Compliance Monitoring

Goal: Detect anomalies in prison inspections, contraband findings, and security issues.

- **Technology:** Custom-built AI model on AWS.
- **Challenges:** Latency in anomaly detection has led to delayed responses to security threats. The incident detection rate is below expectations.

AI Initiative 3: Staffing Prediction

Goal: Predict optimal staffing levels across various prison facilities.

- **Technology:** AWS Lambda-based predictive model.
- **Challenges:** Frequent inaccuracies in predictions have caused both under and over-staffing issues at several facilities.

7. Client Feedback Report

Client Feedback Highlights:

"The new system's SQL queries are returning inconsistent prisoner records, leading to confusion in court appearance schedules."

"The inmate recommender system was supposed to help us reduce costs and make smarter decisions, but it feels like we're constantly second-guessing its suggestions. We've had multiple instances where inmates were sent to the wrong facilities, leading to higher transportation costs and unnecessary logistical headaches."

"Judicial updates are delayed by up to several hours, which is unacceptable given the real-time nature of prison management."

"It's clear this system was built for something much bigger than we needed. We don't have millions of transactions like commercial company. The complexity has made it harder for our staff to use, and we're not seeing any benefit from all this extra horsepower."

"The AI models, especially the recommender system and staffing predictions, are not accurate enough and are causing operational disruptions."

"One of our corrections officers had to escort a prisoner to a prenatal doctor appointment, which took most of the day, when there was another facility about five miles away that had maternity care facilities inside the prison. We can't afford to waste the time of corrections officers, it is costly."

"We are already $20M over budget, and costs keep increasing. This project is becoming unmanageable."

"The whole point of the security monitoring system was to catch potential issues before they become serious, but by the time we get an alert, the incident has already happened. We've had inspections missed and contraband found weeks after it should've been flagged. We might as well go back to manual logs at this point."

"We were promised a predictive staffing model that would help us allocate resources efficiently, but instead we're constantly either overstaffed or scrambling to cover shortages. It's causing confusion on the ground, and it's affecting the safety and operations of our facilities."

"The delay in updating judicial rulings is unacceptable. When a prisoner is set for release, we need that information immediately, not hours later. This lag has already led to one wrongful detention over a holiday weekend, and it's completely eroded our trust in the system."

"We're paying a premium for a system that can't even keep up with basic performance requirements. The latency, the constant errors, and the overall inefficiency make it feel like we've thrown money into a black hole."

"I'm not sure what happened during development, but we're finding errors in basic data entries that should have been caught. How can we trust this system when we can't even get prisoner records or staff schedules right? It's been one issue after another, and it feels like nobody's checking the quality before we're handed these updates."

"The project team seems disconnected from our actual needs. We've tried explaining the problems we face, but their solutions feel like they're tailored for a completely different type of organization. It's as if they're more interested in showing off fancy features than solving our real-world issues."

"I have serious concerns about handing over this system for operations and maintenance. If we can't even get the basics right now, how are we supposed to trust that this will function smoothly for the next five years? I don't want to be stuck in another expensive mess."

8. Project Deliverables:

As the sole deliverable for this project, learners are expected to prepare a report comprising responses to the following:

1. **Assessment Report**
    - **Root Cause Analysis**: A detailed analysis of why the project has failed to meet its schedule and budget.
    - **AI System Gaps**: Identify specific issues in the AI system architecture, algorithm selection, query performance, and resource allocation.
    - **Technical Analysis**: Evaluate the tradeoffs between computational complexity, memory, multi-threading, and performance.
2. **Recovery Strategy:**
    - **Resource Optimization**: Recommendations on how to optimize memory, query expressiveness (SQL vs. alternatives), and compute power (CPU vs. GPU/TPU usage).
    - **AI Model Improvements**: Propose improvements to the recommender system, anomaly detection, and staffing models to reduce errors and improve system accuracy.
    - **Timeline & Budget**: Present a realistic timeline and budget strategy for completing the project on time while addressing current deficiencies.
3. **Recompete Strategy:**
    - **Client-Centered Recovery Strategy**: Identify key issues such as budget overruns, system complexity, latency and underperforming AI features and propose specific fixes for these issues. Outline steps for rebuilding trust through improved communication and early delivery of critical improvements.
    - **Continuous Improvement Plan**: Develop a strategy for continuous system enhancement over the five-year operations and maintenance phase. Explain how technical improvements will address the client's key concerns.
    - **Competitive Differentiation**: Explain why your team's approach will win the re-compete. Highlight key strengths such as experience in system optimization and ability to deliver fast improvement. Define performance indicators to measure success and ensure transparency with the client.

9. Submission Guidelines:

- The solution should consist of a document file converted to PDF. The document should not contain more than 10 pages. In case you want to add a few additional details you can add them to the appendix.

<please refer to the next page for additional details>

Additional Details

- **Database System**
  a. It contains various types of records such as prisoner details, facility details, courts, transportation costs, inspection logs, and staffing-related details. Earlier, the Oracle Data-Mart system was used to maintain the database but as the number of records and facilities increased it was getting difficult to maintain the system due to various reasons like scalability, high infrastructure, and maintenance costs due, to lack of integration with modern tech stack. Due to this, the data was migrated to AWS RDS.
  b. During data migration, data engineers faced various challenges:
    i. The most important challenge was Oracle uses proprietary SQL features, which may not work in PostgreSQL without adjustments, leading to incorrect queries.
    ii. Schema differences and normalization issues.
    iii. Optimize queries for PostgreSQL
  c. It has been found that the facility updates and policy updates are neither being informed nor getting updated in the database system leading to the AI applications working on potentially incorrect data and rules.

- **Inmate Recommender System:**
  a. Inmate assignments are done on a daily basis including assigning new inmates to the right prison facility as well as moving current inmates to newly vacated facilities
  b. Earlier assignments were manual and used to consume a lot of human resources to just manage this, while there were higher costs incurred due to the low frequency of reassignments
  c. When this system was implemented, a one-time activity of reassignment was done to achieve two objectives
    i. minimize unfavorable outcomes like severe health impact due to delay in reaching the emergency medical center, and
    ii. minimize operational costs incurred in moving inmates to various facilities across a month
  d. To monitor the performance of the AI system, two things were done
    i. a prison-wide human reporting SOP (Standard Operating Procedure) was set up to capture instances where a prison personnel marked an inmate as potentially in the wrong prison facility
    ii. a monthly cost reporting was set up to track costs at inmate level
  e. Accuracy was defined as the percentage of inmates where neither a wrong assignment is reported by personnel nor the total cost of that inmate is higher than the set benchmark e.g. accuracy of 85% will mean that 85% of the inmates have neither a wrong assignment reported nor is their cost above the set benchmark
  f. Post-implementation, while there was some immediate benefit realized of having to hold lower human resource capacity, the ongoing performance tracking revealed that the accuracy was hovering around 65% only.

- **Security Compliance and Anomaly Detection**
  a. The important part of the security system is a vision-based model to monitor the activities in the facility and identify suspicious activity.
  b. Additionally, it takes data like motion sensors, security inspection logs,human-reported incidences, metal detectors, door sensors, readings from item-checking machines, etc.

c. The system then flags the incidences where suspicious activity is detected, missed inspections, and contraband activities such that the security personnel can take corrective action. It also highlights the incidences detected during Item checking such as the unusual weight of the item.
d. The model performance is calculated based on human feedback about the number of incidences falsely flagged or the number of incidences that were not flagged by the system. This is mostly based on human feedback about the incorrect flags and non-predicted actual incidences
e. The accuracy is calculated as the percentage of incidences correctly reported out of the total number of actual incidences.
f. The incident Detection Rate is 75% and the Latency in reporting incidents is 12 minutes
g. The system processes a good amount of vision data.
h. Security logs form a significant part of the data but usually come with missing entries. It involves significant data preprocessing.
i. On average 2% of the total activities are detected as suspicious activities or incidences

- **Staffing Prediction System**
  a. The prison has two types of staff those are full-time staff and part-time staff. Correctly identifying the staffing requirement, particularly part-time staffing helps ensure that the required work is getting done with the optimum number of staff.
  b. It predicts the staffing requirement for the facilities based on various factors such as inmate population (demographics and type), routine cleaning, peak visitation period, part-time worker details, etc
  c. It has been observed that during the winter season, the facility required more workers due to more health-related complaints from the inmates.
  d. The staffing requirements are usually predicted at the weekly level so as to plan the coming weeks in advance.
  e. Frequent inaccuracies in prediction have caused either understaffing or overstaffing issues at several facilities.
  f. The model's performance is assessed by calculating the percentage of weeks in which its staffing predictions are accurate within a 10% margin of error compared to actual staffing needs, relative to the total number of weeks analyzed (usually calculated quarterly).