US32 - US32 Latency Cost Budgets

As a product owner, I want visibility into cost and latency budgets.  
  
Details:  
- Track sanctions API call costs, LLM token spend.  
- Alert when exceeding monthly thresholds.  
- Compare p95 latency against SLAs.