Feature Specification: Predictive AI Workflow Trigger

PRD Reference: PRD-2025-11-001
Lead Engineer: [Engineer Name]
Date: 2025-11-15

1. Overview and Rationale

This document details the technical implementation of the **Predictive AI Workflow Trigger (PAWT)** feature. PAWT uses a proprietary time-series model (NexaCore-TSM v1.2) to predict system demand spikes and automatically trigger scale-up actions or preemptive reporting workflows for enterprise clients.

Goal: Proactively mitigate service degradation before a human can react.

2. Technical Design

2.1 Service Architecture

The PAWT will be implemented as a new microservice (svc-pawt) running within the core K8s cluster.

- **Language:** Python (due to existing AI/ML ecosystem).
- **Data Source:** svc-pawt will subscribe to the Kafka topic: core-telemetry-metrics (low-latency stream).
- **Output Action:** Upon prediction threshold breach, svc-pawt calls the **Workflow Engine API** (/v1/trigger) with the payload: {"workflow_id": "[ID]", "trigger_source": "PAWT", "prediction_confidence": [Value]}.

2.2 AI Model Integration

- **Model Name:** NexaCore-TSM-v1.2 (Trained on 12 months of aggregated client telemetry data).
- **Prediction Input:** Requires 5 minutes of continuous data: Average CPU utilization, API call volume (per second), and database connection pool usage.
- **Prediction Output:** A 1-hour lookahead prediction for API call volume.
- **Trigger Threshold:** The workflow is triggered when the predicted API call volume exceeds of the current cluster capacity's historical maximum (P99).

2.3 API Endpoint Details

Method	Endpoint	Description	Authentication

POST	/api/v1/pawt/status	Allows system to check the current health and prediction score.	Internal Token

3. Data & Storage Requirements

- **Storage:** svc-pawt requires a PostgreSQL database instance (db-pawt) to store prediction history and model calibration data.
- **Data Retention:** Prediction history older than 90 days must be automatically archived to cold storage (S3 Glacier) due to volume.

4. Testing Plan

4.1 Unit Testing

Standard unit tests for data transformation and API calls.

4.2 Integration Testing

Verify that svc-pawt successfully subscribes to the Kafka topic and successfully sends the trigger payload to the **Workflow Engine API**.

4.3 Load Testing

Simulate a rapid 5-minute surge in core-telemetry-metrics data to ensure the service can process the stream and generate the trigger event within **10 seconds** of the threshold breach.