Contextual Research Plan[1] into chosen problem

Step 1: Define Areas for Research

• Initial Problem Statement (Unchanged)

Organizations increasingly rely on cloud infrastructure, but managing and optimizing costs across compute, storage, and networking resources is complex and often inefficient. Current tools provide dashboards and static recommendations but fail to adapt to dynamic workloads and organizational priorities. The problem is how to design an AI-powered system that combines optimization algorithms with LLMs to provide intelligent, conversational cloud cost optimization strategies

- What is Known: (Updated)
  - Cloud spending continues to rise with studies estimating that 20-30% of cloud costs are wasted due to inefficiencies (Flexera, 2023)
  - Existing cost-management tools (AWS Cost Explorer, Azure Advisor recommender) provide recommendations but are limited in personalization and adaptability.
  - Optimisation strategies like rightsizing and workload tuning can save up to 65% of costs (Velonov et al. 2023)
  - Graph-based models (Khan et al, 2024) show that cloud resources can be modeled mathematically to find optimal deployment,
  - FinOps driven ML forecasting approaches focused on ML based forecasting can help organizations proactively reserve resoources and save up to 90% in costs (Nawrocki et al, 2024)
  - Foundational models (FMs) extend beyond narrow AI roles as they integrate multimodal data, reason with chain of thought, and enable cross-domain decision making

- What is Assumed (Refined)
  - Organizations prefer automation that both monitors and recommends real-time, actionable cost-saving strategies.
  - LLMs can bridge the gap between optimization models and human operators by generating interpretable recommendations
  - Accesss to cloud usage logs and billing data is feasible (via APIs from AWS,Azure,GCP).
  - Trust and explainability are critical for adoption as users will not blindly accept AI-generated cost changes without transparency

- What is Unknown: (Expanded)

[1] This template follows the steps outlined in Lecture 3 (Sept 10), and also in Section 2.1 Contextual Research Plan of the textbook [Kumar, V. (2012). *101 design methods : a structured approach for driving innovation in your organization* (1st edition). Wiley.]

- How accurate and reliable AI-driven optimizations will be across diverse workloads(batch vs real-time, seasonal vs steady).
- Whether organizations will trust LLM-driven recommendations in sensitive financial decisions
- Best methods to validate AI-generated strategies against historical usage and human expert intuition.
- Which performance metrics matter most to stakeholders (cost savings, efficiency, reduced downtime, carbon footprint)
- Best practices to valiate AI-generated strategies (historical benchmarking, expert oversight, controlled pilots).
- The balance between cost optimization and other objectives (performance, compliance, availability).

- Decompose the Problem into Aspects:

- Processes & Steps:
  - Data ingestion: Collect cloud billing and usage logs.
  - Optimisation: Algorithms generate cost-efficient allocation strategies
  - Interpretation: LLM translates outputs into actionable insights
  - Recommendation: System suggests scaling, rightsizing or switching resources
  - Verification: Users review and approve before implementation
  - Feedback loop: System learns from accepted/rejected suggestions
- People & Roles:
  - Cloud Engineers: Need precise, technical recommendations for scaling and provisioning.
  - Finance Teams (FinOps): Care about budget compliance, savings and ROI
  - Developers: Want performance maintained without disruptions
  - Executives: Look for transparency and measurable savings
  - AI/Optimization Specialists: Design algorithms and fine-tune LLM integrations

- Interactions:
  - Engineers and finance teams interact with the system via dashboards and conversational queries.
  - AI agents interact by combining optimization outputs with LLM explanations.
  - Feedback cycles allow humans to refine the system’s decision-making logic.
  - Cross-department collaboration is facilitated by natural language recommendations that both finance and engineering can understand.

- Environment & Context:
  - Technological: Many organizations now use multi-cloud setups, serverless platforms, and container orchestration (e.g., Kubernetes). While these increase flexibility and scalability, they also make tracking and optimizing costs far more complex.
  - Organizational: Companies face pressure to cut IT spending while still expanding cloud-based services. This creates tension between performance requirements and financial discipline.
  - Economic: Cloud usage charges can swing dramatically month to month due to variable workloads. Such unpredictability makes it difficult for businesses to plan and stick to budgets.
  - Social: Stakeholders increasingly expect sustainable practices. Cloud optimization is not only about saving money but also about minimizing energy use and reducing carbon footprints.

- Analogous Contexts:
  - Smart Grids: Optimizing power distribution mirrors dynamic workload balancing in cloud.
  - Portfolio Management: Allocating assets under constraints is similar to distributing workloads under budget limits.
  - Logistics Optimization: Routing shipments efficiently is analogous to allocating compute/storage.
  - Business Intelligence: Turning raw data into insights parallels turning billing data into actionable strategies.

Step 2: Define Sources and Methods

- Key Publications / Media / Databases: (Revised to reflect academic sources)

Key publication:

- Khan et al (2024) – Graph-based cloud cost optimization
- Nawrocku et al. (2024) – FinOps forecasting for HPC
- Jincai et al. (2025) – Foundational models for intelligent decision-making
- Flexera (2023) – State of Cloud Report
- Velinov et al. (2023) – Rightsizing and workload tuning

- Key People / Stakeholders:
  - Cloud Engineers / DevOps Teams: They deal with scaling and infrastructure costs

  - Finance / FinOps Professionals – track budgets and need cost accountability.
  - Executives / IT Directors – seek transparency in cloud spending and long-term efficiency.

  - Cloud Service Providers (AWS, Azure, GCP) – provide APIs, optimization tools, and best practices.
  - Researchers in Optimization & AI – publish new techniques combining algorithms with AI systems.

- Methods for Collecting Information:
  - Literature Review: Search keywords such as *cloud cost optimization, cloud resource allocation, FinOps, stochastic optimization cloud, LLMs for cloud management, AI cost governance*.
  - Competitor Benchmarking: Review tools like AWS Cost Explorer, Azure Advisor, and GCP Recommender to identify gaps (e.g., personalization, explainability).
  - News Media & Reports: Use Gartner, Flexera, and McKinsey articles for industry trends and cost waste statistics.

Step 3: Create a Research Plan Timeline

Create a rough plan of the tasks to be completed for your Individual Project. Consider that you are working towards Milestone 3 (due October 10), so all work will need to be completed in just over 4 weeks.

Task	Method	Time estimate
Collect academic papers and industry reports	Literature review via Google Scholar, arXiv,Flexera,Gartner	1 week
Extract insights on problem history, causes and impact	Synthesis of literature + industry reports	3-4 days
Bnenchmark existing cloud cost tools (AWS,Azure, GCP)	Competitor analysis through whitepapers, documentation	3 days
Develop visuals (system map of AI-poweredoptimization, comparison table)	Diagramming + tabular analysis	3-4 days
Draft the executive summary (impact, existing solutions, insights)	Write-up based on synthesized findinfs	4-5 days
Submit report	Final polish and formatting	1 day

Repository Link - kasethi23/MSE-302-Individual_Project

Summary of Research Activities So Far

1. Paper 1- *Cost Modelling and Optimisation for Cloud (Khan et al. 2024)*
    - The paper addresses the challenge of rising cloud computing costs, especially in multi-cloud and hybrid environments.
    - It proposes a graph-based approach to model cloud cost elements (compute, storage, network) as nodes and edges.
    - The method incorporates cost, utilisation, performance, and availability to solve the constraint problem of cost optimisation.
    - Evaluations across different user scenarios show the approach effectively optimises cost and scalability, reducing expenses significantly.
    - The approach helps organisations make informed decisions on cloud resource placement, balancing cost and QoS trade-offs.
    - Future directions include extending QoS factors, improving the user interface, and integrating with systems like Kubernetes for resource allocation.

2. Paper 2 - *FinOps-driven optimization of cloud resource usage for high-performance computing using machine learning (Nawrocki et al. 2024)*
    - Introduces a Cloud Resource Usage Optimization System (CRUOS) that predicts long-term cloud resource usage for HPC applications.
    - Uses statistical models, XGBoost, neural networks, and the Temporal Fusion Transformer for forecasting CPU and RAM needs.
    - Achieved up to 31.4% improvement in prediction accuracy with TFT compared to baselines, especially for chaotic resource usage patterns.
    - Developed a Reservation Module to translate predictions into dynamic resource reservation plans aligned with FinOps principles.
    - Qualitative and quantitative assessments show significant cost savings and better SLA compliance through proactive forecasting.
    - Emphasizes that the best predictive model may not always yield the most effective reservation plans, with neural network methods often being more cost-efficient.
3. Paper 3- *Cost Optimization for Cloud Storage from User Perspectives : Recent Advances, Taxonomy, and Survey (Mingyu et al 2023)*
    - The paper surveys techniques and challenges in reducing cloud storage expenses.

- It notes that up to 30% of cloud costs are wasted, driving the need for optimization
- Key strategies include **compression, deduplication, service selection, multi-cloud approaches, edge storage, and caching**.
- The authors propose a taxonomy of optimization methods: storage efficiency,single cloud,multi-cloud, edge-cloud collaboration
- Open challenges include balancing cost and performance, vendor lock-in, and identifying future research directions for cost-effective user-centric storage

4. Paper 4 – *Foundation models and intelligent decision-making: Progress,challenges, and perspectives (Jincai. et al 2025)*

- The paper traces the evolution of decision making systems showing how foundational models (FMs) unify multimodal data to enable context aware decision-making across domains
- It emphasized that FMs can act not only as agents (planner,decision-maker,actor) but also as environments,encoders, and human-machine interactors, expanding their roles beyond traditional AI systems
- Key advances include the integration of reinforcement learning and FMs, multimodal architectures, and prompt-engineering techniques (eg chain-of-thought, tree of thought) that enhance reasoning and adaptability
- Applications span healthcare,finance and science where foundational models can provide cross domain knowledge integration for complex decision making