Part 1 – Problem Brainstorm (max 1 page)

Provide a blurb about your evaluation of the 5 (consider using the various factors discussed in Lecture 2, including factors in the linked Miro board) and rationale for selecting one.

1	Natural Language Query for Analytics- A system that allows users to ask questions in plain language and receive accurate, contextual analytics could democratize access to data as opposed to writing SQL queries.
2	AI-Powered Shopping Layer-. Designing ranking and recommendation systems that are optimized for conversational AI discovery as opposed to traditional SEO.
3	Intelligent Healthcare Symptom Checker- An AI tool that maps natural language symptom descriptions to possible conditions and provide evidence-based next step guidance could improve healthcare access and reduce unnecessary visits
4	AI-Powered Cloud Optimization – A system that combines optimization algorithms with LLMs to analyze cloud usage logs and advise on optimizing costs.
5	AI-Powered Research Synthesis- A multi-agent system that automates literature reviews by assigning sub-agents to read, extract, and analyze findings from thousands of research papers, with a coordinator agent synthesizing results into coherent summaries

Across the five brainstormed ideas, each can be attributed to different and management engineer principles and their feasibility can be considered across multiple factors. Natural language query democratizes analytics by improving decision-making efficiency, while AI-powered shopping anticipates shifting digital markets. The healthcare symptom checker addresses a critical societal issue but faces ethical and regulatory hurdles, and AI-powered Research synthesis supports productivity but is a very niche problem. The most compelling is AI-Powered Cloud Optimization, which is feasible due to accessible cloud usage datasets and a pressing industry need. By combining optimization algorithms with LLMs, it addresses wasted cloud spending and champions optimization which is a core concept in management engineering.It can go on to be a excellent capstone project.

Part 2: Contextual Research Plan[1] into chosen problem (max 5 pages)

Step 1: Define Areas for Research

• Initial Problem Statement:

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:
  - 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 methods can achieve large savings as case studies report up to 65% reduction in costs when strategies like rightsizing, workload tuning and instance optimisation are applied (Velonov et al. 2023)
  - Optimisation algorithms (linear programming, stochastic modelling, reinforcement learning) have long been used in resource allocation problems and are increasingly applied to cloud workloads.
  - FinOps (Financial Operations) is emerging as a discipline to integrate financial accountability with engineering practices, stressing the need for proactive, data-driven optimisation of cloud resources (Nawrocki et al., 2024).
  - LLMs can interpret natural language queries and translate optimization outputs into actionable recommendations, bridging the gap between technical optimization and decision-making.

- What is Assumed:
  - Organizations want automation that not only reports cloud usage but also suggests actionable actionable cost-saving strategies in real time
  - 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).

[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.]

  - Trust and explainability are critical for adoption as users will not blindly accept AI-generated cost changes without transparency
  - Multi-cloud environments add complexity but also make optimization more valuable

- What is Unknown:
  - 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)
  - 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:

Academic Literature:

- *Cost Modelling and Optimisation for Cloud* (Khan et al., 2024).
- *Cloud Cost Optimization: A Comprehensive Review of Strategies and Case Studies* (Deochake, 2023).
- *Optimization of Cloud Costs* (Velinov et al., 2023).
- *FinOps-driven Optimization of Cloud Resource Usage* (Nawrocki et al., 2024).

Industry Reports:

- Flexera *State of the Cloud Report* (2023).
- Gartner *Forecasts on Cloud Spending* (2023–2024).

Media / Blogs:

- AWS, Azure, and GCP cost management whitepapers.
- FinOps Foundation publications.

- 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

APA Citations

Flexera. (2023, April 5). *Cloud computing trends and statistics: Flexera 2023 State of the Cloud Report* [Blog post]. Flexera. https://www.flexera.com/blog/finops/cloud-computing-trends-flexera-2023-state-of-the-cloud-report/ Flexera

Velinov, A., Zdravev, Z., & Nikolova, A. (2023). *Optimization of cloud costs*. *South East European Journal of Sustainable Development*, 7(1), 45-56. https://www.researchgate.net/publication/376687740_Optimization_of_Cloud_Costs

Repository Link - kasethi23/MSE-302-Individual_Project