Analysis and Design Report

Report Period: January 24, 2024 - February 14, 2024

Project Title: ARA (AI-powered Research Assistant)

Guide: Dr. D. J. Chaudhari

Report Prepared By: Kaustubh Warade,
Aditya Deshmukh,
Devansh Parapalli,
Yashasvi Thool

Executive Summary:

This report outlines the requirement analysis and design phase of ARA, highlights the planning and execution of UI design and overall system architecture.

Current Phase Progress:

Task Description	**Scheduled Completion**	**Actual Completion**	**Status**	**Remarks**
Analysis and Design Phase	February 14, 2024	February 14, 2024	Completed	

Accomplishments:

- Conduction of requirement analysis sessions was done to ensure the functionalities required in the application were understood.
- A simplified mockup of the tool's user interface was created to visualize the flow, interactions and overall experience.
- A mixed computing model was developed to ensure sufficient compute resources while minimizing cost.

Challenges & Mitigation:

- Resources required for implementation of “Mixture of Experts Large Language Model” feature were too high for central hosting along with ARA.
  An external provider “Cohere” was outlined for use within ARA.
- Response time of Language Models and AI-search Retrieval were too long for serverless platforms.
  A local “instance” of supervisor was planned for use in long running tasks.
- Large volumes of text must be vectorized and embedded to allow for correct retrieval and subsequently, augmented generation.
  This process is accomplished by utilizing open-source APIs and Deep Learning Models.

Planned Activities for Next Phase:

- Create schemas, connections, setup a database and vectorstore.
- Create a functional application for ARA, along with setting up Chains, Authentication and Storage.
- Deploy ARA to be accessible over any network, while still utilizing computing resources that you own.
- Create unit tests and integration tests for ARA.

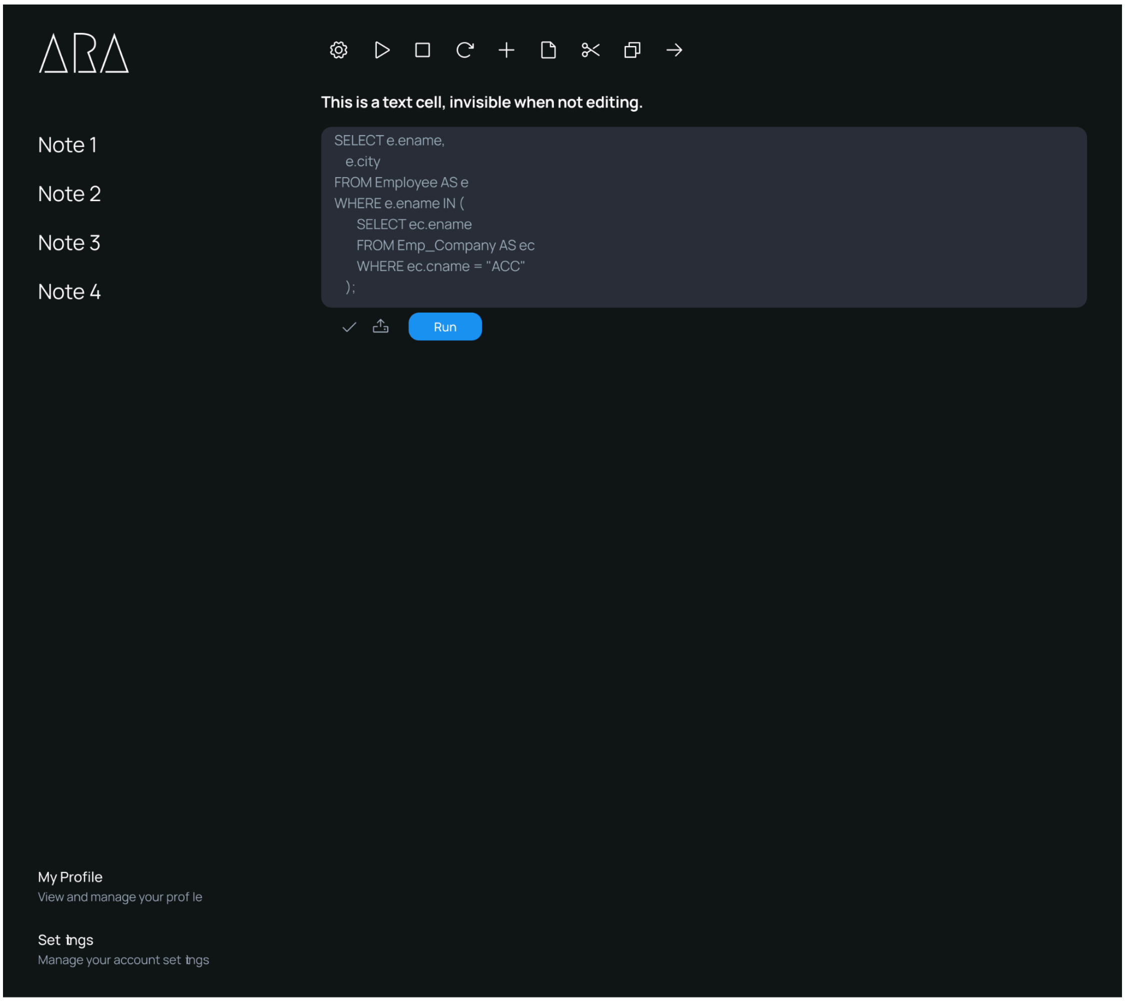
Financial Summary

Description	**Budgeted Amount**	**Amount Spent This Phase**	**Total Expenditure**
VPS (AWS)	₹5000/-	₹0/-	₹0/-

Risk Assessment:

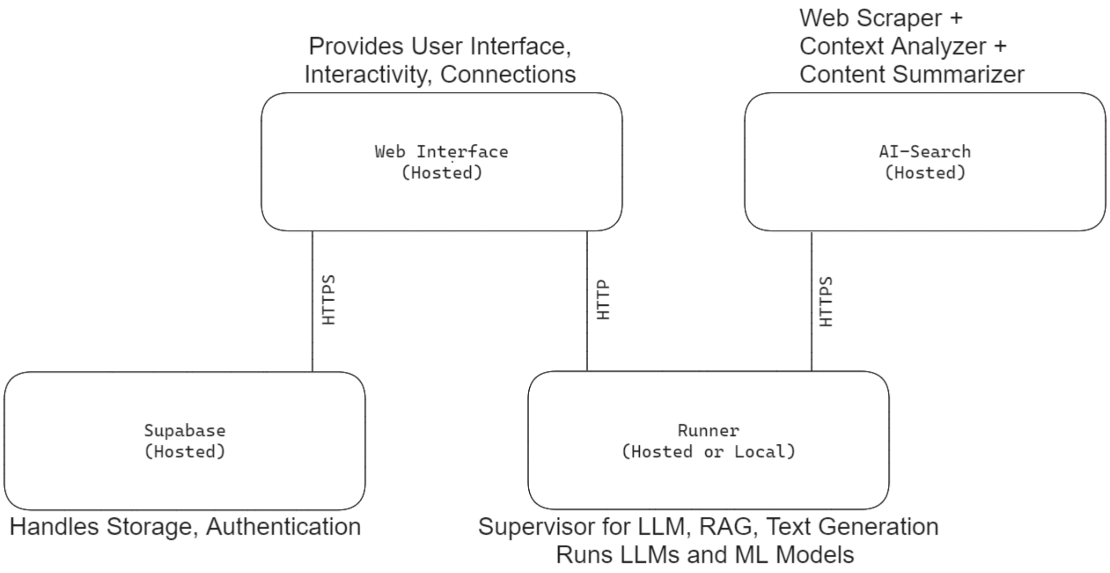
- User Interface Design, if not user-friendly or filled with bugs may reduce user adoption. Users may resist transitioning to a new system.
- As ARA's demand grows, scalability issues may start to prop up, mitigation is done using a local "instance" to ensure computing power never runs out.
- New developments in LLM/LAM and RAG may require massive infrastructure and developmental costs.

Attachments:



This is a text cell, invisible when not editing.

```
SELECT e.ename,
  e.city
FROM Employee AS e
WHERE e.ename IN (
    SELECT ec.ename
    FROM Emp_Company AS ec
    WHERE ec.cname = "ACC"
  );
```



Conclusions and Recommendations:

The Analysis and Design Phase was completed on schedule, establishing a robust foundation for ARA. The requirement analysis sessions laid the groundwork for successful development of ARA. The implementation of a mixed computing model provides for a balance between adequate compute resources and cost-effectiveness.

Approval:

Dr. D. J. Chaudhari

Project Guide

Assistant Professor, CSE Department

Government College of Engineering, Nagpur

Sector-27, Mihan Rehabilitation Colony

Khapri, Nagpur

441108

Date: March 02, 2024