Yannick Lansink, 900102137

Plan van Aanpak

Software Developer, Hogeschool NOVI

5 Juli, 2023

Disclaimer

“Wegens de vertrouwelijke aard van de informatie in het werkstuk/ de afstudeerscriptie en de bijbehorende bijlagen, mag de inhoud noch geheel noch gedeeltelijk op enigerlei wijze worden aangepast, gewijzigd of verveelvoudigd zonder schriftelijke toestemming van de auteur en het betrokken bedrijf. Zowel de docent/ scriptiebegeleider, de examinator alsmede de medewerkers van NOVI Hogeschool worden gehouden de inhoud van het werkstuk/ de afstudeerscriptie als strikt vertrouwelijk te behandelen. NOVI Hogeschool bewaart het document in een afgesloten databank. Gedurende de bewaarperiode kunnen studentdossiers worden ingezien door medewerkers van NOVI, de Examencommissie van NOVI alsmede de Inspectie van het Onderwijs en een visitatiecommissie van de Nederlands Vlaams Accreditatieorganisatie (NVAO, bij (her)accreditaties).“

Inhoud

1. Inleiding

Integratie Business Services (IBS) Toeslagen is bezig met het verjongen van het personeel en door de complexiteit zijn veel medewerkers de dupe van het langdurig opzoeken naar antwoorden op (retorische) vragen. Deze tijdsverspilling gaat de chat applicatie, met ondersteuning van een redeneringsmodel, oplossing.

Ook voor ervaren medewerkers is het complexe en snel veranderende systeem van IBS Toeslagen niet altijd in zijn geheel op te omvatten. Als men aan dit knopje draait valt er aan de andere kant een test om. Vragen komen dan naar boven hoe dit werkt. Om hier een geheel beeld bij te krijgen kan de chat applicatie de gebruiker ondersteuning bieden en helpen met het vermogen om complexe vraagstukken met de bijbehorende bedrijfslogica te beantwoorden. Daarnaast biedt het analyserend vermogen.

Deze verrijking biedt teams binnen IBS toeslagen een beterde efficiëntie en effectiviteit. Ook zal hiermee de tevredenheid onder medewerkers verhogen aangezien een productievere medewerker een grotere bijdrage levert aan het team. Dit zijn 2 van de doelen die IBS Toeslagen heeft opgesteld in het 5 jaren plan (Jansen, 2023).

De auteur gaat het beroepsproduct de komende weken uitbouwen tot een werkend prototype dat wordt ondersteund met een functioneel- en technisch ontwerp. Het is de bedoeling dat het prototype lokaal draaibaar is. De applicatie maakt gebruik van een Azure function, frontend, backend en een vector database. Deze samenhang van componenten biedt een gebruiksvriendelijke manier van interactie en een makkelijker onderhoudbaar stuk software. Verder zoals vermeld kan een Large Language Model (LLM), wat in de backend zit ingebakken, ondersteuning bieden op analytisch vermogen.

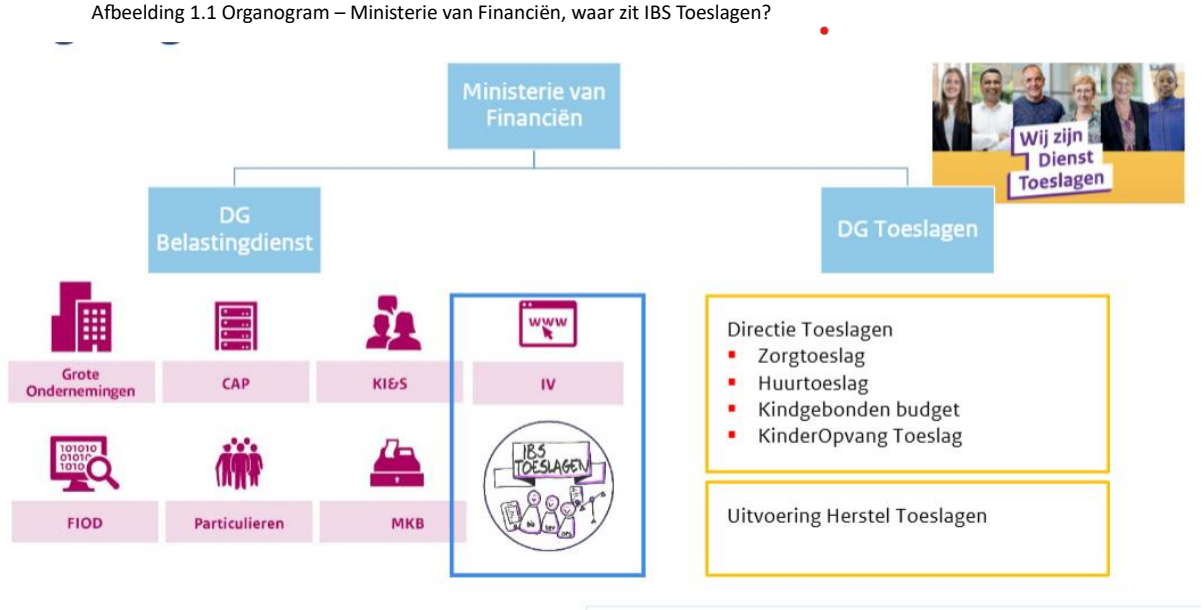
2. Belang organisatie Toeslagen

2.1 Beschrijving van het onderwerp en de context.

Bij het onderdeel Integratie Business Services (IBS) Toeslagen werken ruim 15 teams binnen het Scaled Agile Framework (SAFe) framework, wat neerkomt op +- 150 IT-medewerkers. Binnen en over deze teams wordt veel informatie verzameld. Documentatie wordt op verschillende plekken gedeponeerd: repositories, de wiki, chat kanalen, video gesprekken, online hub en op nog veel meer locaties kom je relevantie informatie tegen. Een onoverzichtelijke en niet controleerbare situatie heeft zich voorgedaan.

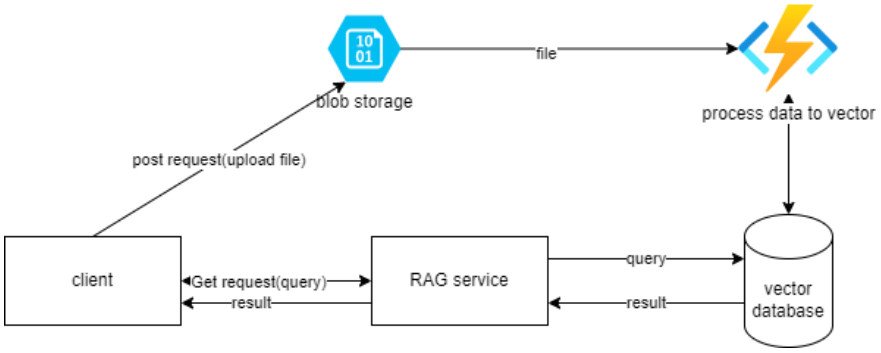
Nieuwe medewerkers zijn vooralsnog uren bezig met het opzoeken van vragen wat normaal gesproken nauwelijks tijd moet kosten. De kanalen worden doorgelezen, repositories worden bekeken en uiteindelijk heeft men een antwoord op de vraag gevonden in de wiki. Dit is omslachtig en een probleem waar de auteur ook tegen aan liep. Het pakje wordt hier uiteindelijk een zakje; het probleem zal groter worden naarmate er meer informatie wordt vergaard door IBS Toeslagen. En dat probleem moeten we onder de knie krijgen en oplossen.

Afbeelding 1.1 Organogram – Ministerie van Financiën, waar zit IBS Toeslagen?



Om deze uitdaging het hoofd te bieden, wordt een chatapplicatie ontwikkeld die medewerkers ondersteunt bij het vinden van informatie. Deze chatapplicatie zal gebruikmaken van een redeneringsmodel in combinatie met een vector database en een Large Language Model (LLM). Het doel van deze technologie is om complexe vraagstukken direct en nauwkeurig te kunnen beantwoorden, gebaseerd op de specifieke bedrijfslogica en documentatie binnen IBS Toeslagen. Daarnaast zal de applicatie analyserend vermogen hebben om contextuele antwoorden te bieden. Dit zal niet alleen de efficiëntie van de teams verbeteren, maar ook de tevredenheid van de medewerkers verhogen, aangezien zij sneller toegang krijgen tot de benodigde informatie.

Diagram 2.1: Technische context van de applicatie



In het overkoepelende diagram (Diagram 2.1) worden de verschillende technische componenten van de chatapplicatie weergegeven en hun onderlinge samenhang. De applicatie bestaat uit de frontend (client) voor de gebruikersinterface, de backend (RAG service) die verzoeken verwerkt, een vector database voor het snel doorzoeken van ongestructureerde data, en een LLM (RAG service) dat door middel van NLP-technieken antwoorden verrijkt en analyseert. Deze componenten werken samen om gebruikers snel relevante antwoorden te geven zonder dat zij verschillende bronnen handmatig hoeven te doorzoeken.

2.2 Uitleg over de relevantie van het onderwerp voor de opleiding.

Het beroepsproduct dat de afstudeerder ontwikkelt, is een applicatie die in een specifieke behoefte voorziet binnen de organisatie IBS Toeslagen. Het is geen verbetering op een bestaand systeem, maar een compleet nieuwe ontwikkeling die tijdsbesparing biedt aan de teams. Het prototype zal aanvankelijk beschikbaar worden gesteld aan team Vos, het team waarin de auteur zich bevindt.

De chatapplicatie zal functioneren als een centraal systeem dat de verschillende informatiebronnen samenvoegt. Door gebruik te maken van meerdere programmeertalen, een cloud-omgeving, een vector database en een LLM, ontstaat een complexe en technisch uitdagende oplossing die nauw aansluit bij de vakgebieden die binnen de opleiding aan bod komen. Het resultaat is een applicatie met een frontend waarmee gebruikers eenvoudig kunnen interacteren, wat medewerkers direct helpt bij hun dagelijkse werkzaamheden.

2.3 Uitgebreide stakeholderanalyse.

Voor een stakeholdersanalyse is het relevant om te weten wie baat en invloed heeft bij het project. Dit zijn leidinggevenden, teamleden, security specialisten en het portfolio team. Al deze personen zullen of gebruik maken van de applicatie, of baat hebben bij de implementatie ervan. Hier moet duidelijk mee overlegd worden zodat het project ook nut heeft. Het kan niet zo zijn dat er een applicatie wordt gebouwd waar uiteindelijk geen belang naar is, dan slaat men de plank mis. Dus een duidelijke communicatie staat hier centraal.

Diagram 2.2: Kracht-Belang Matrix

Stakeholder	**Invloed**	**Strategie**
Leidinggevende	Gemiddeld	Actief beheren met regelmatige update en feedback op vragen.

Teamleden	Hoog	Informeren en feedback verzamelen
Security specialist	Hoog	Advies vergaren
Portfolio team	Gemiddeld	Advies vergaren

Een leidinggevende wilt zien dat het project goed verloopt en dat de gestelde deadlines zijn gehaald. Dit persoon verwacht regelmatige updates en wilt voortuitgang zien.

De teamleden zijn de uiteindelijke gebruikers van de applicatie. Met hun sluit ik kort wat de precieze behoeften zijn. Hieruit stel ik ook de requirements op. Het is van groot belang om dicht bij de klant te staan. Hierdoor zal de feedbackloop verkleinen, waardoor de relevantie van de applicatie niet vervalt en men er zekerder van is dat het project loopt zoals verwacht.

Een poos geleden is de auteur met een security specialist om tafel gegaan om advies te vragen over het gebruik van LLM's. Dit is een type model voor machinaal learning dat verschillende taken op het gebied van natuurlijke taalverwerking of Neuro Linguïstisch Programmeren (NLP) kan uitvoeren, zoals het genereren en classificeren van tekst, het beantwoorden van gespreksvragen en het vertalen van tekst van de ene taal naar de andere (*Wat Is een Llm*, z.d.). Voor IBS Toeslagen is privacy gevoeligheid een van de meeste belangrijke factoren voor dit project. We willen geen data klakkeloos deponeren naar derde partijen die vervolgens de data gebruiken voor verdere trainingsdoeleinden.

Het portfolio management van IBS Toeslagen beheert een verzameling van projecten binnen de organisatie, gericht op strategische doelstellingen. De primaire taak van het team is om te zorgen dat aan de meest relevante projecten wordt gewerkt. Dit prioriteren zij. Voor dit project moet het portfolio management de prioriteit ook bepalen waarbij ze de risico's, middelen en kosten berekenen.

2.4 Belang van het onderwerp voor de organisatie.

De auteur is op het moment van schrijven een half jaar in dienst bij IBS Toeslagen. Bij de instroom, met 9 andere trainees, viel het op dat het landschap complex is. Gemiddeld zijn er 60.000 bezoekers per dag, meer dan 40.000 wijzigingen per jaar en wordt er ruim 2 miljoen keer gebeld per jaar naar Toeslagen. Om dit te ondersteunen is het ICT-landschap uitgegroeid tot een complex architectuur waar 15 teams aan werken met veel processen die onderwater lopen ("Dienst Toeslagen van Wetgeving Naar Uitvoering Onder Architectuur", z.d.). Door deze complexiteit is er veel informatie, die ook nog eens op verschillende plekken zijn te vinden. Voor medewerkers, en ten zeerste nieuwe medewerkers, is het lastig om de relevantie informatie te zoeken die nodig is voor het werk.

In een altijd evaluerende bedrijfsomgeving gaan groei en data beheer hand in hand. IBS Toeslagen is gegroeid tot een complex systeem waardoor data opslag toeneemt. Door hier vol op in te zetten en het probleem op te lossen kan men de efficiënte van haar medewerkers aanzienlijk verbeteren. Gebeurt dit niet, dan zal de complexiteit toenemen en zullen medewerkers nog meer tijd besteden aan het zoeken naar antwoorden.

2.5 Koppeling met het tactische niveau en de strategische doelen van de organisatie.

Nieuwe technologie biedt IBS Toeslagen de kans om de opslag en het beheer van data te verbeteren. De kern van de oplossing ligt in het gebruik van een vector database, die helpt om ongestructureerde data effectiever doorzoekbaar te maken. Traditionele databases zijn vaak beperkt tot het exact matchen van woorden en trefwoorden, wat niet altijd efficiënt is bij grote hoeveelheden complexe,

ongestructureerde data. In een vector database daarentegen worden woorden, zinnen en documenten omgezet in vectoren — numerieke representaties die een semantische betekenis bevatten (What Is A Vector Database, z.d.).

Deze vectoren worden geplaatst in een ruimte waar de afstand tussen twee vectoren correspondeert met de mate van overeenstemming in betekenis. Wanneer een gebruiker een vraag stelt, zet de applicatie deze vraag om in een vector en zoekt deze vervolgens in de vector database naar andere vectoren (documenten of stukken informatie) die inhoudelijk dichtbij liggen, zelfs als er geen exacte woordovereenkomsten zijn. Dit stelt de gebruiker in staat om snel relevante informatie te vinden, zonder dat de exacte woorden in de vraag en het antwoord overeen hoeven te komen.

Dit nieuwe data-opslag en -ophaalproces vormt een cruciale stap voor het aanpakken van het informatiebeheerprobleem binnen IBS Toeslagen. Het opzoeken van ongeorganiseerde informatie is momenteel tijdrovend, maar door data op deze manier op te slaan en doorzoekbaar te maken in een gebruiksvriendelijke applicatie, kan IBS Toeslagen efficiënter werken. Ook zal het de gebruikerservaring verbeteren door het vereenvoudigen van het informatiezoekproces. Dit sluit aan bij de middellange termijn doelen van de organisatie om zowel de medewerkerstevredenheid als de efficiëntie te verhogen (Jansen, 2023).

Deze verbetering in gebruikerservaring en efficiëntie ondersteunt ook de strategische doelstellingen van IBS Toeslagen om in de top 10 favoriete werkgevers te blijven. Daarnaast draagt de applicatie bij aan de verhoging van de functiepuntproductiviteit per team — een meting van de hoeveelheid werk die per team wordt gerealiseerd. Het verhogen van deze productiviteit vertaalt zich naar een verbeterde output en hogere efficiëntie van de teams binnen IBS Toeslagen (Jansen, 2023).

3. Businessdoelstellingen en onderzoeksdoelstellingen

3.1 Formulering van de hoofdvraag en deelvragen

Om de onderzoekskant van dit werk te versterken en een structurele aanpak te waarborgen, wordt de volgende hoofdvraag geformuleerd:

Hoofdvraag: "Hoe kan een chatapplicatie, ondersteund door een vector database en een Large Language Model (LLM), bijdragen aan het verbeteren van de efficiëntie voor het opzoeken naar informatie binnen IBS Toeslagen?"

Deelvragen:

1. Hoeveel tijd besteden medewerkers momenteel aan het zoeken naar antwoorden op vragen, en hoe kan deze tijd worden gereduceerd?
2. Welke eisen en wensen hebben de verschillende stakeholders (zoals teamleden, leidinggevenden, en security specialisten) met betrekking tot de applicatie?
3. Wat zijn de technische en beveiligingsvereisten voor het integreren van een vector database en LLM binnen IBS Toeslagen?
4. Hoe kan de gebruikerservaring van de applicatie worden geoptimaliseerd om een zo effectief mogelijke ondersteuning te bieden aan zowel nieuwe als ervaren medewerkers?

3.2 Formulering van de business- en onderzoeksdoelstelling

De businessdoelstellingen van dit project richten zich op het verbeteren van de efficiëntie, de informatiebeheerprocessen, en het verhogen van de medewerkerstevredenheid binnen IBS Toeslagen. De implementatie van een nieuwe applicatie die snel en effectief toegang biedt tot relevante informatie sluit aan bij de strategische doelen van de organisatie. Deze doelen zijn van belang om de operationele effectiviteit te vergroten en de teams binnen IBS Toeslagen beter te ondersteunen.

Specifieke Businessdoelstellingen:

1. **Efficiëntie verhogen**: Door de ontwikkeling van een applicatie die de informatieopslag en -opvraging optimaliseert, wordt de tijd die medewerkers kwijt zijn aan het zoeken naar antwoorden aanzienlijk verminderd. Dit leidt tot een snellere besluitvorming en meer productiviteit, vooral voor nieuwe medewerkers die doorgaans veel tijd verliezen aan het inwerken in de complexe informatieomgeving.
2. **Medewerkerstevredenheid verbeteren**: Een van de primaire doelen is om de tevredenheid van medewerkers te verhogen door hen te voorzien van gebruiksvriendelijke tools die hun werk vergemakkelijken. Door de toegankelijkheid van informatie te verbeteren, wordt hun werkervaring positiever, wat leidt tot minder frustratie en een hogere motivatie.
3. **Data governance en veiligheid waarborgen**: IBS Toeslagen werkt met gevoelige informatie. Het waarborgen van data privacy en veiligheid, vooral in combinatie met het gebruik van Large Language Models (LLMs), is van essentieel belang. Deze doelstelling richt zich op het implementeren van strikte databeveiligingsmaatregelen in de nieuwe applicatie om te voorkomen dat gevoelige informatie toegankelijk wordt voor derden.
4. **Innovatie en technologische ontwikkeling stimuleren**: De introductie van een cloud-gebaseerde applicatie met een geavanceerde zoekfunctie (op basis van een vector database en NLP-technologieën) draagt bij aan de digitale transformatie van IBS Toeslagen. Dit project zorgt ervoor dat de organisatie up-to-date blijft met moderne technologieën en processen, wat noodzakelijk is om competitief en efficiënt te blijven in een snel veranderende digitale omgeving.
5. **Kostenreductie**: Door de tijd te verminderen die medewerkers besteden aan het zoeken naar informatie, wordt er indirect ook bespaard op operationele kosten. Minder tijdverspilling betekent meer waardevolle arbeidstijd die kan worden ingezet voor productieve taken.

Voordat het project kan beginnen moet er worden aangetoond dat het probleem daadwerkelijk aanwezig is. Dit wordt gedaan door een onderzoek af te leggen. Het onderzoek zal moeten aantonen dat het medewerkers tijd en frustratie kost om informatie te vergaren. Alleen dan kan het van een prototype naar een werkelijk productiewaardig product worden gebouwd.

Het SMART-geformuleerde onderzoeksdoelstelling:
"Het in kaart brengen van de informatiezoekpatronen en -behoeften van medewerkers binnen IBS Toeslagen door middel van een gebruiksanalyse, en vervolgens het ontwerpen en testen van een prototype chatapplicatie die de zoekduur voor nieuwe medewerkers met ten minste 20% vermindert binnen een periode van 6 maanden, waarbij wordt voldaan aan de beveiligingsvereisten en technische eisen van IBS Toeslagen."

3.3 Beschrijving van de aanpak voor het literatuuronderzoek.

Voor het literatuuronderzoek wordt gebruik gemaakt van een benadering waarbij academische artikelen en publicaties worden onderzocht. Ook wordt onderzocht wat de laatste trends zijn omtrent de technologie die van toepassing is. Zo is de wereld van LLM's fascinerend snel, maar dit betekent ook dat je bij moet blijven met de nieuwste vernuftige uitvindingen. Hoe die trends worden opgezocht gebeurt op basis van industrie pioniers. De openbare discussies die plaatsvinden worden met dit persoon worden geanalyseerd en meegenomen in het onderzoek.

Maar de grootste focus ligt op de wetenschappelijke bronnen over zoek technologieën (vector database), de integratie met LLMs en de veiligheid van interne data. De auteur neemt ook voorgaande case studies mee op plekken waar dit concept al is toegepast. Deze soortgelijke implementaties geven inzicht in de problemen waar men tegen aan liep en hier wordt ook geanalyseerd wat de best practises zijn.

3.4 Specifieke zoektermen die gebruikt zullen worden.

- "Vector databases and search functionalities"
- "Natural Language Processing (NLP) in enterprise applications with data protection"
- "Large Language Models (LLM) and data security"
- "Efficiency of information storage in complex organizations"
- "Privacy and data protection in LLM integrations"
- "Azure functions, how to"
- "Information management within the SAFe framework"
- "Openai data integrity"
- "How to run your own llm"
- "How to build a rag service in python"

3.5 Relatie van de zoektermen tot het onderwerp.

De gekozen zoektermen zijn specifiek gericht op de onderwerpen die van direct belang zijn voor de ontwikkeling van de applicatie binnen IBS Toeslagen. Bijvoorbeeld, "vector databases and search functionlity" richt zich op de technische oplossingen die de applicatie efficiënter maken bij het zoeken naar informatie. " Natural Language Processing (NLP) in enterprise applications with data protection" behandelt de mogelijke uitdagingen en oplossingen bij het integreren van LLM's zonder dat dit ten koste gaat van databeveiliging, een cruciaal punt binnen IBS Toeslagen.

3.6 Onderzoeksmethoden per deelvraag

Om per deelvraag gericht onderzoek te doen, wordt voor elke vraag een passende onderzoeksmethode gehanteerd. Dit biedt een gedetailleerde aanpak om de informatiebehoeften van medewerkers en de eisen aan het systeem grondig te analyseren.

Onderzoeksmethoden per deelvraag:

"Hoeveel tijd besteden medewerkers momenteel aan het zoeken naar antwoorden op vragen, en hoe kan deze tijd worden gereduceerd?"

Methode: Observatie. Een observatie van werksituaties biedt inzicht in de tijd die medewerkers besteden aan informatie zoeken.

"Welke eisen en wensen hebben de verschillende stakeholders (zoals teamleden, leidinggevenden en security specialisten) met betrekking tot de applicatie?"

Methode: Interviews met vertegenwoordigers van elke stakeholdergroep. Dit biedt gedetailleerde input over specifieke eisen en wensen en helpt bij het opstellen van de requirements voor de applicatie.

"Wat zijn de technische en beveiligingsvereisten voor het integreren van een vector database en LLM binnen IBS Toeslagen?"

Methode: Documentanalyse van de bestaande veiligheidsprotocollen en interviews met security specialisten.

"Hoe kan de gebruikerservaring van de applicatie worden geoptimaliseerd om een zo effectief mogelijke ondersteuning te bieden aan zowel nieuwe als ervaren medewerkers?"

Methode: Prototype. Gebruikerstests met een prototype van de applicatie zorgen ervoor dat de applicatie gebruiksvriendelijk en intuïtief is voor verschillende gebruikersgroepen binnen de organisatie.

3.6 Toepassing van onderzoeksmethoden en-technieken.

Voor het literatuuronderzoek voert de auteur alleen een kwalitatief onderzoek uit. Enerzijds wordt er een analyse gedaan van bestaande onderzoeken en statistieken over de effectiviteit van vector databases en LLM's in bedrijfscontexten. Anderzijds wordt er een onderzoek uitgevoerd door casestudies en praktijkvoorbeelden te analyseren die toepasbaar zijn op de huidige situatie binnen IBS Toeslagen. Methoden zoals literatuurreview, meta-analyse, en benchmarking worden toegepast om een brede basis van informatie te vergaren.

3.7 Verwachte bronnen en databanken.

Verwachte bronnen zijn:

- Google Scholar: Voor academische artikelen over technologieën zoals LLM's, NLP, en informatiebeheer.
- IEEE Xplore: Voor technische papers over de implementatie van vector databases en cloud-gebaseerde oplossingen.
- Interne documenten van IBS Toeslagen: Voor casestudies, documentatie over de huidige systemen, en interne rapporten over informatiestromen.

- Azure Cloud: Voor de technische implementatie voor de cloud gang van de applicatie
- Open AI: Voor de LLMs die gebruikt worden. Hier lees je over welke modellen er beschikbaar zijn. Ook bestudeer je hier wat voor data protectie mogelijkheden er zijn, zoals Openai Enterprise, wat is dat?
- Langchain: Momenteel de voorloper op het gebied van het abstraheren van LLM api's. Hier leest de auteur de technische implementatie van de applicatie en de nieuwere trends.

3.8 Toepassing van methoden en tools voor ontwerp, bouw en test

Naast de onderzoeksmethoden worden ook methoden en tools gebruikt voor het ontwerp, de bouw en de testfase van de applicatie. Deze omvatten:

- **Programma-ontwerp:** De applicatie wordt ontworpen volgens het *Separation of Concerns* (SoC)-principe om complexiteit te verminderen. Tools zoals Draw.io wordt gebruikt voor de architectuurdiagrammen.
- **Programma-bouw:** Voor de applicatieontwikkeling wordt een mono-repo architectuur gebruikt met Azure Functions en de RAG-service. Versiebeheer wordt gedaan met Git. Voor de backend wordt Python gebruikt met frameworks als LangChain en OpenAI-API voor LLM-integratie.
- **Programma-test:** De applicatie wordt getest op functionele eisen en gebruikerservaring door middel van usability testing met feedback van medewerkers. Tools zoals Postman (voor API testing) worden ingezet om een optimale functionaliteit en gebruiksvriendelijkheid te waarborgen.

4. Ontwerp van het beroepsproduct

4.1 Gedetailleerde beschrijving van het uiteindelijke beroepsproduct.

Het op te leveren beroepsproduct is een chat applicatie met bijbehorende frontend, backend en een database. Een gebruiker upload, zoekt en leest naar antwoorden via de web applicatie. De berichtgeving loopt asynchroon, dat wilt zeggen dat berichten terug worden gestreamd naar de gebruiker zodra de backend die via een derde partij ophaalt.

Het product wordt ondersteund door een redeneringsmodel. Dit is een model dat de context van het antwoord gebruikt om directer en nauwkeuriger antwoord te geven aan de gebruiker. Dit gebeurt onderwater en hier heeft de gebruiker geen invloed op. Zo'n model heet een Large Language Model (LLM) en is in de context van de applicatie is het capabel om met pdf's, tekst en html pagina's te interacteren. Dit ondersteund de applicatie en geeft de gebruiker een rijker antwoord. (*Gpt 4 Paper*, z.d.)

Een vector database gaat hand in hand met een LLM. Dit aangezien data binnen IBS Toeslagen ongestructureerd is; alles wordt momenteel gedumpt op verschillende plekken. Een LLM kan aan deze ongestructureerde data betekenis geven door eerst de gebruiker te ondersteunen in het vergaren van de juiste data en er dan vervolgens een beredenering op los laten.

Een vector database fungeert als een grafiek waarin de documenten die je erin stopt gerepresenteerd worden door punten in een 3 dimensionaal grid, oftewel vectoren (Wat Is een Vector Database en Waarom Belangrijk Voor Generatieve AI?, z.d.). Een vector database kun je zien als een bibliotheek, maar in plaats van boeken op alfabetische volgorde te ordenen, wordt alles gerangschikt op basis van de betekenis van de inhoud. Elke boek wordt dan omgezet in een vector

die in dit 3 dimensionaal grid wordt gezet. Wanneer de gebruiker dan een vraag stelt wordt deze vraag ook omgezet in een vector, waardoor er met een similairity search (bepaalde zoek methode) kan worden gezocht naar boeken (business specifieke informatie) die qua betekenis het dichts in de buurt komen. Hierdoor vind je boeken die relevant zijn op basis van de vraag die de gebruiker stelt, zelfs al zit het exacte trefwoord niet in de vraag.

Dit krachtige duo van een vector database met een LLM wordt veel gebruikt in de industrie van chat applicaties aangezien je business specifieke domein kennis kan meegeven waar de LLM weer een analyse op los laat. Dit ondersteund de gebruik aanzienlijk. Hierdoor wordt het antwoord verrijkt en binnen enkele seconden teruggegeven.

Om het kort samen te vatten kan de gebruiker met de chat applicatie business specifieke vragen stellen en hier binnen enkele seconden verrijkende antwoorden voor terugkrijgen.

4.2 Uiteenzetting van alle onderdelen die opgeleverd zullen worden.

Het uiteindelijke beroepsproduct bestaat uit meerdere geïntegreerde componenten die gezamenlijk de chat applicatie vormen. Dit is gedaan om de als onderdeel van een software principe; seperation of concerns. Hierbij kies je een design principe waarbij de applicatie wordt opgedeeld in individuele en kleinere onderdelen. Hierdoor wordt de complexiteit verminderd en in beter te beheren onderdelen gestopt (GeeksforGeeks, 2024). De onderdelen bij elkaar zorgt voor het realiseren van een efficiënte, betrouwbare en gebruiksvriendelijke ervaring voor de medewerkers van IBS Toeslagen. Hieronder volgt een overzicht van de op te leveren onderdelen:

Vector Database (gebruik maken van een derde partij):
De vector database vormt de kern van het data-opslagsysteem en maakt het mogelijk om ongestructureerde data te organiseren op basis van semantische betekenis. Deze database ondersteunt de applicatie door opgeslagen documenten te transformeren in vectoren, zodat er snel en effectief gezocht kan worden op basis van betekenis, in plaats van exacte trefwoorden. Door gebruik te maken van een derde partij die gespecialiseerde vector databases aanbiedt, wordt de complexiteit van het beheren en onderhouden van een dergelijke infrastructuur geminimaliseerd. De gekozen vector database zal geïntegreerd worden met het LLM om de zoekopdrachten en antwoorden rijker en contextueel nauwkeuriger te maken.

Azure Function (onderdeel van mono repo):
De Azure Function dient als de serverloze backend infrastructuur voor de applicatie. Deze functie zorgt voor de asynchrone verwerking van zoekopdrachten en communicatie tussen de frontend en de derde partij die de data in de vector database opslaat. Door gebruik te maken van Azure Functions kan de applicatie op een schaalbare en kosten-efficiënte manier werken, zonder dat er continu servers hoeven te draaien (Ggailey, 2023). Dit zorgt ervoor dat resources alleen worden gebruikt wanneer nodig, wat zowel de kosten als de technische complexiteit reduceert.

RAG Service (onderdeel van mono repo):
De Retrieval-Augmented Generation (RAG) service is de module die verantwoordelijk is voor het ophalen van relevante informatie uit de vector database en deze vervolgens te combineren met het analytische vermogen van de LLM. Deze service is een cruciaal onderdeel van de architectuur, omdat het ervoor zorgt dat de antwoorden die aan de gebruiker worden gegeven, niet alleen gegenereerd worden door de LLM, maar ook verrijkt worden met de specifieke data en documenten van IBS Toeslagen. De RAG service zoekt eerst naar relevante data in de vector database en laat het LLM vervolgens een inhoudelijk antwoord genereren op basis van die data.

Frontend (onderdeel van mono repo):
De frontend is de web gebaseerde interface waarmee de gebruiker interacteert. Deze zal de mogelijkheid bieden om vragen te stellen, documenten te uploaden en antwoorden te ontvangen. De interface is gebruiksvriendelijk en intuïtief, zodat gebruikers zonder technische kennis gemakkelijk met de applicatie aan de slag kunnen. De berichten worden asynchroon teruggestreamd naar de frontend, wat betekent dat gebruikers de antwoorden direct ontvangen zodra deze beschikbaar zijn vanuit de backend en RAG service.

Deze componenten vormen samen een geïntegreerd systeem dat snel en nauwkeurig antwoorden biedt op basis van de rijke data die binnen IBS Toeslagen aanwezig is.

4.3 Formats van de op te leveren onderdelen.

1. Functioneel Ontwerp Document (FOD)

Format: PDF-document.
Inhoud: Een gedetailleerde beschrijving van de functionaliteiten van de applicatie, inclusief user stories, use cases, en functionele eisen. Dit is essentieel voor de ontwikkeling en voor het duidelijk vastleggen van de verwachtingen.

2. Technisch Ontwerp Document (TOD)

Format: PDF-document.
Inhoud: Beschrijving van de technische specificaties van de applicatie, zoals de architectuur, programmeertalen, frameworks, databases, en integraties met andere (derde partijen) systemen.

3. Applicatie Prototype

Format: Werkende applicatie (lokaal draaibaar)
Inhoud: Een werkende applicatie die de gebruiker lokaal kan draaien om interactie hebben via de gebruikersinterface. Hiermee is lokaal, mits de juiste configuratie is ingesteld, met de RAG-service te communiceren. Het is dan mogelijk om de functionele eisen uit te voeren.

4.4 Samenhang tussen de verschillende onderdelen.

Er is gekozen om de communicatie tussen de frontend en RAG-service te laten verlopen via asynchrone HTTP verzoeken om de berichten terug te streamen naar de gebruiker. De RAG-service kan ook een aanvraag ontvangen en vervolgens er voor zorgen dat deze wordt opgepakt door een azure function. De RAG-service kan ook met derde partijen, aanbieders van taalmodellen, communiceren om de antwoorden te verrijken voor de gebruiker.

Het is verder van hoog belang dat de vector database een lage latentie heeft voor het overdragen van data, vooral naar de gebruiker toe. De database is onderdeel bij elk verzoek van de gebruiker. Bij het stellen van een vraag communiceert de RAG-service met de database en bij het verstrekken van een document door de gebruiker wordt er via de RAG-service een verzoek gestuurd naar de azure function. Voor de communicatie tussen deze laatste zin is het nog onduidelijk wat het effectiefst is. De auteur denkt over een HTTP-verzoek of een event, maar dit wordt nog uitgepluisd.

4.5 Verwachte impact van het beroepsproduct op de organisatie.

IBS Toeslagen zit momenteel in de fase van verjonging. Veel oudere collega's gaan met pensioen en er worden veel nieuwe trainees aangenomen en getraind om het werk op te vangen. De voorgaande

lichting trainees waren met totaal 9 en de hierop volgende gaat een klas met 18 worden. Deze immense verhoging is te voelen bij het team van de auteur. Het afwijzen van nieuwe trainees is bijvoorbeeld niet mogelijk.

Het werk is complex en daarom is geduld een schone zaak bij het inwerken van nieuwe medewerkers. Het vergt veel tijd voor senior ontwikkelaars om beginnend ontwikkelaar te ondersteunen in dit doolhof. De auteur heeft opgemerkt dat hier ook laks mee wordt omgegaan. Trainees worden in het diepe gegooid zonder enig gevoel van richting. Dit op basis van de auteurs eigen ervaring en de lichting trainees waar hij bij zat.

Deze trainees lopen tegen problemen op bij simpele vragen. Er wordt langdurig naar een antwoord gezocht op plekken waar die niet zijn. Door een ingrijpende verandering in het opslaan van informatie in een centrale vector database kan met behulp van een chat applicatie hier verandering in doen aanbrengen. Nieuwe ontwikkelaars zullen binnen enkele seconden antwoord hebben op vragen die voor velen worden gezien als retorisch, maar ook op complexe vraagstukken waar over kan worden beraadslaagd.

Dit geldt niet alleen voor nieuwe medewerkers. De complexiteit van het IBS Toeslagen landschap en de vele veranderingen die er worden doorgevoerd maakt het voor ervaren teamleden ook onduidelijk wat en hoe het nou zit. Ook zullen ervaren medewerkers voordeel krijgen van de beredenering die de applicatie terug streamt. Zowel complexe als retorische vragen zijn mogelijk.

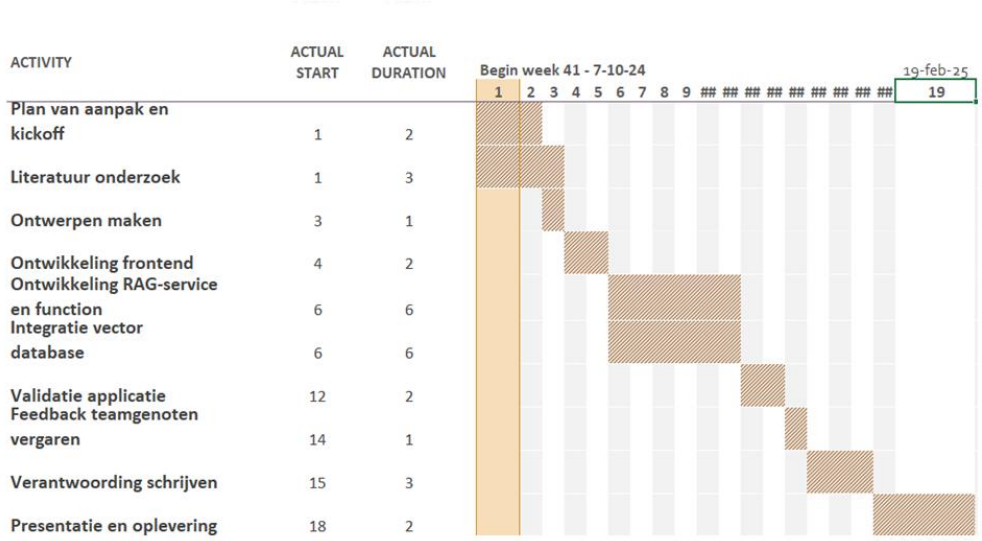
De auteur ziet een grote kans om de effectiviteit en efficiëntie van alle teams binnen IBS Toeslagen aanzienlijk te verhogen. Het zal voor nieuwe medewerkers leiden tot een snellere instroom en voldoening in de vorm van productiviteit aan het team. Voor bestaande medewerkers ziet de auteur ook veel voordeel, vooral complexe vraagstukken kan de applicatie bij helpen. Dit door alle relevante documenten in een handomdraai op te vragen en op basis daarvan een geanalyseerd en beredeneerd antwoord terug te streamen.

5. Planning

Voor de planning wordt gebruik gemaakt van een Gantt-chart. De auteur is van mening dat dit een realistisch beeld schets en tijd geeft om het project succesvol af te ronden. Voor de realisatie van de RAG-service en de integratie van de vector database is extra tijd in beschouwing genomen. Dit loopt overigens ook parallel met elkaar, aangezien er afhankelijkheden nauw liggen tussen de componenten.

Verder ziet de auteur nog een klein risico m.b.t. het vergaren van feedback door teamgenoten. Dit gebeurt laat in het project, waardoor er weinig tijd beschikbaar is om die feedback te verwerken. De auteur kiest er daarom voor om vaker de klant componenten te laten zien, zodat er meer tijd is om op de feedback in te spelen.

Gantt-chart – planning IBS Toeslagen chat applicatie


ACTIVITY
ACTUAL START
ACTUAL DURATION
Begin week 41 - 7-10-24
19-feb-25
1 2 3 4 5 6 7 8 9 ## ## ## ## ## ## ## ## ## 19
Plan van aanpak en kickoff 1 2
Literatuur onderzoek 1 3
Ontwerpen maken 3 1
Ontwikkeling frontend 4 2
Ontwikkeling RAG-service en function 6 6
Integratie vector database 6 6
Validatie applicatie 12 2
Feedback teamgenoten vergaren 14 1
Verantwoording schrijven 15 3
Presentatie en oplevering 18 2

Bronvermelding

Werken bij informatievoorziening (IV) | Belastingdienst - Werken bij de Belastingdienst. (z.d.). Werken bij de Belastingdienst. https://werken.belastingdienst.nl/informatievoorziening

ICT bij de Belastingdienst - Werken bij de Belastingdienst | Werken bij de Belastingdienst. (z.d.). Werken Bij de Belastingdienst. https://werken.belastingdienst.nl/expertises/ict

Wat is een llm. (z.d.). Techopedia. https://www.techopedia.com/nl/definitie/large-language-model-llm

Dienst Toeslagen Van wetgeving naar uitvoering onder architectuur. (z.d.). In *IBS Toeslagen*. IBS Toeslagen.

Jansen, J. (2023). *Lange termijn plannen IBS Toeslagen*. IBS Toeslagen.

gpt 4 paper. (z.d.). Openai. https://cdn.openai.com/papers/gpt-4.pdf

Wat is een Vector Database en waarom belangrijk voor Generatieve AI? (z.d.). https://www.ai.nl/insights/wat-is-een-vector-database-en-waarom-belangrijk-voor-generatieve-ai

GeeksforGeeks. (2024, 13 februari). *Separation of Concerns (SoC)*. GeeksforGeeks. https://www.geeksforgeeks.org/separation-of-concerns-soc/

Ggailey. (2023, 24 mei). *Azure Functions Overview*. Microsoft Learn. https://learn.microsoft.com/en-us/azure/azure-functions/functions-overview?pivots=programming-language-csharp

What is a vector database. (z.d.). Cloudflare. https://www.cloudflare.com/learning/ai/what-is-vector-database/