ONDERZOEKSVOORSTEL

Toepassen van een tijdelijke grafiek transformatie op object traceringsdata voor het trainen van een LLM.

Bachelorproef, 2024-2025

Maarten Van der Schueren

E-mail: maarten.vanderschueren@student.hogent.be
Co-promotor: B. Peirens (Tracked, bart.peirens@tracked.be)

Samenvatting

In deze bachelorproef onderzoeken we hoe we een grafiekmodel kunnen gebruiken om een LLM te ontwikkelen, waarbij we bij bijvoorbeeld klachtenafhandeling snel en efficiënt de oorzaak kunnen gaan vragen. Het onderzoek is verdeeld in twee fases: de eerste fase gaat over het omzetten van events volgens EPCIS naar een grafiekmodel in cosmos DB met behulp van Gremlin API. In de volgende fase trainen we een LLM die ons kan vertellen waar, wanneer, wat en hoe een event heeft plaatsgevonden en anomalieën in het proces kan terugvinden. Het gebruik van deze LLM kan tijds- en geldbesparend zijn door handmatig zoekwerk te onderdrukken. Het is belangrijk in dit onderzoek dat alles schaalbaar en performant is voor zowel bedrijven, nationale of landelijke business use cases. Door deze implementatie verwachten we de klantenervaring en efficiëntie te optimaliseren.

Keuzerichting: AI & Data Engineering
Sleutelwoorden: LLM, OTG, EPCIS, CosmosDB, Assistant Agents

Inhoudsopgave

1. Inleiding

Binnen Arcelor Mittal Gent bestaan er veel verschillende processen die verspreid zijn over meerdere afdelingen. Hierbij bevat elke afdeling een deel van het proces om van grondstof tot een afgewerkt product te komen. Tussen elke afdeling bevindt zich een deel van de toeleveringsketen, waarbij het resultaat van de ene afdeling het beginpunt is voor de volgende afdeling. Dit zorgt ervoor dat iedere stap zijn invloed heeft op de kwaliteit van het eindresultaat.

Elke afdeling is opgebouwd rond een specifiek proces, zoals de hoogoven die ruwe grondstoffen omzet in ruw ijzer en de staalfabriek die ruw ijzer in staal verandert. Omdat deze afdelingen zeer specifieke processen hebben, hebben ze ook zeer specifieke behoeften en datamodellen. Dit maakt analyse of onderzoek over afdelingen heen zeer complex en tijdsrovend.

De scope van deze bachelorproef is tweedelig, waarbij deel één eruit bestaat om samen met Tracked vanuit een gestandariseerd datamodel die zichtbaarheid events bijhoudt om te zetten naar een grafiekmodel. Deel twee bouwt verder voort op dit grafiekmodel om dan een LLM (large language model) te trainen die eenvoudig informatie eruit kan halen en teruggeven.

De centrale vraag die hier gesteld wordt is: “Hoe we efficiënt en snel grafiekmodellering kunnen toepassen om een LLM te ontwikkelen die in staat is om het waar, wanneer, wat en hoe van gebeurtenissen binnen een proces vast te stellen, ter ondersteuning van klachtenafhandeling bij productfouten.“

2. Literatuurstudie

2.1. Cosmos DB

2.1.1. Waarom Cosmos DB gebruiken

Cosmos DB is een NoSQL-database van microsoft.Het biedt een lage latentie, multi-query-api die eenvoudig grote hoeveelheden data kan verwerken en heeft een grote beschikbaarheid zegt van der Put (2020), wat zeer belangrijk is in ons project. Daarnaast is CosmosDB horizontaal schaalbaar, wat betekent dat we op hoogtepunten tot een miljoen lees- en schrijfaanvragen kunnen verwerken door het benodigde aantal servers toe te voegen. De hoge beschikbaarheid wordt

gegarandeerd door replicatie, waardoor we snel kunnen overschakelen als er een probleem is in onze database.

2.2. Grafiek modellering

In een grafiekdatabase slaan we gegevens op in de vorm van verices (knopen) en relaties (edges), hierdoor kunnen we complexe verbanden leggen zoals in ons geval de productieprocessen. Cosmos DB ondersteunt ook grafiekmodellering aan met behulp van Gremlin API. Gremlin is gebasseerd op Apache tinkerpop's Gremlin, dit is een krachtig grafiekverwerkingsframework waarbij we grote grafieken kunnen opslaan, modelleren en doorzoeken via gremlin traversal language (Microsoft, 2024). Dit gebeurt met een snelheid van milliseconden waardoor we een snelle verwerkingstijd zullen kunnen neerzetten. Daarnaast is Gremlin ook schaalbaar en kan het consistentie level gekozen worden om een balans te vinden tussen consistentie, beschikbaarheid en latentie. Binnen zo'n grafiek model hebben we Vertices (of nodes) die een persoon, plaats of event beschrijven zoals in ons geval bijvoorbeeld een slab die wordt verplaatst van A naar B.

2.2.1. Verschil transformatie en aggregatie

Binnen ons project kunnen de bestaansvormen van producten veranderen. Zo hebben we transformatie, dit is een deel van het proces waarbij de verandering van het product onomkeerbaar is. Daarnaast hebben we aggregatie waarbij we wel terug naar de oorspronkelijke vorm zouden kunnen gaan indien nodig. In onze grafiekmodellering moeten we gebruik maken van tijdelijke edges om ook terug te kunnen kijken naar wat voor het transformatie- of aggregatieproces gebeurd is, om zo anomalieën terug te vinden (Jaewook Byun, 2020).

2.3. Retrieval Augmented Generation (RAG)

Om onze chatbot te optimaliseren gaan we gebruik maken van RAG. Dit zorgt ervoor dat de antwoorden die de chatbot zal geven niet outdated zijn en een geldige bron bevatten (zoals onze eigen data) (IBM, 2023). Binnen een LLM waarbij je RAG gebruikt zal de user een prompt geven, daarna gaat het LLM vragen aan RAG om alle bronnen (data) te geven over het gevraagde onderwerp. Zodra hij jouw prompt samen heeft gevoegd met de data die erover beschikbaar is geeft hij een correct antwoord terug. In ons geval kunnen we bijvoorbeeld vragen waar de container staal was op een bepaalde datum, en op dat moment vraagt het LLM aan RAG alle informatie over die container waarna hij een geschikt antwoord kan geven. Deze mehtode wordt ook gebruikt in onder andere ChatGPT en andere LLM's om ervoor te zorgen dat je het model zelf niet herhaaldelijk moet trainen, maar een up-to-date dataset aan koppelt. Kort gezegd is RAG een AI framework dat data gebruikt buiten zijn eigen getrainde data om te voorkomen dat je het model moet blijven updaten.

2.3.1. RAG vs Finetuning

Bij RAG gebruiken we een soort database waarbij het model up-to-date blijft en geldige bronnen gebruikt. Dit is voordelig in ons project omdat er dagelijks heel veel staal geproduceerd wordt en dus veel verschillende processen zijn. Daarnaast zegt IBM (2024), hebben we finetuning waarbij we onze eigen stijl kunnen geven aan het model door bepaalde vakjargon toe te voegen of het model sneller te maken. Maar er moet wel zorgvuldig met finetuning omgegaan worden, want het kan snel veel rekenkracht vragen of zelfs tot overfitting raken waarbij de trainingsdata goed presteerd maar ongeziene data niet zegt EASIIO (2022). Een combinatie van deze 2 methodes zou in ons scenario ideaal zijn. Dit omdat Arcelor Mittal veel vakjargon heeft zoals cokes, slabs,...en de data die we gebruiken specifiek en up-to-date moet zijn om van elk moment op de hoogte te blijven.

2.4. EPCIS

Voor dit onderzoek moet alles voldoen aan de EPCIS (Electronic Product Code Information Services) waarden, dit zijn de wat, wanneer, waar, waarom en hoe. Deze waarden zijn ontwikkeld door GS1 om gegevens over beweging, status en verandering van een item in de toeleveringsketen (supply chain) vast te leggen en te delen binnen en buiten het bedrijf (Devins e.a., 2022). "Met behulp van deze waarden kunnen we real-life objecten omzetten in elektronisch opgeslagen informatie, waarna we dit kunnen communiceren met eindgebruikers." zegt Devins e.a. (2022). Door deze normen toe te passen kunnen we de traceerbaarheid van het product per proces garanderen inclusief de gewenste parameters die opgeslagen worden in ons grafiekmodel zoals tijd (wanneer) en temperatuur (hoe) waar nodig.

2.4.1. Waarom EPCIS?

Huidige Legacy Systemen maken gebruik van ERP, POS, WMS,...Vaak lopen deze niet real-time wat nadelig is als we direct iets willen weten over een product en hoge beschikbaarheid verwachten (Vieweger, g.d.). Doordat iedereen eenzelfde norm gebruikt kunnen we dit schaalbaar houden en makkelijk uitbereiden.

3. Methodologie

Eerst en vooral gaan we aan de slag met een json bestand dat data bevat van Arcelor Mittal waar events in opgeslagen staan. Daarna zetten we dit om naar Cosmos DB en kunnen we via

Gremlin API werken om een grafiek op te zetten met tijdsgebonden edges. Dit dient ervoor om te zorgen dat we in tijd terug kunnen om de historiek op te vragen en te kijken waar, wanneer en hoe een proces gebeurd is. Daardoor kunnen we achteraf bij onder andere klachtenafhandeling terug kijken in het grafiekmodel wat er mogelijks anders was of fout is gelopen.

Dit grafiekmodel moet ook voldoen aan de normen volgens EPCIS en schaalbaar zijn, daarom gaan we zorgen dat de gebruikte data binnen het grafiekmodel zoals een node een id krijgt in de plaats van een statische naam. Daardoor kan Tracked in latere fases of andere business cases dezelfde technieken gebruiken zonder veel aanpassingen te moeten doen op grote schaal.

Zodra we een duidelijk grafiekmodel hebben waarmee we makkelijk in het verleden kunnen kijken, kunnen we dit trainen met copilot van microsoft. Daarna gaan we op zoek naar de beste methode om dit te implementeren met onder andere hulp van de RAG methode en finetuning. Daarnaast gaan we ook nog op zoek naar een geschikt zoekalgoritme om zo snel mogelijk op de juiste data terecht te komen. Daardoor kunnen we dan snel zoeken naar waar mogelijks anomalieën zijn in een proces door te vergelijken met andere batches die hetzelfde proces hebben doorlopen. Als laatste kunnen we dan via copilot makkelijk een implementatie starten voor onder andere microsoft teams waarbij iemand van de klantenservice kan vragen wat er fout is gelopen en dat de chatbot zegt waar, wanneer en hoe er mogelijks een fout is geweest in het proces.

4. Verwacht resultaat, conclusie

In dit onderzoek maken we startlaag aan een groter project waarbij het mogelijk is om uit te bereiden naar bijvoorbeeld een planningstool om de meest efficiënte route te zoeken. Door deze chatbot verwachten we dat medewerkers van Arcelor Mittal eenvoudig een vraag kan stellen en een antwoord kan krijgen op een korte tijd. Zonder deze optie moeten ze elk deel van het bedrijf handmatig opbellen en vragen wat er gebeurd is, dat zijn zaken die zeer tijdsslopend zijn en veel geld kosten. Ik hoop ook dat we dit schaalbaar kunnen opstellen en performant kunnen houden waardoor het ook kan gebruikt worden op landelijk of internationaal niveau. Door de implementatie van dit systeem hoop ik dat we klanttevredenheid en besparingen binnen een bedrijf (in ons geval Arcelor Mittal Gent) kunnen optimaliseren

Referenties

Devins, C., e.a. (2022, juni). *EPCIS standard*. Verkregen november 13, 2024, van https://ref.gs1.org/standards/epcis/

EASIIO. (2022). *LLM De kracht van grote taalmodellen ontketenen*. https://www.easiio.com/nl/fine-tune-llm/

IBM. (2023, augustus 23). *What is Retrieval-Augmented Generation (RAG)?* https://youtube.com/watch?v=T-D1OfcDW1M&t=18s&ab_channel=IBMTechnology

IBM. (2024, september 10). *RAG vs. Fine Tuning*. https://youtube.com/watch?v=00Q0G84kq3M&ab_channel=IBMTechnology

Jaewook Byun, D. K. (2020). *Expert Systems With Applications*.

Microsoft. (2024, augustus 22). *What is Azure Cosmos DB for Apache Gremlin*. Verkregen november 14, 2024, van https://learn.microsoft.com/en-us/azure/cosmos-db/gremlin/introduction

van der Put, M. (2020, augustus 28). *Azure Cosmos DB*. Verkregen november 14, 2024, van https://humandigital.nl/nieuws-en-artikelen/azure-cosmos-db/

Vieweger, T. (g.d.). *EPCIS: Hoe zichtbaarheid van de voorraad verder gaat dan de winkelmuren*. Verkregen november 14, 2024, van https://www.nedap-retail.com/nl/epcis-hoe-zichtbaarheid-van-de-voorraad-verder-gaat-dan-de-winkelmuren/