

12. DECEMBER 2022

AU I IT - AFGANGSPROJEKT

UDVIKLING AF LOGSEARCHER

JAKOB VIGGO HANSEN

DANMARKS RADIO
Copenhagen Business Academy

Indhold

1	INDLEDNING	3
1.1	Problemstilling.....	3
1.2	Problemformulering.....	3
1.3	LogSearchers bestanddele	3
1.3.1	Grafisk Brugerflade.....	3
1.3.2	Database.....	3
1.3.3	Søge-træer	4
1.3.4	LogSearcher mål-systemer	4
2	TEKNOLOGI - OVERBLIK.....	4
2.1	Miljø - udvikling	4
2.1.1	Computersprog – historisk overblik.....	4
2.1.2	Moderne udviklingssprog - 3 kandidater.....	4
2.1.3	Udviklingssprog - opsummering	5
2.2	Miljø - brugerplatform.....	5
2.2.1	Brugerplatform - konklusion.....	6
2.3	.NET i DR.....	6
2.4	Afviklingsmiljø.....	6
2.4.1	VM's vs. containers.....	6
2.4.2	Virtual Machines @ DR.....	7
2.4.3	Containere @ DR.....	8
2.4.4	VM's vs. Containers - Konklusion.....	8
2.5	Databaseteknologi - RDBMS vs. NoSQL	8
2.5.1	Relationel Database Management Systems	8
2.5.2	NoSQL.....	9
2.5.3	Skalering	9
2.5.4	Valg af database - konklusion	9
3	TEKNOLOGI - LOGSEARCHER	10
3.1	Primære udviklings-komponenter	10
3.1.1	eDSL – embedded Domain Specific Language	10
3.1.2	Blazor Server til GUI.....	10
3.2	Drifts-miljø.....	10

3.3	Data-repræsentation – fra kilde til database.....	10
3.4	LogSearchers database.....	12
4	METODE: SYSTEM ANALYSE	14
4.1	ICONIX begreber og processer	14
4.1.1	Funktionel Beskrivelse (trin 1)	14
4.1.2	Domæne Model (trin 2).....	15
4.1.3	Use Cases Packages (trin 3)	16
4.1.4	Grafisk Interface – GUI (trin 4)	18
4.1.5	Use Cases (trin 5).....	18
4.1.6	Robustness Analysis (trin 6).....	19
4.2	System analyse - konklusion	20
5	METODE: UDVIKLING	21
5.1	MVC @ LogSearcher.....	21
5.2	Søgning i store datasæt.....	21
5.2.1	BigO notation & Time Complexity	22
5.2.2	Overblik over nogle indekseringsstrategier	22
5.2.3	Konklusion: Indekseringsstrategi	23
5.3	Udvikling af søgemotor	23
5.3.1	LogParser komponenter	23
5.3.2	Søgning med LogParser	24
6	AFRUNDING	25
6.1	En prototype - afgrænsning af opgavens omfang	25
6.1.1	LogWatcher i WinForm.....	25
6.1.2	LogStore i MSSQL	26
6.1.3	LogParser.....	26
6.1.4	LogSearcher prototype @ Github.....	26
6.2	Afsluttende konklusion.....	26
6.2.1	Produktions version – mod fuld funktionalitet.....	27
6.3	Referencer.....	27

1 INDLEDNING

Jeg arbejder på DR, hvor jeg udvikler applikationer og workflows som understøtter de indholdsskabende afdelinger. Der har jeg haft lejlighed til at erfare at et systems nytteværdi i høj grad er afhængig af hvordan det driftes og supporteres. Når der før eller siden opstår et problem, er muligheden for at spore fejlkæder afgørende for at driftspersonalet hurtigt kan udbedre fejlen.

1.1 Problemstilling

Virksomheder hvis virke er baseret på adskillige samarbejdende IT-systemer, har en stor udfordring med overvågning af drift og stabilitet. Mange systemer kan udsende alarmer i tilfælde af driftsforstyrrelser eller stop. Men enkeltstående alarm-meldinger vil sjældent vise det samlede billede af det påvirkede system. Det er ofte vanskeligt at udlede kontekst og underliggende drift status.

Desuden er diskrete systemer næsten altid blot komponenter i en længere kæde, et workflow eller data-flow. En fejl i ét system skyldes ofte fejl i systemer placeret tidligere i kæden. Systemer senere i kæden kan lide under det aktuelle systems fejl.

Systemerne genererer drifts-logs, ét sæt pr. applikation, ofte adskillige tusind linjer pr. time. Disse logs er nøglen til at udlede et retvisende billede af systemets tilstand, aktuelt og bagud i tiden.

Men det kan være uoverskueligt og tidskrævende at gennemlæse mange tusind loglinjer, og vanskeligt at sammenstille de informationer som fremsøges, **på tværs af tid og systemer**. Overblikket kan drukne i mængden af data.

1.2 Problemformulering

Jeg foreslår at udvikle en applikation, **LogSearcher**, som skal facilitere et hurtigt overblik over flere systemers tilstande, ved at gennemse og analysere deres logs.

LogSearchers brugere er personer som skal supportere og drifte virksomhedens system-park. LogSearcher skal hjælpe dem med at søge efter tekst-billeder på tværs af mange systemers logs. Baseret på søgeord og tidsgrænser skal LogSearcher gøre det nemt at danne sig et overblik over hændelser og deres tidlige distribution.

1.3 LogSearchers bestanddele

1.3.1 Grafisk Brugerflade

LogSearcher skal have en **grafisk brugerflade** (GUI). GUI'en skal facilitere **indsamling** af logs fra forbundne systemer, **gennem søgning** af logs på basis af søge-termer og tidshorisonter, og **udstille** resultaterne fra søgningerne.

1.3.2 Database

Forud for selve søgningen skal de indsamlede logs persisteres til en database, som derefter vil udgøre det data-grundlag som LogSearcher skal traversere.

1.3.3 Søge-træer

Søge-algoritmerne skal være baseret på trie's og binære søgetræer, for at sikre god performance når brugeren potentielt skal søge over mange tusinde linjer rå log-data. Søge-komponenten bør udstille et *eDSL til at facilitere søgning i domænet.

Søge-resultater kan udstilles som tekst-udsnit fra søge-hits, som f.eks:

- top-5 af systemer med søgeordene repræsenteret i deres logs.
- et grafisk billede af den tidslige distribution af et sæt søgeord (f.eks. over de seneste 48 timer)

Ved hjælp fra LogSearcher kan brugerne altså hurtigere danne sig et overblik over begivenheder på tværs af systemer, og få hjælp til at danne en forståelse af hvor de underliggende problemer kan være opstået i den lange kæde af indbyrdes forbundne systemer.

**embedded Domain Specific Language*

1.3.4 LogSearcher mål-systemer

Det er relevant at analysere log-data fra alle de DR-systemer som generer logs. Aktuelt er der mere end 100 systemer.

2 TEKNOLOGI - OVERBLIK

2.1 Miljø - udvikling

2.1.1 Computersprog – historisk overblik

I de allerførste computersystemer som opstod i i 1940-50, var der ingen abstraktion mellem hvordan hardwaren var konstrueret, og hvor systemerne skulle programmeres. Colossus, som brød den tyske Kriegsmarine's kodesystem, Enigma, var specialbygget til dette ene formål, og kunne aldrig udføre andre opgaver end denne ene: at bryde og dechifrere kodede meddelelser. (Colossus, 2022)

Verdens første (elektroniske) "general purpose" computer er generelt anerkendt til at være ENIAC, som blev sat i drift i 1945. ENIAC blev programmeret direkte i binær assembler. (ENIAC, 2022)

Siden da er der opstået adskillige niveauer af abstraktion mellem hardwaren, og den software som vi skriver og anvender til at løse opgaver. CPU'erne har komplekse og omfattende instruksionssæt, som styresystemet kalder for at udstille højere funktioner til applikationerne. (HAL, 2022)

Moderne softwareudvikling sker hovedsageligt på højt niveau af abstraktion. Her anses det som hovedregel for vigtigt, at kode kan afvikles på så mange mål-platforme som muligt. Dette var bl.a. visionen bag udviklingen af Java, "Write Once, Run Anywhere" (WORA, 2022)

2.1.2 Moderne udviklingssprog - 3 kandidater

Fordelt over de 3 programmerings-moduler jeg har taget under uddannelsen, har jeg arbejdet med først **JavaScript**, derefter **C#/.NET** og senest **Python**. De tre sprog ligger pr. 2022 blandt Top-7 Mest populære/anvendte sprog. (Top Programming Languages, 2022)

De tre sprog har som fællestræk at programmets forbrug af hukommelse under kørsel håndteres i baggrunden, af det miljø som sættes op for at afvikle koden. Denne proces kaldes Garbage Collection

(GC), og betyder at allokeret hukommelse automatisk frigøres når et objekts livs-cyklus er slut. (Automatic Memory Management, 2022) (Languages, 2022)

Eksempler på sprog-miljøer uden GC er C og C++, som derfor også foretrækkes til problem-domæner hvor der er begrænset hukommelse eller krav om tids-kritisk eksekvering.

2.1.2.1 JavaScript

repræsenterer det type-svage sprogdomæne, også kaldet dynamisk typing. Det er et fortolket sprog, dvs. det omsættes til maskinvendt kode linje-for-linje under kørsel. En deklareret variabel kan ændre indholds karakter i løbet af sin levetid, dvs. igennem programmets køretid. Variablen A kan starte med at refererer heltals-værdier, dernæst en tekst-streng, og senere et komplekst objekt. JavaScript er i særdeleshed udbredt som de facto normen for at kode interaktivitet på web (sammen med html/css) bl.a pga. dets kompakte og fortættede syntax. I den kontekst køres JavaScript i brugerens webbrowser.

2.1.2.2 C#

repræsenterer det "modsatte" domæne, det type-stærke, kaldet static typing. Et C#-program bliver kompileret til maskine-nær kode før eksekvering. I C# kan alle variable udelukkende referere den datatype de deklarerer mod. C# er stærkt object-orienteret. Det blev udviklet som Microsofts svar på det oprindelige objektorienterede Java fra Sun/Oracle. (Java, 2022)

C# er knyttet til kørsel- og udviklings miljøet .NET, som er Microsofts bud på et kode økosystem til Windows-plattformen (.NET, 2022). I de senere år er .NET Framework blevet afløst af .NET Core, som også er egnet til udvikling mod OSX (Apple), Linux og Android (Google).

2.1.2.3 Python

kan med god vilje placeres imellem JavaScript og C#. Det kan kodes både i et objektorienteret og et funktionelt paradigme, og er typesvagt. Python-syntax er en meget kompakt kodelinje, og er særligt populært til at processere store datamængder i få linjer kode. Python-fortolkere findes i varianter egnet til de mest populære miljøer, Windows, OSX og Unix/Linux. (C# vs Python, 2022), (C# vs Python 2, 2022), (Python for Science, 2022)

2.1.3 Udviklingssprog - opsummering

Hvert enkelt sprog har specifikke styrker og fordele. Python giver mulighed for meget elegant og effektiv rekursion som er vigtige for at bygge de node-baserede søge-træer som udgør søgemotoren. C# understøtter multi-threading (Multithreading in .NET, 2017) indenfor samme app-instans, men det gør Python til gengæld ikke. (TowardsDataScience, 2020) (Python code parallelization, 2016)

2.2 Miljø - brugerplatform

En væsentlig overvejelse i forbindelse med valget af teknologi-plattform for LogSearcher, vil være at undersøge hvilke miljøer den skal eksistere i. En applikation består ikke bare af sin kode, men eksisterer også i kraft af det miljø som understøtter den.

Hvilken platform skal den køre på? - Windows, *Nix, OSX ?

Skal den afvikles lokalt på brugerens klient eller centralt fra en server?

Hvem håndterer daglig drift og problemløsning hvis der opstår fejl?

Kan den virtualiseres eller måske køres i en cloud-løsning?

I DR eksisterer Windows, Linux og Apple OSX side-om-side, og bliver anvendt af overlappende brugergrupper. Alle servere er enten *Nix (Unix-like, 2022) eller varianter af Windows Server, mens klienterne enten kører Windows eller OSX.

Jeg ønsker at brugere med både Win og OSX-klienter skal kunne anvende LogSearcher.

2.2.1 Brugerplatform - konklusion

Det gør det meget oplagt at beslutte at **kernen af applikationen skal afvikles centralt fra en server, og at grænsefladen præsenteres via en web-browser**, baseret på en HTML-side. Derved undgår jeg at skulle udvikle OS-specifikke klienter. Vedligehold og fejlretning kan også bedre fokuseres, hvis applikationen ikke er distribueret over en antal klienter, som alle skal holdes opdateret.

2.3 .NET i DR

På DR anvendes flere forskellige teknologier. Som Public Service Broadcaster (Public Broadcasting, 2022) midt i en digital transition mod web, er en meget stor del af husets **seer/bruger**-rettede kodebase i JavaScript og understøttende frameworks som React og Redux. Men de fleste af de **internt** rettede egenudviklede applikationer er baseret på MS .NET og skrives i C#.

Med få undtagelser, er de fleste af de fundamentale produktions-systemer i huset baseret på off-the-shelf produkter. Men integrationslagene mellem produktion-systemerne er i høj grad egenudviklede. Når data flyttes fra ét system, f.eks. programplanlægningsværktøjet WHATS'ON til programafviklingen (publiceringsplatformen), sker det igennem værktøjer og processer som DR selv udvikler. Den udvikling er baseret på MS .NET og C#.

LogSearcher vil indgå naturligt som et støtteværktøj for personale som varetager driften af integrationslagene. Et godt værktøj skal kunne tilpasses løbende til en dynamisk og foranderlig kontekst, og der må forventes løbende udvikling og tilpasning, også efter idriftsættelse.

Ved at vælge MS .NET som udviklingsmiljø, skabes det bedste potentiale for at forankre værktøjet hos brugerne.

2.4 Afviklingsmiljø

Med en klar beslutning om at basere LogSearcher på en klient/server arkitektur, er næste spørgsmål hvordan miljøet omkring serveren skal bygges.

2.4.1 VM's vs. containers

I moderne IT-drift er det ikke længere praksis at afvikle applikations-instanser på diskret hardware, dvs. en fysisk server pr. applikation/system. I stedet har man en indført op til adskillige lag af abstraktionsniveauer over hardware-niveauet. Det giver mulighed for at afkoble afhængigheden til specifik hardware, og i stedet introducere en generaliseret arkitektur.

Overordnet set er der i dag 2 grene af abstraktion, **Virtual Machine (VM)** og **containers**. Container-teknologier som Docker og Kubernetes bliver oftest anset som cloud-baserede. VM's har hos DR indtil nu været baseret på OnPremise-hardware.

Cloud-computing er et paradigme der faciliterer omfattende adgang til en central pulje af ressourcer. Det har traditionelt været forbundet med VMs, men containers er ved at vinde frem. Man kan betragte containers som letvægts VMs. (Containers Versus Virtual Machine, 2018)

2.4.1.1 Virtual Machine

En VM er i realiteten en simulering af en komplet PC, dvs. hardware som CPU, RAM, storage osv. bliver simuleret for den applikation som skal køres. Ovenpå hardware-simulationen afvikles værts-styresystemet (OS), typisk en variant af MS Windows. Denne løsning giver mulighed for at afvikle adskillige, isolerede instanser af en VM mod samme fysiske hardware. En VM vil typisk være vært for ét komplet system, på tværs af mange processer.

2.4.1.2 Container

En container er et isoleret og relativt letvægts miljø som afvikles på en til formålet egnet platform. Denne platform betegnes som kernel, og sidder oven på enten fysisk eller simuleret hardware. En container sidder så at sige oven på værts-OS, uden at have sit eget dedikerede OS. En container-instans vil normalt kræve færre ressourcer fra værts-systemet end en VM. En meget brugt teknologi til containers er Docker. En container vil som hovedregel kun være vært for én proces eller applikation

2.4.1.3 Isolering som strategi

En stor del af pointen med disse abstraktioner er, udover at optimere brugen af hardware-ressourcer, at opnå isolation mellem instanserne. At undgå at problemer der opstår i én instans, kan have effekt på nabo-instanser. Man taler om at VM vs. containers giver forskellige typer af isolation:

VM: Isolation of the machine
Container: Isolation of the process

2.4.2 Virtual Machines @ DR

Sammendrag af interview med Thomas Borup, DR, MS Infra-team.

Thomas og hans team håndterer drift af alle IT-systemer i DR, både administrations- og produktions-systemer. Deres domæne er Virtual Machines.

2.4.2.1 DR i skyen

"DR har for nyligt besluttet af være "Cloud First", efter at have været OnPremise-orienteret."

OnPremise indebærer at den fysiske hardware er servere placeret i bygningen. Cloud er i DR-kontekst lig med Microsoft Azure. Det indebærer at nye systemer i videst muligt omfang skal deployes til Azure. Azure giver mulighed for både at instantiere komplette VMs, eller modulære komponenter som f.eks en SQL-database. Komponent-tilgangen er favoriseret, fordi der er mindre ressource-overhead på et drifte en komponent snarere end end komplet VM.

"Problemet med Cloud instanser er at latency (svartider) er markant højere mellem lokal klient og cloud-server. Dvs. meget granuleret trafik med mange forespørgsler/svar akkumulerer forsinkelse. Det betyder at vi skal være opmærksomme på om cloud-servicen har meget trafik fra OnPremise systemer, fordi vi i så fald kan opleve nedsat performance."

2.4.2.2 Hardware

"Mit team håndterer 442 fysiske serverer, og 1065 virtuelle maskiner. Af de 442 fysiske servere er 40 værter for virtualiseringsmiljøet. Vi anvender Microsoft HyperV til drift af windows-systemer, og VMWare til linux. I dag bruger vi MS Virtual Machine Manager til håndteringen"

Der er planlagt et teknologiskifte, som indebærer at det eksisterende miljø udskiftes til fordel for MS Azure Stack HCI (Microsoft, 2022), og så kan ressourcer provisioneres mod både Cloud og OnPremise i samme driftsmiljø. Dermed bliver det praktiske skel mellem de to domæner gradvist udvisket.

2.4.3 Containere @ DR

Sammendrag af interview med Sigurd Kristensen, DR, Linux-team.

2.4.3.1 Ny platform

"Vi er ved at bygge vores nye platform. Det bliver Kubernetes-baseret, med adskillige clusters til test, produktion osv. Vi bygger vores deploy-pipeline sammen med DevOps-teamets, dvs. vi får fordel af fælles infrastruktur. Vi vil tilbyde både OnPremise og Cloud-hosting."

Kubernetes er et system til håndtering og drift af container-samlinger og Docker er det mest udbredte container-system. Tilsammen udgør de en platform som giver stor fejl-tolerance, horisontal skalérbarhed og isolation mellem klient system-domæner. (Kubernetes: Up and Running, 2022)

2.4.3.2 Stateless vs. Stateful

"DRs container-strategi indebærer at vi foretrækker 'StateLess' fremfor 'StateFull'. Vi ønsker at kunne rive hele containeren ned når vi re-deployer en app via vores pipeline"

Stateless-paradigmet betyder at en applikation altid deployes i samme tilstand, dvs. med nøjagtigt samme data. Det gør modellen uegnet til f.eks databaser, som netop er kendetegnet ved at skulle reflektere et systems aktuelle tilstand. (Stateless Vs. Stateful, 2019)

2.4.3.3 Foreslået arkitektur

"Jeg vil foreslå at du baserer LogSearcher på en .NET Core container, og en Azure Cloud host til din database. Brug Docker for Desktop til udvikling og debugging."

2.4.4 VM's vs. Containers - Konklusion

Docker/Kubernetes giver mulighed for at skalere drift ifht. brugsbelastning. Det er derfor oplagt at basere LogSearchers komponenter på en docker-platform. Derimod bør databasen hostes på en VM, pga. det 'stateless'-paradigme som DR foretrækker.

2.5 Databaseteknologi - RDBMS vs. NoSQL

Inden vi beslutter hvilken database-teknologi LogSearcher skal benytte, er det passende at diskutere de grundlæggende, dominerende typer af databaser. På tværs af de mange forskellige databaser-produkter ses der 2 grundlæggende typer: relationel og ikke-relationel.

2.5.1 Relationel Database Management Systems

Begrebet RDBMS opstod i 1970'erne. De relationelle databaser betegnes også SQL-databaser (Structured Query Language). SQL er interaktions-syntaxen for RDBMS. SQL er opstået i samme periode som modellen for relationelle databaser og normalisering, men siden forfinet og udviklet.

Det relationelle aspekt opstår ved at alle data gemmes i tabeller, og hvert data-element tildeles en nøgle. Nøglen anvendes til at udtrykke ét tabel-elements relation til et eller flere elementer i en eller flere andre tabeller. RDBMS/SQL stammer fra en periode hvor storage var en bekostelig ressource, og det var vigtigt at reducere data-redundans mest muligt.

Normalisering sigter efter at minimere redundans (data-duplikering) og data-skrøbelighed. Data er skrøbelige, hvis ændringer i databasen kan føre til inkonsistens og lign. anomalier, dvs. ukomplette og "fragmenterede" datasæt. En vel-normaliseret database har meget høj grad af data-integritet, dvs. at den er robust.

Hvor effektivt data-duplikering begrænses, kan beskrives af graden af normalisering i et RDBMS.

1. NF (Normalform): ingen kolonner i tabellen gentager en anden kolonnes værdi
2. NF: Overholder 1NF og indeholder kun kolonner som afhænger af primær nøgle (PK)
3. NF: Overholder 2NF og ingen felter udenfor PK er indbyrdes afhængige.

En database som er tilstrækkeligt normaliseret vil være nemmere at vedligeholde, ændre og udbygge. Fuld normalisering kan have en negativ effekt på performance, fordi data-sæt er fordelt over mange tabeller, og et meningsfuldt svar skal sammensættes på tværs af disse. Men opvejes til gengæld af fleksibilitet og robusthed.

2.5.2 NoSQL

Begrebet dukkede op i begyndelsen af vores årtusinde. Det betegnes som et ikke-relationelt database-paradigme som uafhængigt af tabeller og deres nøgler. Det betyder at en NoSQL-db potentielt kan være bedre egnet til data-sæt som tekstdokumenter, graph-data o.l.

NoSQL er mere dynamisk orienteret end SQL, og er mere modstandsdygtigt overfor ændringer og tilpasninger i systemets datastruktur, fordi det er knap så betinget af rigide relationer. En NoSQL-baseret database tåler gerne løbende strukturelle modifikationer, i takt med at kravene fra aftager-applikationen ændres. De kan også være meget hurtige, fordi svar ikke skal sammensættes på tværs af mange tabeller.

NoSQL databaser kan under-grupperes i 4 strukturer:

- Kolonne-orienterede: data grupperes i kolonner, med frit antal rækker
- Key/Value: HashMap-baseret lookup
- Dokument-orienterede: egnet til JSON, XML ol.
- Graph-orienterede: repræsenterer data i node-baserede strukturer.

2.5.3 Skalering

I takt med at performance-kravene vokser, kendetegnes de to typer (RDBMS/NoSQL) ved forskellig skalerings topografi. Man kan tale om at RDBMS/SQL skal skaleres vertikalt, dvs. at hvis der kræves mere performance, skal der flyttes til en kraftigere server. NoSQL er egnet til horisontal skalering, dvs. over flere diskrete instanser/servere. (SQL vs. NoSQL vs. NewSQL, 2016)

2.5.4 Valg af database - konklusion

Jeg vil anvende et RDBMS, dvs. MSSQL, bl.a. fordi DR's infrastruktur har givet nem adgang til MSSQL instanser i et VM-miljø. Det er også denne type database jeg personligt har mest erfaring med.

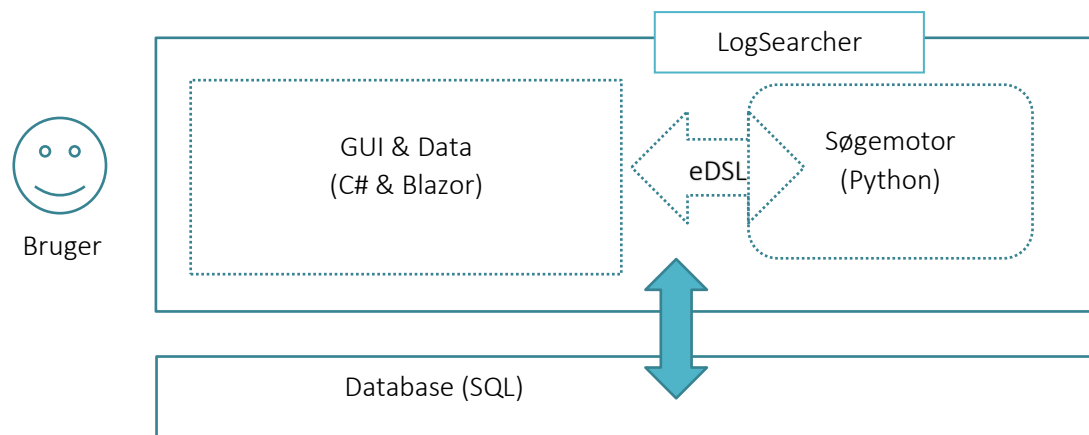
3 TEKNOLOGI - LOGSEARCHER

3.1 Primære udviklings-komponenter

Til LogSearcher vil jeg kombinere C#, Python og MSSQL.

Jeg vil bruge **Python** til de centrale aspekter af søgning i datasæt, søgekomponenten.

Jeg vil bruge **C#** til grafisk brugerflade (GUI) og data-persistering mod databasen.



3.1.1 eDSL – embedded Domain Specific Language

Ved at indbygge et eDSL i søgekomponenten (eDSL, 2022) (Embedded Domain Specific Languages, 2022), kan jeg udstille de primære søge-værktøjer mod det omsluttende C#/.NET lag som udgør den brugervendte applikation.

3.1.2 Blazor Server til GUI

Med .NET Core har Microsoft introduceret et web-orienteret lag til C#, som gør miljøet meget egnet til at udvikle websider. Komponenten hedder Blazor (What is Blazor, 2022), og giver mulighed for en eventdrevet, dobbeltrettet binding mellem grafik-objekter og de underliggende tilstande i applikationen. Blazor er designet til at underbygge SPA-modellen (Single Page Application), hvor brugeren ikke navigerer mellem forskellige sider på samme site, men hvor indholdet af siden i stedet udskiftes og opdateres dynamisk. Jeg vil anvende den hostede model, **Blazor Server**, hvor appen hostes på en IIS eller Kestrel-instans.

3.2 Drifts-miljø

Alle komponenter af LogSearcher skal kunne afvikles i diskrete docker-instanser eller på VMs.

3.3 Data-repræsentation – fra kilde til database

Den fælles kilde-type for alle målsystemer vil være de ustrukturerede rå-data som logfilerne udgør. Dvs. LogSearcher har ikke adgang til strukturerede metadata om logfilens format.

For at sikre en homogen og ensartet kilde til data for søgemotoren, vil jeg vælge at persistere data fra alle målsystemer til en fælles database.

En system-log består af én eller flere linjer af log-events, som mål-applikationen genererer under drift. og de forventes at karakterisere applikationens tilstand. Dvs. at der typisk skrives i loggen, når en væsentlig hændelse har fundet sted.

F.eks. gennemført brugerinteraktion, eller indlæsning af eksterne data.

Eller når der opstår en fejl eller fejltilstand.

På basis af logs fra de 3 eksempler, har jeg identificeret 4 primære elementer som karakteriserer et log-event.

TimeOfEvent: hvornår fandt begivenheden sted

Severity: hvor alvorlig vurderes hændelsen at være, fra simpel information til fatal fejl

Event-description: en mere detaljeret beskrivelse af fejlen og evt. relateret tilstand

Source-module: navnet på det modul som fejlen eller meddelelsen stammer fra.

Målsystemerne anvender forskellige formateringer for logningen. Event-tid (TimeOfEvent) kan f.eks være angivet i UTC eller DK-tid. Niveauet for log-linjens alvorlighed (Severity) kan være angivet umiddelbart efter event-tid, eller slet ikke. Event-beskrivelsen kan indeholde tekst. Der kan evt. optræde en reference til det kode-modul som er kilden til log-linjen.

Her er 3 eksempler fra kurante DR-systemer:

Galaxy Site Selector	2022-11-03 13:05:34.3721081 [INF] @GetDaletServiceIsInstalled, looking for service: "DaletService" (SiteSelector.Domain.Session.SharedAgents.CheckGalaxyInstalled) 2022-11-03 13:05:34.3738653 [WRN] Couldn't find service: "DaletService" - "Service 'DaletService' was not found on computer '.'." (SiteSelector.Domain.Services.ServiceHandler)
	TimeOfEvent er angivet med to komponenter, dato og tidspunkt, adskilt af mellemrum. Tidszone er ukendt. Severity er angivet i næste kolonne, adskilt af mellemrum og omgivet af klammer []. Event-beskrivelsen er i næste kolonne, adskilt af mellemrum. Kildemodul er sidste kolonne, adskilt af mellemrum og omgivet af parenteser ().
MA Ingest	2022-10-04 00:14:29,010 INFO Common.Comms.MaPersist.MaCom.MaRequest [0] - <?xml version="1.0" encoding="UTF-8"?> <fault><reason>Error 401: HTTP 401 Unauthorized</reason><detail></detail></fault> 2022-10-04 00:14:29,010 ERROR Common.Comms.MaPersist.MaCom.MaRequest [0] - MaCom, Un-authorized 2022-10-04 00:14:29,011 INFO MaSync.ProcessTasks.Domain.ProcessTask.ProcessNew [0] - Failed to authorize with user superadmin 2022-10-04 00:14:29,011 INFO Common.DbOps.FailedTasksDbOps.StoreFailedTaks [0] - Storing 0898b419-7cfb-4f39-b044-dba2edda9c94 in failed Tasks.
	TimeOfEvent er angivet med to komponenter, dato og tidspunkt, adskilt af mellemrum. Tidszone er ukendt. Severity er angivet i næste kolonne, adskilt af mellemrum. Kilde-modul er næste kolonne, adskilt af mellemrum. Event-beskrivelsen er resten af linjen.
VizController	2022-11-01T19:25:41.207Z superState Handling state: resultat.sceneOutDone 2022-11-01T19:25:41.209Z superState Error in checkForBlokke: Error: Failed to match currentResultGuid in active playlist. 2022-11-01T19:25:41.209Z superState Checking for next result...
	TimeOfEvent er angivet med to komponenter, dato og tidspunkt, adskilt af mellemrum. Tidszone er UTC* Kilde-modul er næste kolonne, adskilt af mellemrum Event-beskrivelsen er resten af linjen

* (ISO 8601, 2022)

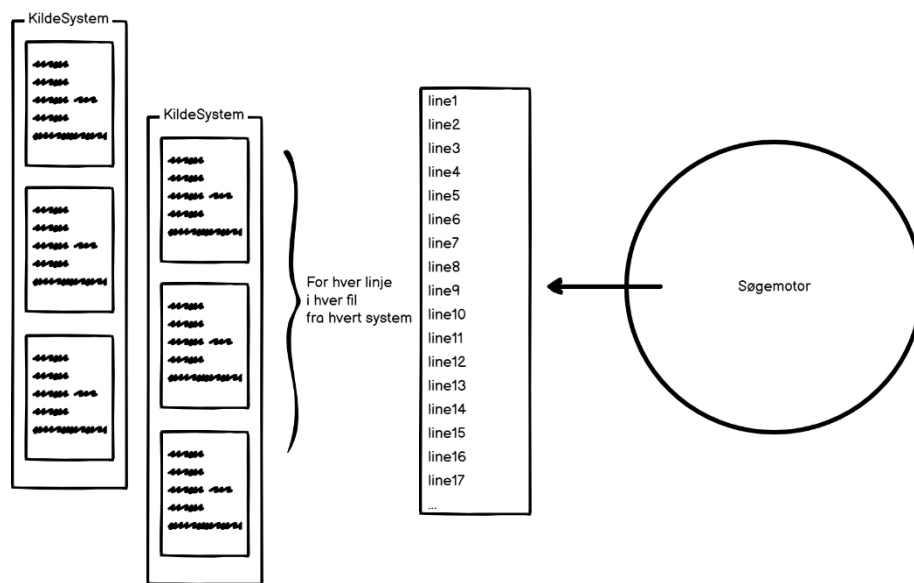
Vores database bør anvende ét TimeOfEvent-format, ét Severity-format, ét format til event-beskrivelse og ét format til data om kilde-modul.

Hvis målsætningen om at databasen skal indeholde homogene data skal mødes, er det derfor nødvendigt at filtrere og re-formatere log-data inden persistering.

3.4 LogSearchers database

Der er to distinkte faser i database-interaktionen. Først skal log-data pre-processes og indlæses i databasen. Derefter skal der læses fra databasen ind i søgemotoren.

For at optimere denne proces og minimere flaskehals-effekten, skal databasens struktur understøtte en effektiv transformation fra kildefiler til data-records, og samtidig facilitere hurtig og smidig tilgang fra søgemotor.



Søgemotoren arbejder på ord-forekomster i tekst-linjerne, og når der er fundet et match, skal det kunne spores tilbage til kildefil og -system.

Det medfører denne tabel-struktur:

Table: **SourceSystems**

ID (int)	Name (string)	SourceFolder (string)	LineTemplate (string)
10	MA-Ingest	C:/logs	0-15^

Table: **LogFiles**

ID (int)	SourceSystemID (int)	FileName	FileHash
1	10	Logfile.txt	xxxxxx

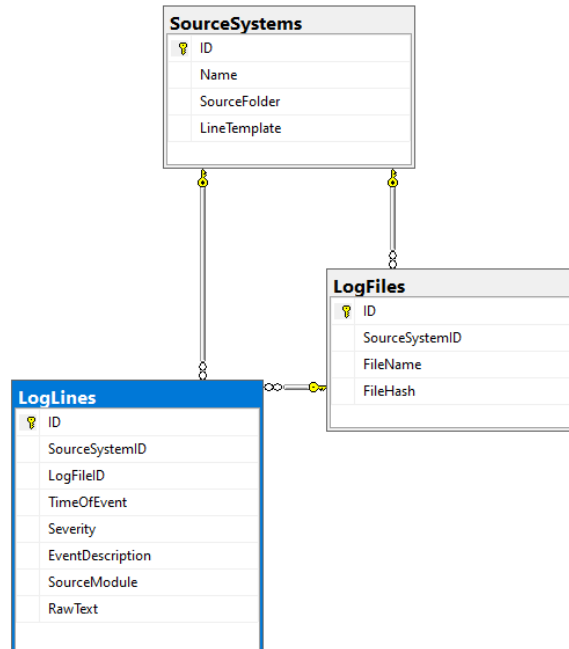
Table: **Loglines**

ID (bigint)	SourceSystemID (int)	LogFileID (int)	TimeOfEvent (dateTime)(nullable)	Severity (string)(nullable)	EventDescription (string)	SourceModule (string)(nullable)	Rawtext (string)
10000	10	1	---	ERROR	Failed while uploading user data	Main.GetData	

Entity Relations diagrammet for databasen.

Tabellen **LogFiles** har en FK til SourceSystems ID, og tabellen **LogLines** har en FK til hhv. SourceSystems ID og LogFiles ID.

Databasen er således fuldt normaliseret til 3. normalform.



4 METODE: SYSTEM ANALYSE

Systemudviklings-fasen er primært orienteret mod det brugervendte lag af LogSearcher.

Jeg vil til denne facet af udviklingen benytte design-metoden ICONIX (Doug Rosenberg, 2007) som det blev gennemgået i CPH modulet "Systemudvikling".

ICONIX metoden er en minimalistisk, effektiv tilgang baseret på UseCase-drevne UML-modeller. Unified Modelling Language – UML (Fowler, 2003), er et sæt af symboler som er velegnede til, på abstrakt niveau, at beskrive og designe processer i software under udvikling.

ICONIX og UML er tæt knyttet til det objektorienterede udviklings-paradigme (OOP, 2022).

4.1 ICONIX begreber og processer

Første trin består i at samle en såkaldt **Funktionel Beskrivelse (trin 1)**, som i klart sprog beskriver hvordan applikationen forventes af præsentere sig overfor brugeren, og hvordan interaktion foregår.

På basis af den funktionelle beskrivelse, kan der dannes en såkaldt **Domæne Model (trin 2)**. Denne bruges til at identificere objekter som er relevante at modellere i softwaren, og ikke mindst deres indbyrdes relation.

Næste trin er at bygge **Use Case Packages (trin 3)**, som i bulletpoint-form lister de primære trin for hver case. Desuden bygges et diagram over trinenes indbyrdes relation og afhængighed.

Dernæst skabes en design-skitse for softwarens **GUI (trin 4)**, gerne i en forsimplet form som lægger vægt på funktion snarere end form. Overvejelse om grænsefladen kan også påvirke strukturen af domæne-modellen.

Nu skrives komplette **Use Cases (trin 5)**, som i klart sprog og abstrakte detaljer beskriver hvordan en facet af brugerens interaktion med softwaren skal foregå.

Robustness Analysis (trin 6) er et ICONIX-specifikt begreb som sigter efter at koble Use Cases og objekter. Vi søger at identificere domæne-modellens objekter i de analyserede Use Cases.

ICONIX-processen lægger op til at hele design-fasen er iterativ, hvor en opdagelse på ét niveau kan påvirke tidligere beslutninger og kræve at de revurderes og tilpasses.

4.1.1 Funktionel Beskrivelse (trin 1)

De med rødt fremhævede navneord er særligt interessante ifht. domæne-modellen:

Vi skal udvikle en web-app som giver sin bruger mulighed for at udvælge et eller flere målsystemer, og gennemse logs fra dette.

Brugeren kan oprette nye målsystemer og tilføje dem til en **målsystem-liste**. Oprettelse indebærer definition af af **kildefolder**, dvs. hvor på netværket findes logfilerne. Brugeren definerer hvordan en **loglinje** fra målsystemet er struktureret, dvs. opretter en **template** for målsystemets **loglinjer**.

Brugeren kan markere et **målsystem** og aktivere indlæsning af logs.

Brugeren kan afgrænse en **søgning** ved at angive ét eller flere **nøgleord** i en **nøgleords-liste** som skal forekomme i en loglinje, og angive et **tidsinterval** som log-begivenhedens **tidspunkt** skal ligge indenfor.

Når søgningen er gennemført, og der er fundet **loghits**, skal disse loglinjer præsenteres for brugeren i en **hitliste**. Hvis brugeren klikker på en linje i hitlisten, vises hele linjen.

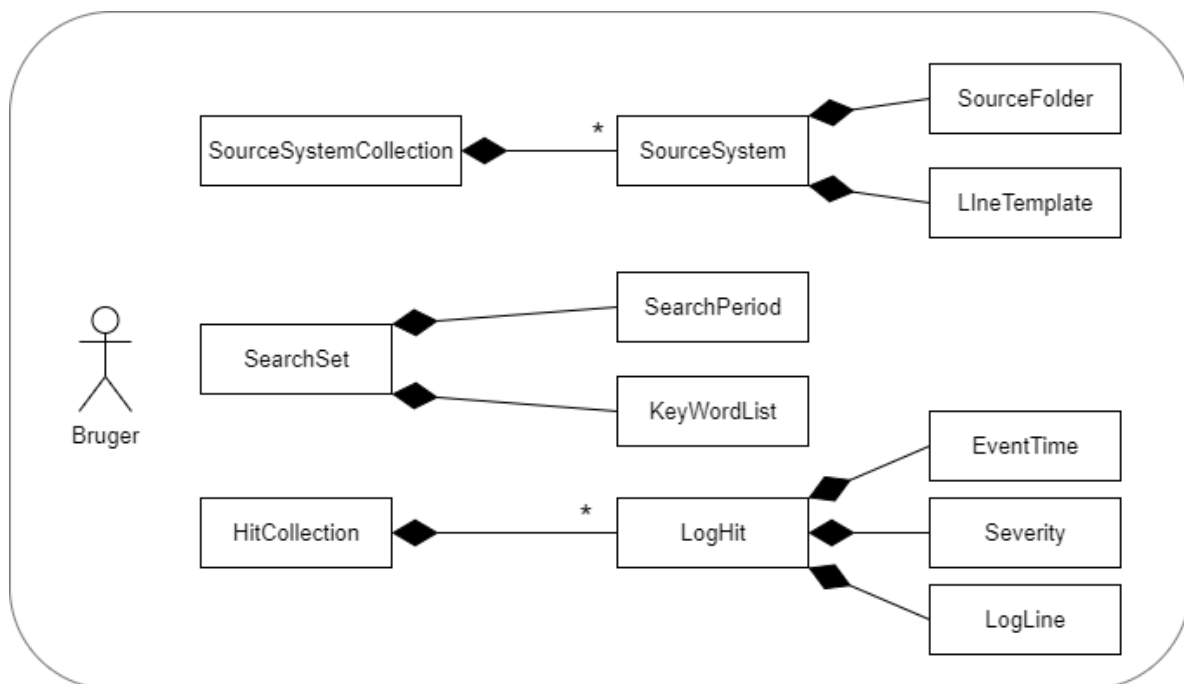
4.1.2 Domæne Model (trin 2)

Den funktionelle beskrivelse identificerer nøglebegreber som er objekt-kandidater til domænemodellen.

Nøglebegreber	Domæne model objekter
Målsystemliste	SourceSystemCollection
Kildefolder	SourceFolder
Loglinje	LogLine
Template	LineTemplate
Målsystem	SourceSystem
Søgning	SearchSet
Nøgleord	KeyWord
Nøgleordsliste	KeyWordList
Tidsinterval	SearchPeriod
Tidspunkt	EventTime
Loghits	LogHit
Hitliste	HitCollection
Severity	Severity

Selve domænemodellen optegnes på basis af disse domænemodel-objekter. Her bliver objekternes relation tydelig.

4.1.2.1 LogSearcher – Domæne model

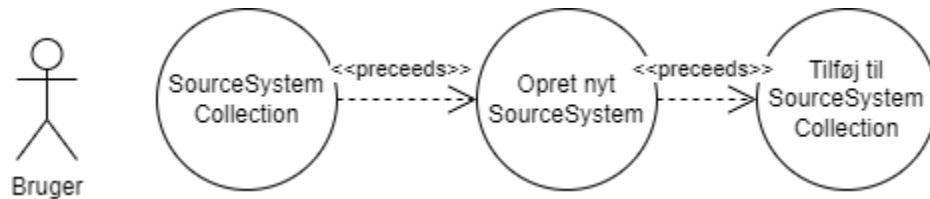


4.1.3 Use Cases Packages (trin 3)

Kortfattet overblik over Use Cases, domæne objekter og interne relationer.

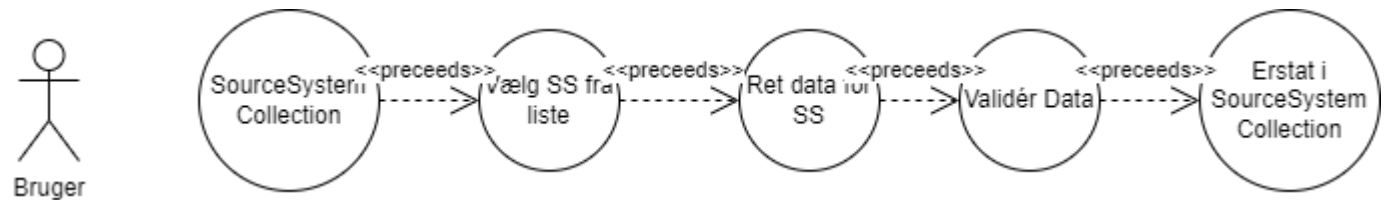
4.1.3.1 Create New Source System

Opret nyt målsystem



4.1.3.2 Edit an existing Source System

Tilføj et nye målsystem til listen



4.1.3.3 Select an existing Source System

Vælg ét målsystem fra listen



4.1.3.4 Update logs for a selected SourceSystem

Vælg ét eksisterende målsystem.

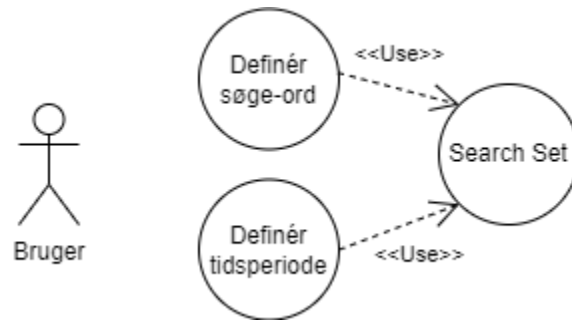
Opdater logs for det pågældende system



4.1.3.5 Define SearchSet for a search

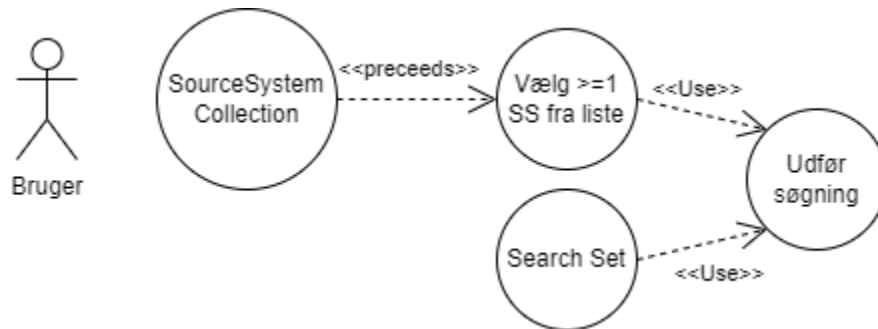
Indtast 0 eller flere søgeord.

Angiv en start og slut-periode for søgningen.



4.1.3.6 Execute search

Brug de anførte søgekriterier til at gennem søge de valgte målsystemer



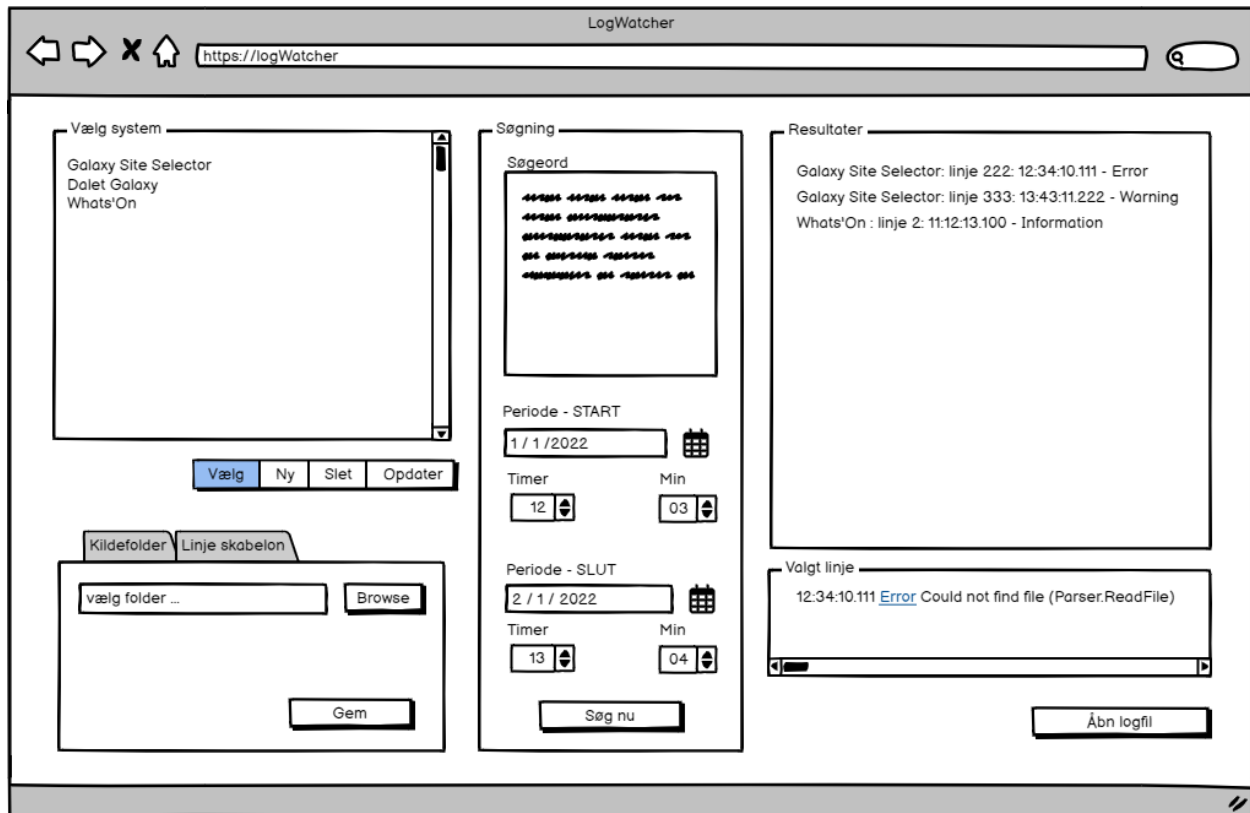
4.1.3.7 View returned hits from search

Efter gennemført søgning, vises listen over returnerede loglinjer



4.1.4 Grafisk Interface – GUI (trin 4)

Simpel skitse for interaktion med bruger.



(<https://balsamiq.cloud/skqgr93/p6284we/r6B57>)

4.1.5 Use Cases (trin 5)

Ved at udvælge og beskrive brugerhandlinger kan vi sætte domænemodellen i spil.

4.1.5.1 Create New Source System

Brugeren ser en liste over allerede eksisterende **SourceSystem**'s som er oprettet i **SourceSystemCollection**. Hvis listen er tom, vises intet. Brugeren klikker på knappen "Ny", og der vises felter for indtastning af **SourceFolder**, og definitionen af målsystemets **LineTemplate**.

Hvis de indtastede data passerer validering, aktiveres "Gem"-knappen. Hvis brugeren trykker på "Gem", bliver det ny system tilføjet **SourceSystemCollection**.

4.1.5.2 Edit an existing Source System

Brugeren ser en liste over allerede eksisterende **SourceSystem**'s som er oprettet i **SourceSystemCollection**. Brugeren vælger ved at klikke i listen.

Objekterne **SourceFolder** og **LineTemplate** for den valgte liste vises i et separat vindue. De kan nu redigeres. Når brugeren trykker på "Gem" bliver informationerne valideret, og data skrives tilbage i objekterne. Derefter opdateres **SourceSystemCollection** med det rettede **SourceSystem**.

4.1.5.3 *Select an existing SourceSystem*

Brugeren ser en liste af **SourceSystems** indeholdt i **SourceSystemCollection**. Brugeren vælger ved at klikke i listen. Listen giver mulighed for at vælge ét eller flere systemer (multiselect).

4.1.5.4 *Update logs for a selected SourceSystem*

Brugeren ser en liste af **SourceSystemer** indeholdt i **SourceSystemCollection**. Brugeren vælger ved at klikke i listen. Hvis kun ét system er valgt, bliver knappen "Opdater" aktiv.

SourceFolder for det valgte system bliver scannet, og evt. nye filer bliver indlæst i databasen.

4.1.5.5 *Define SearchSet for a search*

Hvis brugeren har klikket på mindst ét **SourceSystem**, bliver søge-vinduet aktivt.

Brugeren kan skrive et antal ord som bliver tilføjet **KeywordList**. Der kan også anføres et tidspunkt for tidligste forekomst, og seneste forekomst. Tidspunkter anføres i datoformat, og med mulighed for at angive time/min. Tidspunkter tilføjes til **SearchPeriod**. **KeywordList** og **SearchPeriod** føjes til et **SearchSet**-objekt.

4.1.5.6 *Execute search*

Hvis søgevinduet er aktivt, kan brugeren klikke på knappen "Søg nu". De betingelser som er defineret i **SearchSet**-objektet bliver anvendt til at begrænse søgeresultatet.

4.1.5.7 *View returned hits from search*

Hvis en søgning er gennemført, og hvis **HitCollection** indeholder elementer, bliver hvert **LogHit** i **HitCollection** vist i "Resultater"-vinduet. Et **Loghit** vises i reduceret form.

Brugeren kan klikke på et **LogHit** i vinduet, og så vises hele linjen i vinduet "Valgt linje".

Hvis brugeren har valgt en linje, bliver knappen "Åbn logfil" aktiv. Hvis brugeren klikker, bliver hele kilde-log-filen åbnet i et nyt vindue.

4.1.6 Robustness Analysis (trin 6)

I en Robustness Analysis tilstræber vi at reducere tvetydighed i domænemodellen og Use Cases. Det primære værktøj til formålet er opbygning af et **Robustness Diagram** (RD).

I et RB indgår en række symboler som repræsenterer specifikke aspekter af det modellerede system: Boundary, Entity og Controller.



4.1.6.1 *Definition: Boundary Object*

Denne objekttype repræsenterer grænseflader. Dvs. entiter som optræder i f.eks. GUI, og udstiller et aspekt af systemets tilstand, og/eller tillader brugeren at interagere med aspektet.

4.1.6.2 *Definition: Entity Object*

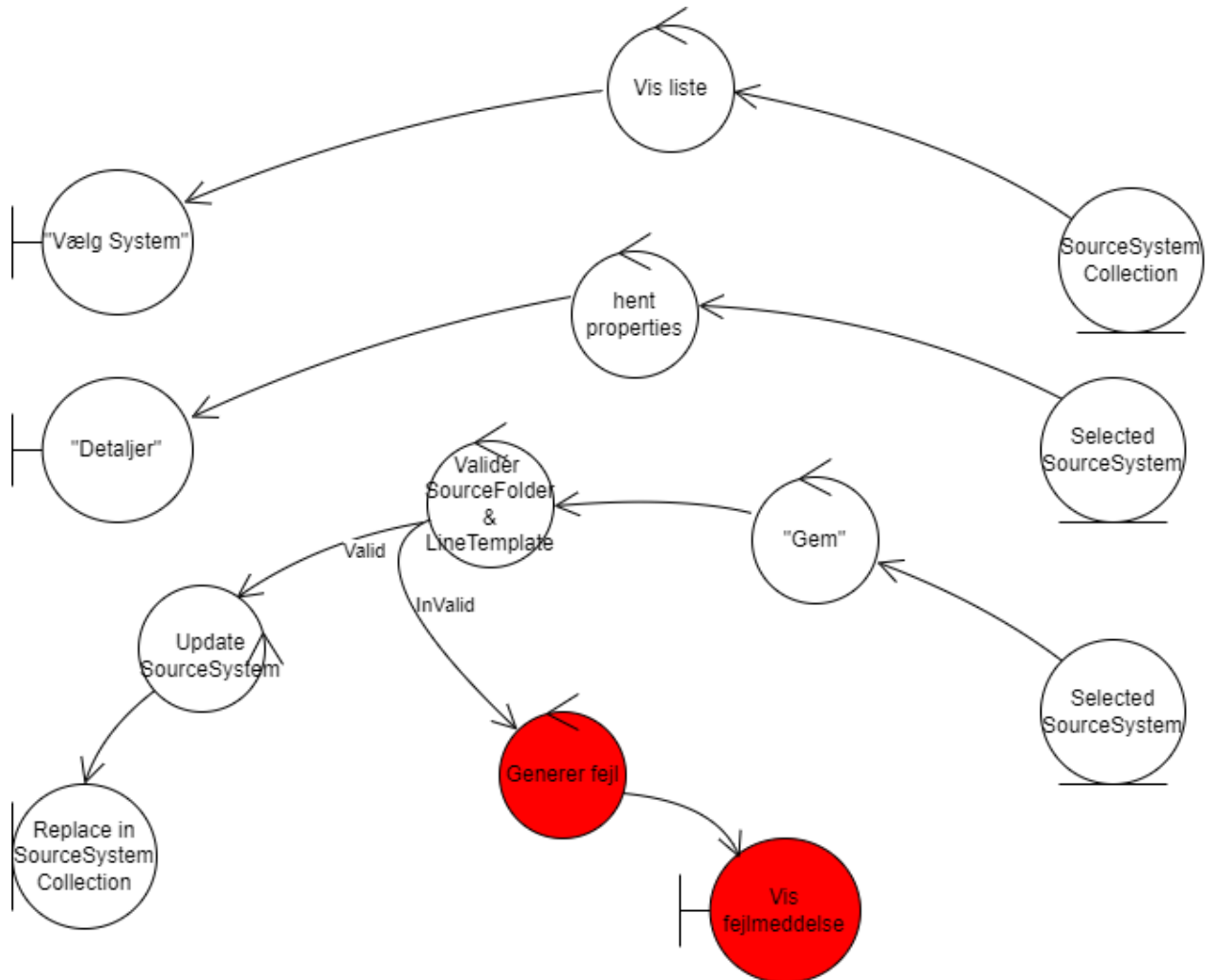
En repræsentation af et domænemodel-objekt, dvs. et objekt som repræsenterer systemets tilstand.

4.1.6.3 Definition: Controller

Repræsenterer typisk en process, dvs. en handling som udføres eller sæt data som transformeres.

4.1.6.4 Robustness Diagram

Jeg har valgt at afgrænse analysen til kun at omfatte casen "Edit an existing sourcesystem".



Under analysen viser RD behovet for at håndtere fejl-scenariet hvor de nye data for **SourceFolder** og **LineTemplate** ikke kan valideres. Ved fejl bliver der genereret en bruger-rettet fejlmeddelelse.

4.2 System analyse - konklusion

Med analysen i hånden er processen med at kode front-end til LogSearcher blevet reduceret til hvordan selve implementeringen skal realiseres. De svære beslutninger om struktur og interaktivitet er taget, og integrationen mod applikationens øvrige komponenter er også blevet klarere.

5 METODE: UDVIKLING

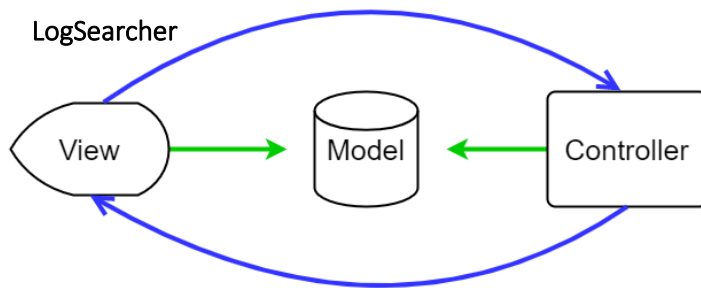
5.1 MVC @ LogSearcher

LogSearcher udgøres af 3 distinkte komponenter eller lag. En brugervendt komponent, et persisterings-lag, og en søge-komponent. Denne struktur er inspireret af MVC-modellen, Model-View-Controller.

View: det brugervendte lag, GUI, som udstiller applikationens funktionalitet.

Controller: det lag som håndterer selve databearbejdningen.

Model: det lag som repræsenterer data og deres struktur.

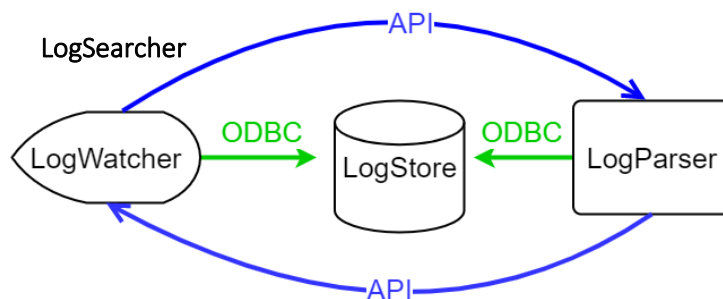


View og controller-lagene bruger den fælles database til at persistere og hente data.

Når data er indlæst, starter brugeren søgningen. Derved sendes en kommando til controlleren, som henter det relevante søgesæt, og behandler det. Når søgningen er klar, sendes

søgeresultatet til brugeren.

Det giver mening at navngive komponenterne ifht. deres rolle i denne illustration.



View: **LogWatcher**

Model: **LogStore**

Controller: **LogParser**

Både LogWatcher og LogParser anvender ODBC-protokollen til at skrive og hente data hos LogStore.

Interaktion mellem LogWatcher og LogParser faciliteres af et API.

ODBC er en udbredt standard for tilgang af RDBMS-systemer (ODBC, 2022). API (Application Programming Interface) refererer til en http-baseret udveksling af data via REST API (REST, 2022).

LogWatcher er i denne sammenhæng forside af LogSearcher-applikationen, som udgøres af 3 sammenkædede komponenter. Denne struktur giver flere fordele, bl.a. stor kompatibilitet med den foreslåede drifts-model. Hver enkelt komponent kan afvikles i et diskret miljø, som kan være fysisk afkoblet fra de øvrige.

5.2 Søgning i store datasæt

LogSearchers relevans er baseret på at kunne finde specifikke ord i store datasæt, og knytte hvert sæt af forekomster til en kildefil/kildesystem.

Den naive tilgang til denne søgning er at gennemgå hvert linje ord-for-ord og danne en forekomst-liste. Hver gang der søges, skal alle linjer chekkes for forekomst. Hvis ét check tager x millisekunder, vil y antal

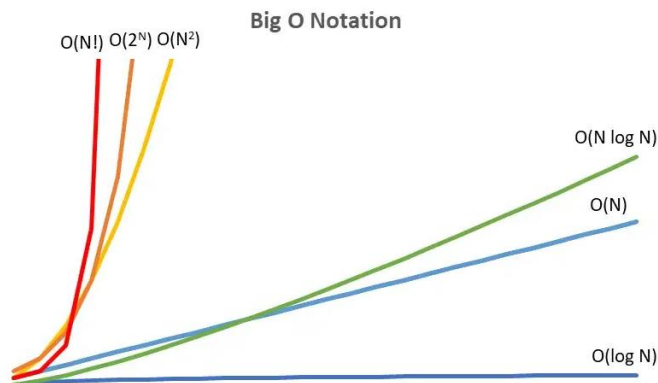
linjer tage $x \cdot y$ millisekunder. Altså en lineær relation mellem søgesættets størrelse og varigheden af søgning.

Til at udtrykke denne type relationer anvendes normalt BigO notation.

5.2.1 BigO notation & Time Complexity

"Big O" notation er en måde at måle en algoritmes effektivitet. Det udtrykker køretid i forhold til inputmængde. Altså hvordan funktionen skaleres.

Der er 2 primære aspekter af notationen: Time Complexity og Space Complexity (hhv. køretid og hukommelsesforbrug). De angiver den øvre grænse for vækstraten af funktionen.



Big O introduction

BigO er altså en måde at tale om en algoritmes effektivitet, da det giver en målestok for køretid/hukommelsesforbrug i forhold til det datasæt som anvendes til input.

Som regel er vi mest interesseret i Time Complexity. (Big O introduction, 2021)

BigO versionen af eksemplet fra 5.2 er: $O(n)$, hvor n er antallet af linjer.

Her illustreres gængse forekomster af relationen mellem input-mængde og processtid. X-aksen repræsenterer input, Y-aksen tidsforbrug.

5.2.2 Overblik over nogle indekseringsstrategier

Der findes et antal strategier til at håndtere sortering og søgning af datasæt. Nogen af de mest brugte er de følgende.

5.2.2.1 Binary search på en sorted list

Udgangspunktet er en liste af værdier, **sorteret i størrelsesorden** (enten størst-først eller størst-sidst). Ved at starte i midten af listen, kan vi konstatere om søge-værdien er i øvre eller nedre halvdel. Derefter reduceres søgesættet til denne halvdel. Processen gentages indtil værdien enten er fundet, eller ikke eksisterer i sættet.

TC, Sortering af liste: $O(n \log n)$, søgning: $O(\log n)$

5.2.2.2 Hashtable

Hvert ord i søgesættet omsættes til en talværdi(hash) og placeres i en tabel af typen linked list. En linked list har for hvert element i listen en pointer til det næste element. Ved søgning omsættes søgenøglen til en hash efter samme metode. Tabellen traverseres indtil den søgte hash er fundet. Afhængig af hashing-metoden kan der forekomme kollisioner, hvor 2 input tildeles samme hash.

TC, Insert/retrieve: $O(1)$

5.2.2.3 Trie (prefix tree)

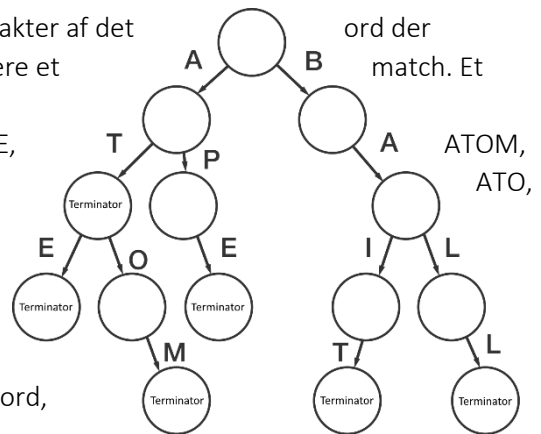
Ordet trie kommer af **retrie**ve, at hente. Det er en træ-baseret datastruktur, som er velegnet til at organisere store data-mængder, typisk tekst-streng. Et trie har en tom rod-node (root) og referencer til til under-noder (child-nodes).

Hver enkelt forbindelse mellem noder repræsenterer en karakter af det indlæses. Ved at traversere noderne i træet, kan vi identificere et komplet ord indikeres ved at sidste node markeres som

terminator/slut-node. I figuren ser vi at ordene {AT, ATE, APE, BAIT, BALL} er markeret som afsluttede ord. Hvorimod {AP, BA, BAI, BAL} ikke er.

Tries anvendes ofte som erstatning for hashtables, bl.a. fordi de undgår problemet med kollisioner, og er nemme at ændre størrelse på.

TC, Indsæt: $O(mn)$, søgning: $O(an)$ (m = karakterer i længste ord, n=antal ord, a=længden af længste ord)



5.2.3 Konklusion: Indekseringsstrategi

Jeg har valgt at implementere tries til søgemotoren, LogParser, primært for at undgå hashtable-kollisioner og problemer med at ændre størrelsen. Desuden er søgetiden i et indlæst trie svarende til en hashtable.

Terminatoren skal rumme en reference til ordets kilde, dvs. kildesystem:logfil:loglinje. Derved kan vi identificere kilden for hvert ord-hit.

5.3 Udvikling af søgemotor

LogParser er udviklet i Python, fordi det er velegnet til den type datahåndtering som Tries og Binary Search Tree kræver.

5.3.1 LogParser komponenter

LogParser-komponenterne bliver instantieret når LogWatcher sender en query.

Her følger et overblik over klasserne og deres metoder.

5.3.1.1 Shell, QueryParser, Query, PrepareTrie, Loader (workers)

De 5 primære workere arbejder i en kæde hvor 1. led starter og konsumerer de følgende.

Shell modtager søgekommandoer, og delegerer konstruktionen af den aktuelle query som brugeren sender. **QueryParser** modtager query, og aktiverer de relevante metoder. **Query** er wrapper for selve søgefunktionaliteten. **PrepareTrie** kalder **Loader** for at populere søge-træerne.

5.3.1.2 LogTrie, SearchTrie, BST

LogParser anvender 2 forskellige trie-træer og et Binary Search Tree.

LogTrie indeholder for hver forekomst af et ord, pointere til kilde-loglinjen(1 eller flere).

SearchTrie holder styr på antallet af hits når der søges efter flere ord.

LogTrie bruger **terminatorer** til at markere ordgrænser. En terminator definerer en reference til kildesystem:logfil:loglinje. For hvert hit på samme ord, tilføjes et terminator-element.

Der bruges også et binært søgetræ (**BST**), som kun anvendes til at sortere hits efter tid. BST traverseres InOrder for at udtrække det sorterede sæt.

5.3.1.3 *SearchSet, SourceSystem, SearchPeriod, LogLine (models)*

De databærende klasser. LogLine rummer metoder til at udtrække tidspunkt og tekst-elementer fra en loglinje. SearchSet, SourceSystem og SearchPeriod repræsenterer samme objekttype som i LogWatcher.

5.3.1.4 *Database, Api (wrappers)*

Database udstiller ODBC-værktøjer til database-interaktion med LogStore. Api tilføjer en http-grænseflade til LogParser.

5.3.2 Søgning med LogParser

LogParser bliver udstillet til LogWatcher via et minimalt API. Dette API er grundlæggende en wrapper for indgangspunktet for LogParser, Shell.

API'et udstiller 2 endpoints: /search og /research.

/search anvendes hvis søgningen er ny, dvs. hvis gruppen af loglinjer er ny.

/research anvendes hvis det er samme sæt loglinjer, men nye søgeord.

LogWatcher sender LogParser et json-dokument som en serialiseret version af LogWatchers SearchSet-objekt. På basis af søge-argumentet, startes en forespørgsel mod LogStore. De modtagne loglinjer loades i LogTrie. Derefter forespørges LogTrie på hits for søgeordene, og de tilsvarende terminatorer parses for at identificere loglinjerne.

Når søgningen er afsluttet, dannes et svar-datasæt: et array af (sourcesystemid, logfileid, loglineid).

Svaret udstilles via API, som LogWatcher query'er og derefter danner en liste over hits. LogStore forespørges på elementernes ID, og LogWatcher danner hitlisten og udstiller dem for brugeren via GUI.

5.3.2.1 *LogParser eDSL*

QueryParser giver mulighed for at anvende en liste af metoder og argumenter til at konstruere den endelige forespørgsel. QueryParsers metoder kaldes via refleksion. (Reflective Programming, 2022)

Shell bygger en sekvens af query-elementer som sammenstilles før de endeligt gives som argument til QueryParser, som derefter udfører søgningen.

Denne struktur anvendes f.eks. til at differentiere søgninger fra Api, mellem at anvende hhv. nyt søgegrundlag og et eksisterende.

@ API, ny søgning (/search)

```
def ReSearch(self, searchSet: SearchSet):
    if (self.shell == None):
        raise HTTPException(status_code=400,
detail="Previous search not valid")
    self.results = None
    self.shell.do_find(searchSet.KeywordList)
    self.shell.do_sort('True')
    self.results = self.shell.do_run()
```

```
def Search(self, searchSet: SearchSet):
    self.results = None
    self.shell = Shell(searchSet)
    self.shell.do_find(searchSet.KeywordList)
    self.shell.do_sort('True')
    self.results = self.shell.do_run()
```

@ Api, eksisterede (/research)

6 AFRUNDING

I afsnit 1-5 har jeg defineret problem-domænet, og beskrevet rammerne for en mulig løsning ift. problemformulering. I denne proces har jeg tegnet både et konceptuelt og et praktisk billede af LogSearcher som en løsning.

Den nærværende opgave beskriver den fuldt udfoldede implementering, samt overvejelser derom og visse konkrete, praktisk realiserede aspekter. Det medfølgende produkt repræsenterer derimod en tilnærmelse til den fuldt implementerede version.

6.1 En prototype - afgrænsning af opgavens omfang

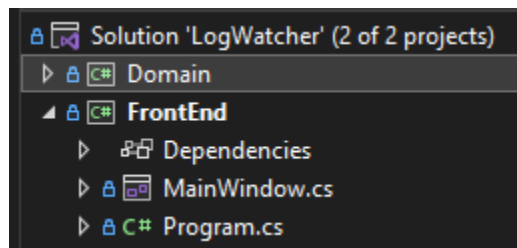
Til min opgavebesvarelse har jeg udviklet en prototype af den fulde funktionalitet og implementeringsmodel som indtil nu er beskrevet.

Den foreslåede struktur og jobfordeling mellem komponenterne er bevaret i denne prototype.

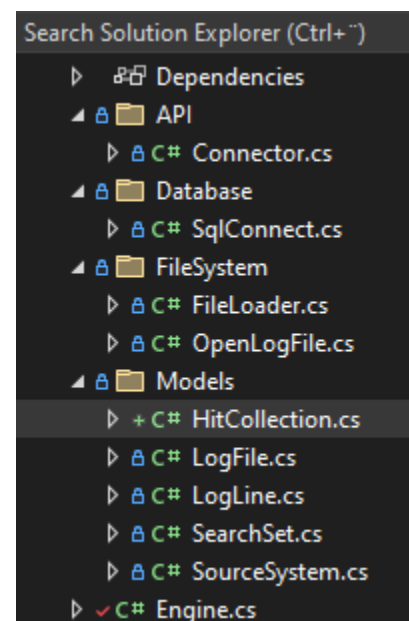
6.1.1 LogWatcher i WinForm

LogWatcher prototypen er realiseret i Windows Forms .NET 6.

Funktionaliteten er distribueret mellem projektet **FrontEnd** og library'et **Domain**. FrontEnd varetager GUI og brugerinteraktion, og Domain indkapsler bl.a. database interaktion og et minimalt http-interface.



Domæne-typerne som blev beskrevet i afsnit 4, er implementeret i Domain/**Models**.



Klassen **API.Connector** håndterer det minimale http-interface.

Database.**SqlConnect** udgør database-grænsefladen, og **FileSystem.FileLoader** udstiller metoder til at indlæse logfiler.

Klassen **Engine** orkestrerer og udstiller funktionalitet mod FrontEnd.

Koden i FrontEnd består primært af event-håndtering (når brugeren klikker på knapper osv).

Jeg valgte at udføre protypen i WinForms, fordi formatet er meget velegnet til hurtigt at bygge et grafisk interface og understøtte bruger-interaktion. Det giver mig mulighed for at fokusere på at udvikle modellen for hvordan komponenterne skal fungere sammen, uden at skulle opbygge et fuldt website.

6.1.1.1 Kørsel af komponent

Til brug for prototypen startes blot build-outputtet (.exe) af FrontEnd i et windows10 miljø.

6.1.2 LogStore i MSSQL

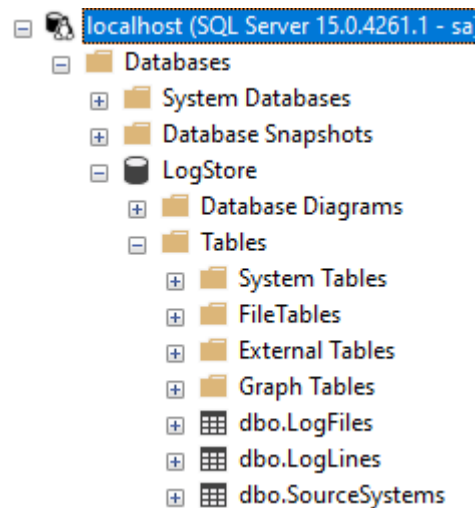
LogStore er realiseret i en MSSQL-instans.

I realiteten er databasen umiddelbart klar til produktionsbrug. De tre tabeller som systemet er afhængigt af, kræver meget lidt tilpasning før udrulning.

6.1.2.1 Kørsel af komponent

Jeg har til prototypen valgt at afvikle en MSSQL-instans i en Docker container i mit udviklings-miljø.

Instansen afvikles via Docker Desktop.



6.1.3 LogParser

Komponenten er ~95% realiseret i sin endelige form.

LogParser udstiller mere funktionalitet end der bringes i spil i denne prototype. Bl.a. giver den mulighed for at finde start/slut perioder for forekomst af sæt af søgeord, f.eks. for at finde start/stop af længere-kørende processer.

I den endelige implementering skal LogParser kobles sammen med en webserver, for at kunne udstille sit API mod LogParser. Denne feature er ikke udviklet til prototypen.

6.1.3.1 Kørsel af komponent

Men for at kunne simulere funktionaliteten og interaktionsmodellen for komponenterne af LogSearcher, har jeg valgt at bruge uvicorn (Uvicorn, 2022) som stand-in webserver.

6.1.4 LogSearcher prototype @ Github

Denne opgavebesvarelse er tilgængelig på https://github.com/FaderVader/CPH_Afgangsprojekt_2022

Selve koden til hhv. LogWatcher og LogParser findes under /src

6.2 Afsluttende konklusion

Med nærværende opgave, samt det tilhørende stykke praktisk implementering, har jeg demonstreret hvordan en løsning for applikationen LogSearcher kan udformes. Jeg har vist hvordan applikationens elementer kan bygges, og hvordan de bringes i samspil.

Prototypen realiserer 100% af den funktionalitet som jeg foreslog i problemformuleringen.

Selve driftsmodellen for prototypen afviger for den foreslåede model, hvor LogWatchers server-komponent og LogParser afvikles som docker-instanser, og LogStore på en VM.

Til prototypen har jeg valgt at hoste databasen i en Docker-container, bl.a. fordi det gør det meget nemt og hurtigt at nulstille hele databasen når der under udviklingen opstår fejl 😊 Det har også givet mig erfaring med og forståelse for drift af Docker-containere, som er nyttig hvis løsningen skal skaleres op til den foreslåede implementering i produktion.

6.2.1 Produktions version – mod fuld funktionalitet

Prototypen anvender en meget rudimentær proces til pre-processering og udskilning af loglinje bestanddele. Når LogSearcher skal behandle data fra forskellige systemer, skal denne proces udvides.

Det er nærliggende at udvide LogWatcher til at udstille LogParsers mulighed for at lede efter sæt af start/slut søgeord, og deres tidslige distribution, f.eks Top5/Bottom5.

Muligheden for at vise et grafisk overblik over den tidslige distribution af søge-hits er også en oplagt feature.

Hvis produktions-løsningen skal understøtte et flerbruger-scenarie, skal LogWatcher som Blazor Server udbygges med navngivne bruger-sessioner. LogParser skal instantieres individuelt for hver bruger-session, hvor hver session har eget diskret søgesæt.

Det skridt vil også introducere behovet for at håndtere sikkerhedsaspekter, f. eks at integrere og håndtere brugeradgang mod DRs Active Directory.

Jakob Viggo Hansen
Nørrebro, 2022

6.3 Referencer

Embedded Domain Specific Languages. (7. 11 2022). Hentet fra Wikipedia: https://en.wikipedia.org/wiki/Domain-specific_language

.NET. (7. 11 2022). Hentet fra Microsoft: <https://dotnet.microsoft.com/en-us/learn/dotnet/what-is-dotnet>

Automatic Memory Management. (7. 11 2022). Hentet fra Microsoft: <https://learn.microsoft.com/en-us/dotnet/standard/automatic-memory-management>

Big O introduction. (22. 3 2021). Hentet fra Towards Datascience: <https://towardsdatascience.com/introduction-to-big-o-notation-820d2e25d3fd>

C# vs Python 2. (7. 11 2022). Hentet fra litsLink: <https://litslink.com/blog/csharp-vs-python-choosing-right-language-for-your-project>

C# vs Python. (7. 11 2022). Hentet fra Hackr.io: <https://hackr.io/blog/c-sharp-vs-python>

CI. (7. 11 2022). Hentet fra Wikipedia: https://en.wikipedia.org/wiki/Continuous_integration

Colossus. (7. 11 2022). Hentet fra Britannica: <https://www.britannica.com/technology/Colossus-computer>

Comparative Analysis of Horizontal And Vertical Scaling. (6 2016). Hentet fra Academia.edi: <https://d1wqtxts1xzle7.cloudfront.net/52711973/IJSARTV2I63514-with-cover-page-v2.pdf?Expires=1668261640&Signature=fRPR9P9Aa47XxQD0tVfIF9eApzeR4eSNZRhtU9u8j2hKPgllvpEHq4HvWC61uoHVM7~FoogigIDN5vnFucZOz3FCFnHEDDoXhn2AgdyiRkTOZA8LTrC-QYMot9ZgVRdgoztBy1yNmP>

Containers Versus Virtual Machine. (2. 9 2018). Hentet fra Springer Link: https://link.springer.com/chapter/10.1007/978-981-13-1501-5_12#citeas

Continuous Integration. (7. 11 2022). Hentet fra Atlassian.com: <https://www.atlassian.com/continuous-delivery/continuous-integration>

Doug Rosenberg, M. S. (2007). *Use Case Drive Object modelling with UML*. Apress.

DSL Idris. (7. 11 2022). Hentet fra Type-driven.org.uk: <https://www.type-driven.org.uk/edwinb/papers/dsl-idris.pdf>

eDSL. (7. 11 2022). Hentet fra Haskell.org: https://wiki.haskell.org/Embedded_domain_specific_language

Embedded Languages. (7. 11 2022). Hentet fra cse.chalmers.se: <http://www.cse.chalmers.se/edu/year/2016/course/TDA342/secret42/lecture2.html>

AU i IT Afgangprojekt – Udvikling af LogSearcher

- ENIAC*. (7. 11 2022). Hentet fra Wikipedia: <https://en.wikipedia.org/wiki/ENIAC#Programmers>
- Fowler, M. (2003). *UML Distilled: A Brief guide to Standard Object Modelling Language*. Addison-Wesley Professional.
- HAL*. (7. 11 2022). Hentet fra Wikipedia: https://en.wikipedia.org/wiki/Hardware_abstraction
- ISO 8601*. (8. 11 2022). Hentet fra Wikipedia: https://en.wikipedia.org/wiki/ISO_8601
- Java*. (7. 11 2022). Hentet fra Wikipedia: [https://en.wikipedia.org/wiki/Java_\(programming_language\)](https://en.wikipedia.org/wiki/Java_(programming_language))
- Kubernetes: Up and Running*. (2022). Hentet fra Google Books: https://books.google.dk/books?hl=en&lr=&id=KeB-EAAAQBAJ&oi=fnd&pg=PT17&dq=kubernetes&ots=V9SXEoolU6&sig=jBTY4WiUsz-tDeVhPTa19Pp582o#v=onepage&q=kubernetes&f=false&redir_esc=y#v=onepage&q=kubernetes&f=false
- Languages*. (8. 11 2022). Hentet fra Memory Management: <https://www.memorymanagement.org/mmref/lang.html>
- Microsoft*. (2022). Hentet fra Azure Stack HCI: <https://azure.microsoft.com/en-us/products/azure-stack/hci/>
- Multithreading in .NET*. (17. 7 2017). Hentet fra jcomputers: <http://www.jcomputers.us/vol13/jcp1304-07.pdf>
- ODBC*. (03. 12 2022). Hentet fra Wikipedia: https://en.wikipedia.org/wiki/Open_Database_Connectivity
- OOP*. (12. 11 2022). Hentet fra Wikipedia: https://en.wikipedia.org/wiki/Object-oriented_programming
- Public Broadcasting*. (7. 11 2022). Hentet fra Wikipedia: https://en.wikipedia.org/wiki/Public_broadcasting
- Python code parallelization*. (2016). Hentet fra Eso.org: <http://www.eso.org/~jagonzal/ADASS-2016/P6-11.bkp/P6-11.pdf>
- Python for Science*. (7. 11 2022). Hentet fra IBM: <https://developer.ibm.com/blogs/use-python-for-scientific-research/>
- Reflective Programming*. (4. 12 2022). Hentet fra Wikipedia: https://en.wikipedia.org/wiki/Reflective_programming
- REST*. (3. 12 2022). Hentet fra Wikipedia: https://en.wikipedia.org/wiki/Representational_state_transfer
- SQL vs. NoSQL vs. NewSQL*. (10 2016). Hentet fra COE: <https://caeaccess.com/archives/volume6/number1/binani-2016-cae-652418.pdf>
- Stateless Vs. Stateful*. (2019). Hentet fra Forbes: <https://www.forbes.com/sites/forbestechcouncil/2019/09/18/stateless-vs-stateful-the-devils-in-the-details/>
- Survey on Horizontal and Vertical Scaling*. (6 2016). Hentet fra Academia.edu: <https://d1wqtxts1xzle7.cloudfront.net/52711973/IUSARTV2I63514-with-cover-page-v2.pdf?Expires=1668261499&Signature=A5-xcyvRc2PkRD7UbiO9ZdGyQeSr9xC2NxKy77eV3f3zXprb26J1u9J9l6w4F94SouZm9nBNVOW8s3ubmiEnPM2rS3W~qM10bWe5rAbQ6SPbFLhLwkCUhQL85ZLUZGps3jU1XUFqW9sEO>
- Top Programming Languages*. (7. 11 2022). Hentet fra IEEE: <https://spectrum.ieee.org/top-programming-languages-2022>
- TowardsDataScience*. (9. 7 2020). Hentet fra Python multi-processing: <https://towardsdatascience.com/python-multi-threading-vs-multi-processing-1e2561eb8a24>
- Unix-like*. (8. 11 2022). Hentet fra Wikipedia: <https://en.wikipedia.org/wiki/Unix-like>
- Uvicorn*. (4. 12 2022). Hentet fra Uvicorn: <https://www.uvicorn.org/deployment/>
- What is Blazor*. (7. 11 2022). Hentet fra Blazor University, Microsoft: <https://blazor-university.com/overview/what-is-blazor/>
- WORA*. (7. 11 2022). Hentet fra Wikipedia: https://en.wikipedia.org/wiki/Write_once,_run_anywhere