



UNIVERSITATEA DIN
BUCUREȘTI

FACULTATEA DE
MATEMATICĂ ȘI
INFORMATICĂ



SPECIALIZAREA TEHNOLOGIA INFORMAȚIEI

Lucrare de licență

MICROSERVICII ÎN DEVOPS

Absolvent

Bahrim Dragoș

Coordonator științific

Conferențiar Doctor Kevorchian Cristian

București, iulie 2023

Rezumat

Microserviciile sunt o arhitectură de sisteme distribuite ce a luat amploare în urma structurării organizațiilor către piață. Aceasta a fost facilitată de schimbări în modul în care lansăm produsele prin dezvoltarea platformelor cloud și a ce a dus la îmbunătățiri la timpul de livrare dar și a consistenței. Obiectivul este prezentarea acestor concepte și implementarea acestora prezentând modul de gândire ce influențează ciclul produsului. În acest scop, mă documentez legat de domeniu și încerc să îmi fac o idee de ansamblu asupra aspectelor ce trebuie luate în construirea unui sistem iar apoi o să aplic aceste informații. Aceasta lucrare ar trebui să servească ca un prim contact cu microserviciile dar și a practicilor DevOps.

Abstract

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Fusce vitae eros sit amet sem ornare varius. Duis eget felis eget risus posuere luctus. Integer odio metus, eleifend at nunc vitae, rutrum fermentum leo. Quisque rutrum vitae risus nec porta. Nunc eu orci euismod, ornare risus at, accumsan augue. Ut tincidunt pharetra convallis. Maecenas ut pretium ex. Morbi tellus dui, viverra quis augue at, tincidunt hendrerit orci. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aliquam quis sollicitudin nunc. Sed sollicitudin purus dapibus mi fringilla, nec tincidunt nunc eleifend. Nam ut molestie erat. Integer eros dolor, viverra quis massa at, auctor.

Cuprins

1	Introducere	5
2	Preliminarii	8
3	Microservicii	9
3.1	Concepte generale	9
3.2	Design	12
3.2.1	Interconectare	12
3.2.2	Domain Driven Design	13
3.2.3	Concepte de programare orientată obiect în microservicii	14
3.2.4	Migrarea	15
3.3	Tehnici de implementare	17
3.3.1	Tipuri de comunicare	18
3.3.2	Tehnologii de comunicare	20
3.3.3	Schimbări în funcționalitate	22
3.3.4	Partajare între microservicii	23
3.4	Build	28
3.5	Testare	31
3.6	Deployment	34
3.7	Monitorizare	38
3.8	Securizare	41
3.9	Evoluție	46
3.9.1	Stabilitate	46
3.9.2	Scalare	49
3.9.3	Caching	50
3.10	Organizare	52
3.10.1	Consum	52
3.10.2	Structură	53
4	Orchestrare și infrastructură	56
4.1	Concepte generale	56

4.2	Tipuri de deployment	57
4.2.1	Mașină fizică	57
4.2.2	Mașină virtuală	57
4.2.3	Container	58
4.2.4	Aplicație în container	58
4.2.5	Platform as a Service	59
4.2.6	Function as a Service	59
4.3	Infrastructure as a Service	60
4.3.1	General	60
4.3.2	Extinderea limitelor	60
4.3.3	Microsoft Azure	62
4.4	Docker	68
4.4.1	General	68
4.4.2	Componente	69
4.4.3	Utilizare	70
4.4.4	Simpla orchestrare	71
4.5	Kubernetes	72
4.5.1	General	72
4.5.2	Obiecte de bază	73
5	Îmbunătățirea metodelor de dezvoltare software	79
5.1	Schimbari în livrarea produselor	79
5.2	DevOps și aspecte ale comunității	79
5.3	Ce înseamnă să aplici DevOps	79
5.4	Unelte	79
6	Aplicație	80
7	Concluzii	81
	Bibliografie	82

Capitolul 1

Introducere

Domeniul informaticii, calculatoarelor și al tehnologiei informației este unul care va fi mereu în continuă evoluție. Dorința companiilor de a-și servi clienții cât mai rapid a dus la un avans tehnologic în care mereu apare ceva nou ce are ca scop îmbunătățirea proceselor actuale. Această arie și-a început dezvoltarea aproape acum o sută de ani iar avansul poate fi apreciat numai gândindu-ne la modul în care scriam cod și livram produse acum, acum zece ani și acum două zeci de ani. Se observă o diferență și la accesibilitatea unităților de calcul, actual majoritatea persoanelor au acces la un dispozitiv ce se poate conecta la Internet, ceea ce le permite să devină consumatori la diferite servicii.

Piața pentru unități de calcul a evoluat. În 1943, Thomas Watson, director la IBM menționează că „I think there is a world market for maybe five computers” (Cred că există o piață globală pentru poate cinci computere). În 1999, Michael Barr afirmă în „Programming Embedded Systems in C and C++” „One of the more surprising developments of the last few decades has been the ascendance of computers to a position of prevalence in human affairs. Today there are more computers in our homes and offices than there are people who live and work in them. Yet many of these computers are not recognized as such by their users.” (Una dintre cele mai surprinzătoare dezvoltări a ultimelor decenii a fost ascensiunea computerelor într-o poziție predominantă în afacerile oamenilor. Astăzi sunt mai multe computere în casele noastre și în birouri decât persoane care trăiesc și lucrează în ele. Însă multe dintre acestea nu sunt recunoscute de către utilizatorii lor). Cele două fraze evidențiază modul cum s-a schimbat prezența computerelor în viața noastră. La început acestea erau foarte costisitoare, greu de administrat și de operat însă acestea au devenit din ce în ce mai mici iar recent aproape orice are nevoie de semiconductori datorită integrării ce le fac „smart”.

Aceste afirmații se aplică și pentru produsele software, ce au ca menire să ofere asistență altor aplicații sau să fie consumate direct. Indiferent de locul în care mă duc, probabil persoana cu care aș interacționa folosește un computer, fie că îmi cumpăr ceva de la un magazin făcând o plată cu cardul sau că mi se livrează un colet iar curierul marchează pe AWB că a fost livrat. Christopher Little afirmă că „Every company is a technology com-

pany, regardless of what business they think they are in. A bank is just an IT company with a banking license.” (Orice companie este o companie ce se axează pe tehnologie, indiferent de mediul de afaceri în care se află. O bancă este doar o companie IT cu licență de a funcționa ca o bancă.)

Desigur nu toate companiile sunt la fel și nu toate concurează cu toate companiile, însă chiar dacă nu ești o companie ce activează în tehnologie, probabil folosești componente tehnologice pentru a-ți îmbunătăți randamentul sau să oferi clienților un avantaj față de competiție, care probabil gândește în același mod.

Însă pentru fiecare companie obiectivul este identic, să livreze clienților produse de calitate și cât mai rapid. Pentru aceasta trebuie să ne asigurăm că avem inginerii necesari pentru implementarea cererilor noi, însă momentul în care aceasta este terminată este doar începutul procesului de integrare și ulterior lansare, și uneori acesta este cel care reprezintă un blocaj din mai multe puncte de vedere ce limitează numărul de câte ori putem aduce o îmbunătățire produsului nostru.

În încercarea de a livra mai rapid sau pentru a ușura modul în care manevrăm un influx sau o lipsă de trafic, s-a încercat folosirea unei arhitecturi paralele asupra aplicațiilor creând arhitecturi orientate pe servicii, iar în momentul în care aceste servicii sunt concentrate pe un număr redus de funcționalități și pot fi lansate independent și să își păstreze funcționalitate putem vorbi de o arhitectură bazată pe microservicii.

Însă schimbând arhitectura sau practiciile din modul de desfășurare al livrării produsului nu este suficient pentru a asigura performanța, întrucât ambele dintre ele aduc dezavantaje ce trebuie tratate iar uneori acestea sunt mai mari decât avantajele pe care schimbarea le-ar aduce, însă implementarea corectă în locurile ce ar beneficia de o astfel de abordare poate să aducă performanțe ce nu ar fi posibile cu procesele vechi.

Afinitatea mea pentru microservicii provine de la lipsa de cunoștințelor pentru a putea alege dintre mai multe limbaje de programare, framework-uri sau platforme. De exemplu, ce avantaje mi-ar face să aleg ca pentru server-ul meu să folosesc o aplicație în Node.JS sau una în Go? Chiar dacă putem să cunoaștem niște lucruri informative la început, cei mai buni indicatori sunt monitorizările proprii în producție. De asemenea pot fi cazuri în care îmbunătățirile aduse unei platforme pe parcurs o fac mai bună față de ce era inițial (dezvoltarea TypeScript poate să influențeze alegerea inițială a unei platforme SpringBoot ce este type-safe). Astfel, pe o arhitectură bazată pe microservicii putem să folosim limbajul potrivit pentru serviciul căruia îi aduce cele mai multe beneficii. Acest lucru se extinde și pentru baze de date. Poate în timpul evoluției serviciului apar tehnologii noi, însă având un singur lucru central ar face migrarea mult mai grea, de asemenea uneori ar trebui să facem compromisuri pe care alegerea unui alt tip de baze de date ar face să o dispară.

Dezvoltarea tematicilor DevOps pornește de la ușurința cu care ne putem crea o bază de date folosind containere, întrucât instalarea unei baze de date pe calculatorul

personal în timp ce lucram la proiectele de facultate a făcut să am experiențe în care la finalul fiecărui an îmi reinstalam sistemul de operare doar pentru că o dată ce instalam o bază de date crea suficiente servicii, foldere care după o dezinstalare încă rămâneau. Folosind Docker, pot crea o bază de date cu o singură linie de cod, care în același timp să fie preconfigurată și să o pot împărtăși cu colegii de proiect ca să nu pierdem timp în configurarea mediului de lucru.

Lucrarea mea va urma o structură în care prezint noțiunile teoretice la început iar la final încerc să aplic aspectele prezentate în crearea unei aplicații. Inițial pornesc cu prezentarea microserviciilor continuând cu diferitele moduri în care acestea pot fi lansate, imediat după cu dezvoltarea noțiunilor de DevOps ce au ca scop accelerarea dezvoltării, iar la final o prezentare generală a modului în care îmi construiesc aplicația.

Capitolul 2

Preliminarii

Cel puțin până la momentul la care am decis ce temă să aleg pentru lucrarea mea de licență niciun curs de bază nu a abordat în detaliu tema pe care vreau să o dezvolt, din acest motiv lucrarea mea de licență nu este doar aprofundarea unui concept și crearea unei aplicații, ci încercarea familiarizării cu obiectele de lucru în același timp. Desigur, unele aspecte s-ar putea desprinde și în urma practicii în industrie dar nu a fost suficient ca să pot să consider că ma cunosc de bază în acest domeniu și doar încerc să găsesc lucruri cât mai specifice.

De asemenea, microserviciile sunt o arhitectură destul de nouă ce a luat amploare datorită limitărilor din punctul de vedere al dezvoltării și scalării aplicațiilor monolitice, astfel în opinia mea, este destul greu să vizualizezi anumite lucruri fără experiență la prima mână cu limitările acestea. Întrucât serviciul pe care îl dezvolt nu ar necesita neapărat o arhitectură bazată pe microservicii, ar funcționa complet normal și aș reduce foarte mult din complexitate dacă ar fi dezvoltată în mod normal.

La fel și cu practiciile DevOps, acestea se axează mai mult companiilor care au nevoie de timp de penetrare a pieții foarte rapid, însă eu doar încerc să mă familiarizez cu setarea și folosirea serviciilor pentru uzurile mele proprii. Un lucru important ar fi ca ce folosesc să nu mă limiteze, de exemplu infrastructura de care am nevoie va proveni din Azure doar pentru că primesc credite gratuite ca și student, însă în absența lor probabil aș încerca lucruri cu cost minimal sau chiar nici să nu le fac, și probabil aș avea același randament.

Capitolul 3

Microservicii

3.1 Concepte generale

Microserviciile sunt o arhitectură de sisteme distribuite ce a devenit populară în ultimele decenii datorită necesității scalării. Ceea ce o separă de o arhitectură orientată pe servicii este faptul că microserviciile sunt mult mai specifice, acoperind o singură parte din afacere și faptul că pot fi lansate în execuție independent, astfel microserviciile ar trebui să depindă cât mai puțin de altele. Comunicarea se face direct către interfețele expuse de fiecare, iar informațiile reținute de fiecare (de exemplu baza de date de date sau diferite fișiere) pot să fie modificate doar prin metodele de comunicare oferite ci nu direct.

Un concept important în microservicii este ascunderea de informații, astfel din exterior fiecare ar trebui să fie tratat ca o cutie neagră ce face anumite lucruri și expune anumite date, nu contează detaliile de implementare și nici tehnologiile folosite în implementare, ci doar capabilitățile fiecăruia. Acest lucru permite să modificăm microserviciul în funcție de cerințe, chiar și refactorizarea metodelor inițiale cât timp se respectă capabilitățile pe care microserviciul ar trebui să le îndeplinească.

Cel mai important aspect al microserviciilor este capacitatea de a lucra independent de celelalte. Acestea trebuie să funcționeze independent de celelalte iar dacă aducem schimbări într-unul atunci nu suntem obligați să facem schimbări în alt microserviciu, în caz contrar ar apărea dificultăți în lansare. Obținerea independenței poate să aducă contribuții majore în timpul de deployment însă implementarea este mult mai complicată datorită necesității comunicării cu alte servicii.

Un alt aspect de bază este acoperirea unui singur segment din afacere prin intermediul unui microserviciu. Alături de caracteristica precedentă, dacă afacerea și-ar dezvolta necesitățile ar fi mult mai dificil să modificăm mai multe microservicii și ulterior să coordonăm lansarea lor.

Independența microserviciilor este importantă, în acest scop, ele ar trebui să fie singu-

rele care ar putea să își modifice starea. Întrucât starea se referă la elementele componente, de exemplu o bază de dată, acestea nu ar trebui să fie împărtășite și nici accesate de alte microservicii. De asemenea ar trebui evitate modificarea interfețelor declarate pentru a evita necesitatea modificării altor microservicii. „Dimensiunea microserviciilor ar trebui să fie redusă prin oferirea unui număr restrâns de funcționalități declarate pentru a fi ușor de înțeles și întreținut. Însă important este cât de multe microservicii pot fi administrate întrucât un număr mai mare de microservicii, deși mici ca funcționalități, aduc dificultăți în comunicare, scalare și depanare.

Microserviciile sunt agnostice din punct de vedere al tehnologiilor care ar putea fi folosite. Din acest motiv, acestea oferă flexibilitate însă prețul plătit este crearea a mai multor puncte vulnerabile.

Orice prezentare al acestui tip de arhitectură nu ar fi completă fără prezentarea conceptului de „monolith” (monolit), întrucât este considerată o arhitectură veche reprezentând modul de funcționare din trecut. În cel mai simplist mod de funcționare, sistemul monolitic este reprezentat de un singur proces „gigantic” ce acoperă toată funcționalitatea sistemului, în general conectat la o singură bază de date ce și ea este reprezentată de un singur proces. Putem intui problemele cu această organizare, dacă conexiunea către procesul monolitic cade sau baza de date are probleme, întreg sistemul este afectat. Microserviciile încearcă să acopere aceste probleme prin independență, astfel dacă un microserviciu este căzut, celelalte nu sunt afectate. Procesul poate fi însă separat pe module dezvoltate separat însă care la final se cuplează formând un singur proces. Există conceptul de sistem monolitic distribuit în care acesta este la fel separat în module independente însă care nu ar funcționa dacă nu sunt toate active, acest tip de sistem oferă toate dezavantajele sistemelor distribuite dar și a monolitului ceea ce îl face destul de rar întâlnit.

Chiar dacă o arhitectură monolitică prezintă dezavantaje evidente ce sunt ușor de văzut, acestea nu sunt la fel de mari pentru companiile mici. Numărul redus de aplicații ce trebuie lansate și monitorizate, posibilitatea de reutilizare a codului mult mai ușoară, obținerea infrastructurii pentru proces și uneori chiar și scalarea, deși este limitată poate să facă o arhitectură monolitică mult mai atractivă. Nu ar trebui să asociem arhitectura monolitică ca un lucru antic, ci să o considerăm ca o opțiune. Însă pentru o organizație mică, acesta ar trebui să fie doar un punct de plecare, eventual după ce serviciul oferit devine mai popular s-ar putea să ajungem să cunoaștem limitările arhitecturii și să ne orientăm tot către microservicii, însă dacă nu avem succes, eliminăm multe complexități apărute în urma introducerii microserviciilor.

Principalele avantaje ale microserviciilor provin de la baza acestora, o arhitectură distribuită, însă cuplat cu conceptul de ascundere a informației și domain-driven design (design orientat domeniu) aduc multe alte avantaje asupra altor arhitecturi distribuite. În cadrul unui sistem format din microservicii putem folosi orice tip de tehnologie pentru

program și orice tip de bază de date, întrucât ascundem implementarea de exterior. Astfel putem alege tehnologii ce aduc avantaje în dezvoltarea microserviciului. Acest avantaj duce la capacitatea de a folosi tehnologii noi fără a afecta tot sistemul, însă limitează capacitatea de a împărtăși cod (de exemplu prin librării interne, întrucât acestea ar trebui să fie rescrise pentru fiecare limbaj). Un sistem distribuit rezistă mult mai ușor la căderi, întrucât acestea pot fi tratate ca să se prevină un lanț, însă cu un număr mare de puncte slabe devine greu de aproximat cât de mult este afectat un sistem. Scalarea unui sistem monolitic este ușoară, doar replicăm procesul, însă la microservicii scalarea poate fi mult mai concentrată pe punctele care chiar au nevoie să fie scalate. Lansarea unei schimbări într-un microserviciu este mult mai ușoară întrucât nu necesită crearea de timp mort în aplicație, întrucât nu tot sistemul este înlocuit ci doar o mică parte din acesta, ce poate fi chiar și mai controlată prin diferite aplicații de orchestrare.

Adaptarea microserviciilor vine cu un cost, pe lângă dificultățile de implementare a design-ului apar și alte probleme, de exemplu dezvoltarea locală. Atunci când dezvoltăm un microserviciu acesta, posibil, este nevoit să comunice cu alte microservicii care la rândul lor comunică cu alte microservicii și putem continua, toate acestea microservicii trebuie să funcționeze în același timp pe laptop-ul dezvoltatorului, întrucât nu putem folosi microserviciile din producție, însă nu e posibil să rulăm foarte multe microservicii pe un singur computer depinzând de configurație. Adoptarea microserviciilor, în general nu poate avea succes decât dacă este combinată cu elemente de DevOps, ceea ce înseamnă ca dezvoltatorii și operatorii să învețe tehnologii noi, ceea ce este un impediment în momentul în care vrem să lansăm un produs rapid în piață și nu avem experiența necesară. Microserviciile necesită mult mai multă infrastructură pentru a fi rulată, mai multă tehnologie ce trebuie să fie folosită pentru ca dezvoltarea și livrarea să decurgă eficient, însă pe termen lung acesta poate să fie prevăzută dacă arhitectura este folosită cum trebuie prin livrare mai rapidă și mai eficientă deci profitul ar putea să crească. Monitorizarea sistemului devine mai complicată, într-un proces monolitic toate fișierele de jurnalizare (logging) sunt în același loc. Într-un sistem distribuit, acestea trebuie să fie colectate, și ulterior ansamblate pentru a avea sens, acest lucru devine mai dificil când intră în discuție fenomenul de replicare al unei instanțe. Securizarea unui sistem distribuit este mai grea, traficul de date se face mult mai mult prin rețele sau prin diferite căi alternative ceea ce necesită un grad ridicat de atenție. Testarea unui sistem format din microservicii este mult mai dificilă, mai ales când trebuie testate mai multe microservicii în același timp, acesta duce la creșterea timpului necesar pentru a primi răspuns. Sistemele distribuite cresc timpul de răspuns întrucât acestea nu se mai face în cadrul unui singur proces, deci datele trebuie codificate și decodificate atunci când trec dintr-un mediu în altul. În cadrul unui sistem distribuit o altă problemă care apare este consistența datelor, dată de faptul că în timpul unei cereri, datele dintr-un microserviciu se pot schimba, astfel tranzacțiile devin dificile mai ales când înainte ne bazam pe atomicitatea și consistența oferită de

baza de date.

Astfel, microserviciile funcționează cel mai bine pentru organizațiile mari ce vor ca dezvoltatorii lor să livreze facilități noi foarte repede care să nu fie întârziate de compatibilitatea cu caracteristici deja existente. Prin natura variată a acestei arhitecturi se pot folosi tehnologii noi dar și folosirea platformelor cloud, astfel se oferă multă flexibilitate cu prețul creșterii complexității. Pentru companiile mici, al căror rată de succes este greu de calculat, overhead-ul produs de adaptarea unei arhitecturi în acest stil nu ar aduce suficient de multe avantaje și ar trebui să opteze pentru o arhitectură tradițională ce este stabilă și oferă colaborare mult mai ușoară în cadrul aceluiași proiect la scale mici.

3.2 Design

Partea de creare a schemei arhitecturale a unui sistem este una dintre cele mai importante faze ale dezvoltării. Acesta este momentul în care putem descoperii că microserviciile nu ar fi chiar o alegere potrivită pentru cerințele noastre. Din acest motiv este important să nu ignorăm potențiale probleme ci să se formuleze toate problemele pe care le putem aștepta în implementare

3.2.1 Interconectare

Deși microserviciile pot fi lansate și pot funcționa independent, acestea trebuie să funcționeze în cadrul unui sistem complex. Astfel trebuie definite modul de comunicare și interfețele expuse. În definirea interconectării întâlnim câteva concepte de bază prezentate anterior ce provin din descompunerea în module.

Ascunderea de informații este procesul de separare al detaliilor de implementare pentru a evita modificarea în alte locuri în cazul în care acestea se schimbă. Acest concept oferă varii beneficii precum îmbunătățirea timpului de dezvoltare datorită posibilității de dezvoltare în paralel, creșterea înțelegerii sistemului mult mai ușor întrucât fiecare modul poate să fie înțeles independent și flexibilitate oferit de faptul că putem schimba funcționalitate fără a fi necesar ca alte părți ale sistemului să fie modificate. David Parnas spune că „The connections between modules are the assumptions which the modules make about each other.” (Conexiunile create între module sunt ipotezele create de module față de celelalte) [12]. Astfel, cu cât alte microservicii folosesc un anumit microserviciu mai mult se crează mai multe locuri în care devine mai dificil de modificat, astfel cu cât întâmpinăm problema aceasta mai puțin cu atât putem lansa acel microserviciu fără a lua în considerare modificări care să pastreze compatibilitatea cu elementele existente.

Un alt concept important este coeziunea, scopul microserviciilor este de a funcționa independent, astfel vrem ca lucrurile care interacționează să fie grupate iar atunci când se schimbă ceva să nu trebuiască să modificăm multe locuri. Elementele care nu

interacționează ar trebui să fie separate pentru a nu crea probleme de regresie.

În sisteme distribuite apare cuplarea, atunci când un microserviciu depinde mai mult sau mai puțin de altul pentru a avea funcționalitate completă. Un sistem cu cuplare mică poate să fie lansat fără să trebuiască modificări în locuri multiple.

Cele două concepte definesc lucruri asemănătoare, astfel ele se pot rezuma ușor ca „A structure is stable if cohesion is strong and coupling is low.” [4] (O structură este stabilă dacă coeziunea este ridicată și cuplarea este scăzută). Coeziunea se referă la lucrurile din interior iar cuplarea se referă la exterior iar acest lucru este greu de balansat.

Cuplarea poate apărea sub diferite forme, cuplare de domeniu în care un microserviciu interacționează cu alt microserviciu pentru că are nevoie de funcționalitate ce depășește atribuțiile lui. Astfel de cuplare, în general, nu poate fi evitată într-un sistem distribuit. Acesta este cel mai scăzut nivel de cuplare însă trebuie să evităm cazul în care un microserviciu depinde un centru pentru alte microservicii întrucât ar însemna ca acesta are prea multe responsabilități. Există conceptul de cuplare temporară, adică atunci când un microserviciu așteaptă ca alt microserviciu să execute ceva, astfel cuplarea se întâmplă pentru că cele două operațiuni se întâmplă în același timp. Cuplarea Pass-Through apare atunci când un microserviciu este nevoit să trimită date unui alt microserviciu întrucât acesta trebuie să trimită datele unui microserviciu la un nivel mai mic, astfel un microserviciu are rolul doar de a transmite date. Acest lucru este problematic întrucât dacă formatul acestor date ar trebui să se schimbe, ar trebui să modificăm mai multe microservicii. Evitarea acestui tip de cuplare se face prin schimbarea modului prin care comunicăm de exemplu microserviciul ce emite date ar putea să trimită direct datele către microserviciul ce are nevoie de ele sau ca microserviciul intermediar să trimită datele în format pur, fără să îl intereseze o formă anume. Cuplarea comună apare atunci când mai multe microservicii încearcă să acceseze o resursă comună, de exemplu o bază de date sau locația unui fișier. Problema în acest caz este că dacă modul în care această resursă interacționează se schimbă, mai multe microservicii trebuie modificate. Sau dacă unul din microservicii are nevoie de un comportament special, alte microservicii ar trebui să se adapteze ceea ce crează probleme în dezvoltare. Soluții posibile de evitare pot fi crearea unui alt microserviciu ce funcționează ca o interfață pentru resursa comună ce va funcționa după principiile unui microserviciu. Cuplarea de conținut apare atunci când un microserviciu modifică starea internă a altui microserviciu, de exemplu prin modificarea bazei de date a acesteia. Acest lucru este periculos întrucât pot apărea modificări nedorite prin utilizare incorectă, astfel se încalcă procesul de ascundere de informații.

3.2.2 Domain Driven Design

Domain Driven Design este un concept introdus de Eric Evans în cartea sa [5]. Acesta se bazează pe crearea codului în jurul afacerilor pentru ca acesta să fie mai ușor de

înțeles prin creșterea interacțiunii cu analiștii de afaceri, acest lucru este benefic pentru microservicii întrucât acestea se pot orienta către mediul de afaceri pentru a se dezvolta alături de acesta fiind mult mai eficient.

Unul din concepte este limbajul universal ce cuprinde ca codul din interiorul unui microserviciu să folosească termeni folosiți de către utilizatori și de către analiști, acest lucru este benefic întrucât face comunicarea mult mai eficientă, o persoană nou venită poate să înțeleagă afacerea și codul în același timp. Dacă cele două sunt diferite, atunci pot exista probleme în comunicarea problemelor.

În DDD, un agregat este o unitate ce are o anumită stare ce se poate modifica în timp și are ca scop modelarea unui concept din domeniul real. În general un agregat este cuprins în întregime de către un microserviciu chiar dacă acesta poate conține mai multe microservicii. Un agregat poate fi schimbat din exterior însă poate să își controleze singur starea pentru a preveni stările ilegale

Mărginirea contextului se referă la ascunderea mai multor activități ce țin de implementare cu scopul de a defini anumite responsabilități sigure pe care le îndeplinește. Acest lucru se poate îndeplini prin ascunderea modelului, atunci când vrem să expunem o entitate nu trebuie să oferim toate informațiile despre acesta ci doar cele necesare. Distribuirea unui model, adică a unei entități între mai multe microservicii se face prin traducerea acestei entități în domeniu propriu.

DDD în microservicii este un concept fundamental întrucât ambele pun la centru afacerea și dezvoltarea în jurul ei. Astfel conceptele prezentate în DDD pot fi aplicate microserviciilor mărginirea contextului are ca bază ascunderea de informații și duce la crearea de formate standard.

3.2.3 Concepte de programare orientată obiect în microservicii

Multe din problemele cu care ne confruntăm în microservicii pot fi tratate folosind concepte ce au ca scop ușurarea modului în care creăm sisteme prin intermediul claselor în programarea orientată obiect. Deși diferite în natură, ambele au ca rol definirea unor interacțiuni între elemente ce sunt făcute să funcționeze separat sau prin intermediul unuor interfețe.

Începând cu conceptele de bază, abstractizarea se referă la ascunderea implementărilor interne și afișarea doar facilitățile oferite de clasă ce pot fi accesate din exterior. Acest lucru face referire la conceptul de ascundere al informațiilor descris anterior.

Encapsularea se referă la agregarea de date și attribute, adică funcționalități pentru a crea un tot unitar, acest principiu este compatibil cu microserviciile și nevoia acestora de a funcționa independent folosind datele proprii și permite o securizare mai bună a stării interne.

Moștenirea se referă la capacitatea unei clase de a prelua attributele și funcționalitățile

unei clase deja existente, deși mult mai greu de imaginat în microservicii, moștenirea în microservicii ne permite să creăm microservicii ce extind pe cel existent fără a cauza probleme.

Polimorfismul este un concept ce face referire la capacitatea accesării a unor multiple entități prin intermediul unei singure interfețe, în contextul microserviciilor putem spune că un microserviciu ar putea să înlocuiască alt microserviciu dacă are aceeași interfață de comunicare și duce la crearea a unor aceleași rezultate.

Pentru microservicii putem aplica și concepte mai complicate de OOP precum principiile SOLID. Principiile SOLID sunt un set de reguli cu scopul de a oferi dezvoltatorilor o metodă de a scrie cod, astfel încât acesta să respecte cerințele actuale și să permită dezvoltări/modificări ulterioare cu ușurință fără a avea dificultăți cu metodele deja existente. SOLID este un acronim iar fiecare literă face referire la un principiu. Principiul responsabilității unice (Single Responsible Principle) se referă la necesitatea ca fiecare clasă și metodă să aibă o singură funcționalitate, făcând asta reducem complexitatea dar și eventuale redundanțe create prin folosirea aceluiași cod în mai multe locuri făcând ca fiecare porțiune din cod să aibă un singur motiv de a exista. Principiul Deschis – Închis (Open Close Principle) conduce la faptul că entitățile scrise trebuie să fie deschise pentru moștenire dar să nu poată fie modificate, astfel asemănător cu principiul anterior, codul deja scris nu va fi afectat întrucât noile modificări se aduc asupra unei porțiuni noi. Principiul substituției al lui Liskov (Liskov Substitution Principle) se referă la faptul că clasele derivate trebuie să acopere funcționalitățile clasei pe care o moștenesc astfel încât acestea să poată să le înlocuiască fără modificări adiționale. Principiul separării interfețelor (Interface Segregation Principle), clasele ce implementează interfețe nu ar trebui să implementeze metode de care nu au nevoie, astfel dacă este nevoie să adăugăm o metodă nouă unei interfețe aceasta trebuie să fie implementată doar de către obiectele care au nevoie, dacă acest fapt nu este satisfăcut trebuie ca interfețele să fie separate ca să nu forțăm alte clase să implementeze metode noi. Principiul inversării dependențelor (Dependency Inversion Principle) se referă că modulele de nivel înalt nu ar trebui să depindă de cele de nivel scăzut, ele trebuie să fie abstractizate astfel acestea pot fi modificate în orice moment dacă vrem să aducem un alt modul cu funcționalități similare dar cu implementări diferite.[3]

3.2.4 Migrarea

În planificarea urbană există conceptul de dezvoltare pe spațiu verde. Aceasta se referă la dezvoltarea pe un loc în care la momentul începerii lucrărilor este spațiu verde ce nu necesită costuri adiționale pentru curățare. Atunci când vrem să construim un sistem bazat pe microservicii pornind de la zero, apare acest fenomen. Însă de multe ori există un sistem existent ce ajunge la anumite limitări ce fac dorința de trecere către

microservicii. Acest lucru face referire la conceptul de dezvoltare pe spațiu maro, adică pe un spațiu post industrial în care există costuri de demolare sau restaurare și ce necesită o atenție mult mai ridicată datorită posibilității alterării compoziției pământului ce ar putea să fie toxic.

Migrarea unei aplicații monolitice nu ar trebui să fie făcută cu scopul a obținerii unei aplicații formate din microservicii ci pentru că există o funcționalitate ce este împiedicată de a funcționa la capacitățile necesare pentru funcționare optimă, astfel migrarea unui sistem monolitic ar trebui să fie făcută incremental și posibil nu complet dacă nu este nevoie.

Începutul migrării ar trebui să fie făcut cu o analiză a sistemului și identificarea dacă avem posibilitatea de a întreține microserviciile, acest lucru înseamnă să avem prezentă infrastructură și oameni care să știe să configureze uneltele necesare precum cele de CI/CD pentru a livra cu aceeași consistență. Ulterior putem analiza sistemul pentru a observa locurile unde se formează coeziune fapt ce duce la posibilitatea creării unui microserviciu cu cod grupat. Pentru acest lucru se pot folosi unelte precum CodeScene ce pot identifica porțiunile de cod ce se schimbă împreună. Acest lucru ajută în crearea unui context mărginit și să identificăm modul în care acesta ar funcționa cu exteriorul.

O dată ce identificăm lucrurile ce vrem să le separăm putem separa funcționalități și elementele conținute, UI-ul, backend-ul și baza de date. Am fi tentați să ignorăm o parte din aceste componente însă microserviciile funcționează cel mai bine atunci când toate acestea sunt separate prin simplul motiv ca chiar dacă UI-ul nu este 1:1 cu backend-ul, acestea tot vor fi modificate în același timp. În acest scop, există două modalități în care putem face migrarea, pornind de la cod, însemnând că separăm funcționalitatea însă în continuare folosim datele din baza de date monolitică. Acest tip de abordare este destul de simplu, întrucât obținem rezultate bune pe termen scurt, însă există limitări întrucât este posibil să nu putem separa datele din baza de date, fapt ce înseamnă că munca depusă pentru separare ar fi inutilă. Separarea datelor mai întâi se referă la crearea unei baze de date separate pentru datele pe care vrem să le folosim, acest tip de abordare aduce mai multe probleme la început întrucât ne face să ne gândim la probleme precum pierderea integrității datelor.

Atunci când facem modificări de tipul acesta trebuie să avem grijă cum orchestrăm schimbările. O modalitate ce provine de la un arbust tropical „Strangler Fig” se referă la crearea unui strat de interceptie ce verifică dacă funcționalitatea apelată a fost implementată într-un microserviciu, și dacă este atunci va ruta traficul către microserviciu. Acest tip de implementare permite să nu modificăm aplicația monolitică și să putem să ne întoarcem la ea în cazul în care avem probleme. De asemenea, dacă datele sunt comune putem folosi load balancing între cele două la nevoie. În rularea paralelă, microserviciul și aplicația monolitică funcționează în paralel și ne permite să le comparăm ca să ne asigurăm că avem funcționalitatea dorită. Metoda comutatorului ne permite să avem cele două

sisteme aplicația monolitică și microserviciul să funcționeze în același timp și să comutăm între cele două foarte ușor printr-un proxy.

În momentul în care începem separarea datelor, apar probleme ce nu pot fi prezise cauzate de lipsa unei baze de date monolitice. De exemplu, performanța unei baze de date este extrem de bună când este vorba de crearea de cereri ce conțin operații de alăturare de tabele, astfel noi trebuie să mutăm aceste operații extrem de optimizate în interiorul microserviciilor. O dată cu lipsa unei baze de date comună pentru toate tabelele apar problema lipsei de constrângeri, nu ne mai putem baza pe baza de date să forțeze aceste constrângeri și trebuie să verificăm și să tratăm erorile în cazul în care apar erori. Tranzacțiile nu mai sunt garantate. O dată cu separarea datelor se pierde ACID-itate, o altă problemă a microserviciilor atunci când avem tranzacții distribuite între mai multe microservicii. În funcție de necesități ar fi necesar să rulăm cereri direct în baza de date sau să oferim acces către persoane din exterior către aceasta, însă prin conceptul de ascundere de informație acest lucru nu ar fi posibil. Ca să rezolvăm această problemă putem crea o copie a bazei de date, read only, pe care microserviciul o va popula o dată ce populează baza de date inițială.

Astfel, migrarea unui sistem monolitic este extrem de dificilă și vor apărea probleme, din acest motiv este recomandat să folosim o abordare incrementală în care mutăm funcționalități una câte una, acoperindu-ne obiectivul pe care ni-l dorim prin introducerea arhitecturii formate din microservicii.

3.3 Tehnici de implementare

Trecerea de la o aplicație monolitică la una bazată pe microservicii uneori nu este ușoară. Unul din principalele motive este o posibilitate de a regândii modul cum accesăm anumite elemente. Atunci când totul se află în cadrul unui singur proces, accesarea datelor devine mai grea întrucât nu avem acces la ele și uneori nu este rentabil să căutăm după ele. Atunci când apelăm o funcție local, timpul necesar pentru executarea acesteia este neglijabil, atunci când această funcție trebuie să apeleze un alt microserviciu, lucrurile se complică și trebuie decis dacă chiar avem nevoie. Un apel făcut printr-o rețea mereu va avea o întârziere, de asemenea cresc locurile în care acesta poate eșua. În dezvoltarea aplicațiilor mobile întrucât ne așteptăm ca aplicația să reacționeze la atingeri, nu putem să facem anumite operații pe thread-ul principal, acestea pot fi apelarea lucrurilor de pe Internet sau accesarea unei baze de date (chiar și dacă este de tipul SQLite3 fiind locală), aceasta pune în perspectivă impactul pe care acest tip de apeluri le pot avea.

Atunci când modificăm modul în care comunică un microserviciu trebuie să avem grijă la modul cum ar afecta alte microservicii, sau atunci când acest lucru este necesar trebuie să orchestrăm aplicarea acestor schimbări. Un alt concept ce într-o arhitectură monolitică se tratează ușor este tratarea erorilor, întrucât toate se întâmplă în același mediu de lucru

putem avea o viziune asupra ce poate să cadă și în ce mod. Într-un sistem distribuit, pot apărea mai multe tipuri de erori care se datorează rețelor, de exemplu crash-uri de microservicii ce necesită resetări, erori de omisie de mesaje, când încercăm să apelăm un microserviciu însă acesta nu trimite un răspund, în același timp acest răspund poate să fie mult prea rapid și să nu putem să îl recepționăm sau prea târziu și să considerăm că nu am primit, dar și erori arbitrare precum unele în care ne așteptăm la un răspund și nu primim ce avem nevoie.

Toate aceste aspecte trebuie tratate atunci când lucrăm cu microservicii, însă fiind vorba de o arhitectură paralelă ce oferă independență avem flexibilitatea de a alege ce folosim și ce nu. De aceea este important să alegem tehnologii și stiluri de comunicare care oferă performanțele dorite încercând să minizăm timpul de răspund și tipurile de avarii ce pot apărea.

3.3.1 Tipuri de comunicare

La fel ca în comunicarea între procese sau în cadrul unui proces, există comunicare sincronă ce introduce blocare la un anumit nivel și comunicare asincronă ce permite procesele să trateze cererile în momentul în care pot, aducând îmbunătățiri ale performanței dar crescând dificultatea înțelegerii și a modurilor în care aplicația poate cădea.

Cel mai simplu de înțeles mod de comunicare este comunicarea sincronă bazată pe cerere și răspund. În cadrul acestuia un microserviciu face o cerere (indiferent de formă) către un alt microserviciu și așteaptă un răspund. Acest scenariu este extrem de familiar, semănând cu cererile către o bază de date sau către un API extern, ceea ce le face ușor de implementat însă nu este o tactică eficientă în cadrul unui sistem distribuit întrucât aduc dezavantaje precum apariția acestui blocaj, creșterea latenței dar și apariția acestui cuplaj temporar întrucât se poate întâmpla ca unul dintre microservicii să nu fie același ca cel care a răspund întrucât a fost necesară o resetare. Acest tip de comunicare poate fi folosită pentru microservicii simple dar trebuie evitat crearea unui șir întreg de cereri întrucât crește timpul total de răspund dar și apariția unui eșec în cascadă.

Însă, o arhitectură paralelă are posibilitatea creării a unui tip de comunicare asincronă ce nu blochează funcționarea microserviciilor, adică ele nu așteaptă un răspund direct. Avantajul unui tip astfel tip de comunicare este posibilitatea tratării operațiunilor ce ar dura mai mult întrucât se elimină conceptul de cuplare temporară. Dezavantajul este creșterea nivelului de complexitate și creșterea dificultății de monitorizare și de a manevra erorile, însă ar trebui să fie utilizată atunci când avem procese de lungă durată sau pentru a comunica cu mai multe microservicii în același timp.

Cel mai simplu mod de a implementa comunicarea asincronă este crearea de resurse partajate. Un microserviciu adaugă date, fie într-o locație de stocare, adăugăm un fișier sau la o bază de date adăugăm alte date, iar celelalte microservicii ce vor să acționeze în

urma primirii acestei informații doar verifică apariția de date noi, iar atunci când apare o prelucrează. Dezavantaje includ necesitatea tratării cazurilor în care microserviciile ce scanează sunt căzute, iar în momentul în care vor reveni sunt nevoie să trateze date noi dar și faptul că se realizează o formă de cuplare asupra domeniului partajat. Această formă de comunicare poate fi folosită atunci când suntem restricționați de tehnologiile folosite, întrucât orice tip de sistem poate să citească date dintr-un fișier dar și atunci când vrem să trimitem un volum de date ridicat.

Un alt tip de comunicare este cel sub formă cerere-răspuns, acesta poate fi sincron sau asincron. În mod blocant, microserviciul ce face cererea va aștepta un răspuns iar în funcție de acesta își va continua activitatea. În mod asincron, se va proceda în mod asemănător însă de data aceasta nu se va mai aștepta un răspuns, sau cel puțin nu unul cu informația cerută. Acest lucru este mai dificil, întrucât pentru a răspunde cererii, microserviciul trebuie să trimită unui microserviciu specific răspunsul creat și există posibilitatea ca microserviciul ce inițiază cererea să nu mai existe astfel am putea încerca ca orice microserviciu să poată să trateze orice cerere. Putem folosi acest mod de implementare atunci când avem nevoie să avem o confirmare la cererile făcute și de asemenea este destul de ușor de implementat, cel puțin varianta blocantă.

Una dintre comunicările reprezentative pentru microservicii este cea făcută prin evenimente, este asincronă în natură și se bazează că atunci când în cadrul unui microserviciu se întâmplă un lucru pe care consideră că ar trebui să îl comunice în exterior, acesta emite un eveniment cu diferite informații astfel celelalte microservicii pot asculta iar în momentul primirii acestea vor reacționa independent. Este un opus al modelului anterior întrucât în acest tip de comunicare, în general, nu avem cunoștințe dacă evenimentul a ajuns către celelalte părți sau dacă cineva ascultă pentru acest tip de evenimente ci eventual dacă a fost procesat de către Event Bus. Comunicarea prin evenimente oglindește modul de organizare al unor echipe autonome din cadrul companiilor, acestea anunță atunci când produc un rezultat nou și nu primesc feedback direct dacă s-a acționat după mesajul lor în mod implicit, întrucât comunicarea poate fi un mail general la o listă de abonați, aceștia ar primi doar confirmarea că serverul ce procesează mail-urile a validat cererea. Implementarea constă în crearea unei metode de a trimite evenimente și de a recepționa. Un prim mod este folosirea de programe specializate precum Apache Kafka sau RabbitMQ ce livrează platforme dar și implementări pentru ambele probleme ce vin cu avantajul că toate problemele care le-am avea precum garantarea livrării și permit scalare însă sunt un sistem în plus de administrat. Un alt mod de implementare include folosirea specificațiilor declarate în Atom pentru a crea un flux web de resurse ce poate fi folosit pentru a trimite și consuma evenimente însă nu oferă latență mică. În general, ar fi bine să nu implementăm comunicarea prin evenimente de la zero întrucât este dificil de administrat și sunt multe probleme greu de tratat, de aceea putem folosi produse existente. Avantajul unui astfel de tip de comunicare este posibilitatea ca un eveniment să

fie tratat de mai multe microservicii în același timp, astfel evenimentul trebuie să conțină informații suplimentare pentru ambele microservicii, deci apare riscul de a distribui mai multe informații ceea ce induce diferite probleme de securitate, astfel putem folosi acest tip de arhitectură acolo unde putem profita de acest avantaj.

Deși comunicarea asincronă aduce îmbunătățiri în decuplarea sistemelor și în timpul de răspuns în cazul în care vrem să comunicăm cu mai multe evenimente acesta aduce complexități ce devin greu de depanat, astfel avem nevoie de monitorizare și moduri de identificare a modului evoluției cererii implementate corespunzător.

3.3.2 Tehnologii de comunicare

Am menționat anterior că microserviciile ne cumpără flexibilitate, însă din acest motiv avem și mai multe probleme întrucât mereu când alegem un tip de tehnologie trebuie să limităm dezavantajele și în același timp să maximizăm avantajele pe care le poate aduce. Astfel putem să alegem tehnologii mai vechi care au fost testate dar limitate sau tehnologii noi care promit că aduc îmbunătățiri dar e posibil să nu fie lipsite de erori ce s-ar rezolva în timp.

Când alegem o tehnologie, trebuie să avem grijă că aceasta trebuie să fie compatibilită cu serviciile pe care le avem, de asemenea trebuie să nu ne limiteze în alegerea tehnologiilor viitoare. În acest scop modul de comunicare al microserviciilor ar trebui să fie clar definit întrucât nu putem fi siguri ce elemente sunt folosite și care nu. De asemenea tehnologia aleasă nu ar trebui să inhibe comunicarea cu alte microservicii.

Remote procedural call (RPC) (Apel procedural la distanță) este un mod de apelare de funcții local însă care se execută într-un serviciu la distanță. Pentru a face asta posibil sunt necesare definirea unor scheme explicite precum SOAP sau gRPC definite ca limbaje de definit interfețe (interface definition language). Definirea unor scheme explicite face ca folosirea tehnologiilor diferite să fie mai ușoară însă dacă nu vrem să folosim astfel de scheme putem folosi Java RMI care necesită folosirea aceleași tehnologii însă fără schemă definită. Folosind o astfel de tehnologie ne permite generare de cod pe care clientul le poate folosi pentru a comunica cu serverul făcând implementarea pe partea clientului mai ușoară întrucât nu trebuie să facă ceva în mod special. Limitările acestei tehnologii reprezintă restricționarea tehnologiilor folosite (mai ales în cazul Java RMI) însă celelalte menționate funcționează cu diferite tehnologii. Ascunderea faptului că prin aceste metode se face un apel peste Internet poate fi înșelător și ascunde complexitatea. De asemenea atunci când producem modificări asupra codului serverului trebuie să generăm codul pentru client, însă dacă adăugăm sau eliminăm câmpuri din interiorul funcțiilor, clientul trebuie să facă mai multe modificări deși el poate să nu folosească acele părți din cod. Acest tip de abordare este potrivit pentru comunicarea blocantă între două sisteme asupra căruia știm când se modifică lucrurile și în ce fel, astfel cunoaștem cât de multe modificări vor trebui

să fie făcute.

Representational State Transfer (REST) prin HTTP, în acest tip de arhitectură avem definite niște resurse ce pot fi accesate prin link-uri sub diferite forme ce permite identificarea și modificarea acestora în funcție de modul de accesare. Acest tip de abordare funcționează foarte bine cu verbele HTTP predefinite (GET, POST, UPDATE) întrucât știm la ce comportament ne putem aștepta indiferent de resursa pe care o accesăm. Faptul că folosim HTTP ne permite să folosim toate elementele dezvoltate pentru a îmbunătății sistemul precum caching, echilibrări de trafic, unelte de monitorizare existente dar și elemente de securitate precum autentificare și certificate. O altă trăsătură adusă în REST este Hypermedia as the engine of application state (HATEOS) ce permite să nu cunoaștem întreg URI-ul către resursă ci doar cel de început. O dată intrat pe prima intrare, aceasta va afișa alte intrări pe care putem intra și ce am putea accesa dacă intrăm pe ele. Acest lucru este util pentru că nu trebuie să definim niște căi explicite ci acestea pot fi căutate însă pentru a face asta cresc numărul de cereri pe care trebuie să le faci doar pentru a executa o singură comandă. Atunci când folosim această tehnologie nu avem o modalitate simplă de a crea librării pe care clienții serverului le pot folosi, din acest motiv s-a creat specificația OpenAPI din Swagger ce permite ca din anumite informații să generezi automat o documentație dar și cod pentru o varietate de limbaje însă nu a fost adoptat la fel de mult întrucât ar fi dificil de implementat pentru proiectele existente. Prin acest format putem trimite o varietate de formate de date însă din cauza cerințelor adăugate de HTTP atunci când creăm o cerere ar putea să împiedice performanțe ridicate atunci când nu avem date multe de trimis. Putem folosi acest mod de comunicare cam în toate situațiile în care avem nevoie de comunicare unu la unu între servicii, fiind un stil comun de comunicare este suportat de majoritatea tehnologiilor deci ne garantează ca am avea compatibilitate.

Ambele tipuri de apeluri au aceeași problemă, în unele cazuri ar trebui să facem mai multe apeluri pentru aceeași resursă întrucât nu avem suficiente date, sau în cazul în care un câmp dintr-o cerere se schimbă acesta ne-ar forța să modificăm în mai multe locuri. Astfel de probleme sunt rezolvate de GraphQL ce permite să specificăm exact ce câmpuri avem nevoie fără a fi nevoie să facem mai multe cereri. Deși ar ușura cererile la nivelul clientului, dezavantajele apar la server, în primul rând acesta este un Query Language, deci dacă cererea este complicată atunci server-ul va trebui să o proceseze, dacă se adună multe cereri de tipul acesta ar scădea performanța. De asemenea, întrucât de fiecare dată se face un query, caching-ul este mai dificil de realizat și nu poate fi folosit la fel de eficient pentru a scri informații.

În arhitectura bazată pe evenimente putem folosi agenți de mesaje (Message Brokers), aceștia funcționează ca un intermediar pentru comunicațiile între microservicii, astfel în loc să comunice direct acestea comunică prin intermediar trimițând orice tip de mesaje. Acestea oferă diferite facilități în modul în care trimit mesaje, printr-o coadă în care se

pun mesajele iar consumatorii doar preiau direct din coada aceasta sau prin topică ce permite consumatorilor să se aboneze la un anumit tip de mesaj iar atunci când un mesaj vine ce îndeplinește criteriul, acesta va fi trimis la toți abonații. Principalul motiv pentru care am folosi o astfel de tehnologie este pentru că oferă diferite caracteristici care ar fi greu de implementat precum garanția livrării, astfel dacă nu există nici-un microserviciu care să accepte un mesaj, atunci el va fi livrat la primul care va fi disponibil să prelucreze acel mesaj. Alte caracteristici includ garantarea ordinei în care mesajele sunt primite sau să îți permită să trimiți același mesaj către mai multe topic.

Asupra acestor diferite moduri de comunicare apar diferite moduri de a trimite un mesaj, referindu-ne la compoziția acestora. Astfel, acestea pot fi sub formă de text sau sub formă binară. În general, putem trimite sub formă text adică JSON sau XML însă dacă o să ne intereseze să eficientizăm viteza atunci va fi necesar să encodăm mesajele trimise și să le trimitem sub formă binară însă acesta nu este un proces ce ar aduce suficient de multe îmbunătățiri.

Atunci când am vorbit de comunicarea între microservicii am menționat importanța creării unei interfețe sau a unei scheme pentru a afișa modul cum comunică microsericiul cu exteriorul. Acest lucru permite consumatorilor să știe la ce să se aștepte și să implementeze mult mai ușor crearea legăturii iar pentru cei ce administrează microserviciul ar cunoaște ce câmpuri sunt expuse și ce trebuie să facă. Deși unele tehnologii necesită prezența unei scheme sub o formă sau alta, schemele ne ajută și în crearea unei documentații consistente ce ajută la implementare. Cu ajutorul unei scheme putem găsi probleme mult mai ușor, asemănător cu modul în care funcționează limbajele de programare ce necesită specificarea unui tip atunci când declarăm variabile și limbajele interpretate ce pot avea orice valoare, avem un nivel în plus de siguranță, cel la nivelul la compilare, astfel într-un limbaj în care acest tip de verificare ar lipsii, unele probleme apar doar în producție. Atunci când trebuie să modificăm o schemă aceasta se poate face în doua feluri, structural când câmpurile se schimbă, astfel clienții trebuie sunt nevoiți să administreze noul câmp, sau semantice atunci când câmpul nu își schimbă numele însă modul de calculare este diferit. Acest ultim tip este destul de periculos întrucât este ușor de ignorat dacă nu apar probleme evidente. În general este recomandat să creăm scheme de tipul acesta, mai ales dacă microserviciile sunt deținute de echipe diferite, în cazul acesta nivelul de comunicare dintre cele două este îmbunătățit iar problemele sunt rezolvate mult mai ușor.

3.3.3 Schimbări în funcționalitate

Atunci când este vorba de microservicii, este inevitabil ca funcționalitatea acestora să se schimbe la un moment dat, iar în cazul cel mai fericit în care cuplarea și coeziunea sunt favorabile, atunci cel mai probabil nu trebuie să modificăm alte microservicii, însă de multe ori acesta nu este cazul.

O primă parte sunt metodele de evitare ale acestor probleme, cel mai simplu prin asigurarea că funcționalitatea microserviciului nu este schimbată ci doar se adaugă funcționalitate. De asemenea, putem implementa în consumatori ca aceștia să fie mai toleranți și să încerce să își găsească datele de care au nevoie chiar dacă se schimbă schema, acest lucru se poate face și prin prelucrarea strictă a câmpurilor de care avem nevoie și nu de întreg fișierul. Putem folosi tehnologii care să permită adaptare ușoară, de exemplu atunci când serializăm un obiect în Java acesta are o anumită versiune, însă dacă modificăm obiectul adăugând un câmp această versiune s-ar schimba deși câmpurile de care am avea nevoie ar rămâne la fel, acest lucru creând o incompatibilitate, însă dacă am folosi ceva asemănător cu HATEOS ar fi mult mai ușor de adaptat fără nicio modificare. Este important ca orice modificare adusă datorită unei schimbări în funcționalitatea unui microserviciu să fie prinsă cât mai rapid, de asta sunt importante testele iar cu ajutorul schemelor definite explicit putem folosi diferite unelte ce pot valida diferențe între modificări la nivelul protoalelor, schemelor sau a specificațiilor OpenAPI.

Însă uneori este imposibil să mitigăm toate problemele și este nevoie să adăugăm schimbări pe care și alte microservicii sunt obligate să le facă, astfel apar dificultăți în orchestrarea ordinii acestor schimbări. Una dintre soluții este anunțarea schimbărilor pe care vrem să le facem și să lăsăm toți consumatorii să lanseze versiuni noi ce tolerează schimbările și după aceea să lansăm varianta schimbată. Acest lucru nu este în general suficient întrucât datorită modului de planificare aceste apariții acestor schimbări în consumatori poate să fie destul de lentă. O altă metodă de administrare a schimbărilor este de a lansa două servicii în paralel, una ce conține modificările și una care nu le conține pentru microserviciile care încă nu s-au adaptat. Acest lucru este de evitat din același motiv ca precedentul dar și faptul că crează multe dificultăți să ai mai multe servere de același timp lansate. De asemenea se poate păstra versiunea veche și doar crearea unei noi versiuni a endpoint-ului. Versiunile mai vechi pot fi schimbate să fie compatibile cu versiunea nouă și eventual retrase când suntem siguri că toți clienții au primit modificările necesare. Acest lucru se poate face prin monitorizarea traficului pe interfață. Acesta este tipul preferat de adus schimbări întrucât este cel mai convenabil pentru consumatori dar și pentru server.

3.3.4 Partajare între microservicii

Unul dintre cele mai simple concepte de programare ce aduce îmbunătățiri în scrierea codului este DRY, „Să nu te repeți” (Don't Repeat Yourself) acesta la prima vedere are ca sens să nu duplici codul întrucât dacă ar trebui să schimbi o funcționalitate e posibil să fie nevoie să modifice în mai multe locuri din cadrul sistemului ceea ce face ca modificările să fie mai greu de făcut.

În universul microserviciilor, acestea sunt separate din natură, astfel repetarea codului

în cadrul mai multor microservicii este inevitabilă, de asta am fi tentați să creăm librării comune (de exemplu pachete de NPM sau dependențe în Maven) însă chiar dacă inițial ar fi mai ușor de scris acest lucru generează o cuplare ridicată între aceste sisteme deoarece atunci când o librărie comună este actualizată atunci sistemele trebuie să implementeze versiunea nouă, însă nu de multe ori este posibil acest lucru întrucât nu toate microserviciile ar putea fi actualizate în același timp, în acest scop trebuie să înțelegem că librăria creată de noi va fi consumată sub forma unor versiuni multiple de clienți diferiți.

Un mod avansat de a folosi diferite microservicii este crearea unei librării pentru acel microserviciu pe care mai târziu o oferim altor consumatori. Când este vorba de echipe mici, este dificil de implementat întrucât logica de server ar trebui să fie diferită de logica ce se găsește în librăria de client, când librăria ar trebui să funcționeze ca o abstractizare al API-ului pe care îl oferă microserviciul, astfel e indicat ca cele două să fie separate. La fel ca librăriile trecute trebuie să ne obișnuim cu ideea că librăria va fi consumată în mai multe versiuni.

Când lansăm microservicii trebuie să ne asigurăm că acestea pot să comunice între ele. Acest lucru, în general, este dificil de implementat întrucât microserviciile pot rula pe servere care dispar și apar din nou constant, mai ales când e vorba de infrastructură oferită ca serviciu. Sisteme de orchestrare precum Kubernetes ne oferă o metodă simplă de configurare a microserviciilor ce se află într-un singur cluster însă atunci când introducem mai multe, devine mai complicat. Însă chiar și așa nu ne putem baza pe serviciul de orchestrare să ne facă legătura întrucât dacă avem un număr mic de microservicii nu ar fi suficient de rentabil să rulăm o astfel de tehnologie.

O metodă simplă folosită și pentru Internet este Domain Name System (Sistem de nume de domeniu) prescurtat DNS. Acesta ne permite să asociem unui IP un nume, astfel nu ar trebui să cunoaștem exact IP-ul sistemului cu care vrem să comunicăm ci doar numele acestuia. Avantajele acestei metode este familiaritatea tuturor persoanelor cu acest lucru, din acest motiv ansamblarea unui sistem nu ar trebui să fie grea, însă în cadrul microserviciilor apar mai multe probleme datorită faptului că acestea pot să fie înlocuite mereu de instanțe noi. Fiecare intrare de DNS are o perioadă în care nu este actualizată (Time to live - TTL), și poate fi cached sub diferite forme. Din acest motiv putem opta pentru varianta în care păstrăm în DNS doar calea către un load-balancer sub care se află microserviciile de care avem nevoie, însă această soluție aduce problema în care nu știm exact cu cine comunicăm iar dacă unul din ele are probleme acesta nu va fi evitat.

Limitările DNS-ului au adus la apariția tehnologiilor sub nume de Registrii dinamice de servicii Dynamic Service Registries ce presupune ca microserviciul să fie cel care se înregistrează la un serviciu central iar pentru ceilalți ce vor să îl acceseze vor lua intrarea din cadrul acestui centru. Astfel de sisteme includ ZooKeeper ce necesită rularea mai multor noduri într-un cluster ce poate funcționa ca un sistem de configurare și permite

crearea unei ierarhii în care se pot insera, modifica și interogă noduri sub diferite forme. Consul este asemănător cu tehnologia precedentă însă are mai multe facilități printre care include o interfață RESTful prin HTTP ce face integrarea cu diferite tipuri de tehnologii mult mai ușoară sau prin faptul că se pot genera fișiere de configurare ce vor fi actualizate în timp real, ceea ce permite ca programul să nu interacționeze cu Consul. O altă metodă am menționat-o anterior, adică prin Kubernetes, modul în care lucrează acesta este că prin faptul că microserviciile sunt lansate prin această platformă, aceasta poate să înregistreze modul în care a evoluat și să își păstreze un serviciu ce poate identifica locația pod-urilor lansate prin el. Kubernetes oferă multe servicii fără nicio configurare ce permite să evităm platforme dedicate ca cele două menționate anterior.

Serviciile Cloud în general se împart în regiuni ce oferă servicii. În cadrul acelor regiuni putem avea mai multe resurse ce ar fi grupate sub același centru. Putem defini ca trafic Nord-Sud, traficul ce se află în afara centrului de date (de exemplu clienți) și Vest-Est traficul ce se produce în interior.

Pentru a facilita comunicarea Nord-Sud există conceptul de API Gateway (Pasarelă API) ce funcționează ca un intermediar între persoanele din exterior ce vor să acceseze microserviciile din interior. Un segment de piață care s-a dezvoltat recent este acela de a monetiza API-uri create prin diverse metode precum abonamente cu un număr limitat de cereri. Când vrem să oferim un astfel de serviciu apariția unui intermediar ce poate să îndeplinească diferite facilități precum monitorizare, limitare de trafic sau funcționarea ca un reverse proxy și poate să ofere un fel de panou de control, însă în general este folosit de către companii pentru a-și accesa microserviciile din aplicațiile dezvoltate, acest lucru se întâmplă pentru că în general un API Gateway este destul de complex și uneori nu avem nevoie de toată funcționalitatea pe care ne-o oferă, de asemenea pot să fi și alte soluții care să fie mai eficiente. De exemplu, dacă vrem să externalizăm microservicii ce rulează în Kubernetes putem folosi soluții precum Ambassador. Dacă vrem un API Gateway întrucât ne așteptăm la mult trafic, atunci probabil nu avem nevoie de ceva central ci de mai multe. Un API Gateway poate funcționa însă trebuie să știm dacă se pliază cu funcționalitatea de care avem nevoie și să folosim alte tehnologii mai specializate dacă nu găsim.

Comunicarea Vest-Est se poate îmbunătăți folosind Service Meshes, aceasta îmbunătățește comunicarea între microservicii prin adăugarea unui proxy, astfel microserviciile nu ar comunica direct ci fiecare ar avea propriul proxy cu care comunică iar acestea vor comunica între ele. Ele sunt administrare de servicii de control și îmbunătățește traficul întrucât proxy-urile pot fi în aceeași rețea locală. Acest lucru simplifică dezvoltarea întrucât microserviciile doar fac cereri normale însă de fapt sunt rutate de către proxy. Deși utile, acestea induc multă complexitate în infrastructură ceea ce e bine să fie evitată.

Îmbunătățirea metodelor de comunicare între microservicii nu este suficientă pentru a crea un sistem complet pentru care este ușor să dezvoltăm. Am menționat importanța

schemelor explicite, acestea au ca rol îmbunătățirea comunicării capabilităților unui unct de acces prin prezentarea schemei însă nu este suficient, de asta e necesară documentație explicită. Standardul OpenAPI este eficient în crearea eficientă a documentației combinând cu Ambassador pentru Kubernetes poate să descopere automat puncte de acces OpenAPI și să afișeze documentația automat. Documentația pentru comunicarea bazată pe evenimente nu este la fel de dezvoltată însă avem opțiuni precum AsyncAPI sau CloudEvents.

Atunci când avem multe sisteme devine din ce în ce mai greu să cunoaștem exact ce face fiecare sistem, de asta e important să știm starea fiecăruia. Folosind ID-uri de corelare pentru tranzacții putem să vedem modul cum interacționează. Folosind servicii de descoperire precum Consul putem să vedem unde rulează microserviciile, OpenAPI ne arată ce capabilități are fiecare sistem și sistemele de monitorizare ne arată starea sistemelor. Toate informațiile acestea sunt disponibile prin intermediul diferitelor API-urilor de care dispunem atunci când folosim aceste tehnologii. De asta este destul de important să avem pagini dedicate care se actualizează în timp real ce arată nu doar documentația microserviciilor ci și starea acestora. O astfel de unealtă a fost construită la Financial Times, Biz Ops, ce oferă starea întregului sistem și calculează pentru fiecare o metrică ce ar indica dacă acestea au nevoie de schimbări. Astfel de unelte există disponibile și liber, precum Backstage de la Spotify.

O arhitectură distribuită ne permite creșterea eficienței prin distribuirea sarcinilor către mai multe sisteme de calcul. Într-o aplicație monolitică putem să obținem performanțe mai bune prin eficientizarea algoritmilor, prin folosirea thread-urilor în interiorului sau prin crearea unor procese copil, însă toate acestea rulează sub un singur procesor, ceea ce înseamnă că indiferent că se întâmplă aproximativ concurrent, nu este chiar în totalitate. Însă cu ajutorul microserviciilor, acest lucru este posibil.

Însă atunci când apare un lanț lung de instrucțiuni, înseamnă că sunt multe puncte în care acesta se poate întrerupe și să aibă nevoie să se facă o analiză a lucrurilor care s-au întâmplat și să se întoarcă la o versiune sigură. Acest concept apare în contextul prezenței unei baze de date, în general relaționale însă există și versiuni care nu sunt relaționale care promit același lucru. O tranzacție reprezintă un grup de instrucțiuni ce ar trebui să se execute împreună, dacă una din ele eșuează din diverse motive, atunci toate acestea sunt anulate și ne întoarcem la un punct de recuperare de la începutul tranzacției. Acest lucru se referă la „aciditatea” bazei de date ce fac parte din acronimul „ACID” ce reprezintă un set de proprietăți ce se referă la consistența datelor dintr-o bază de date. Consistența se referă că în cadrul unei operațiuni cu baza de date, la finalul acesteia este lăsată într-o stare validă pe care se poate continua să se execute operațiuni. Izolarea se referă ca multiple tranzacții se pot întâmpla în același timp fără ca cealaltă să fie afectată, acest lucru este asigurat de separare stărilor intermediare către o altă tranzacție. Durabilitatea se referă că la finalul unei tranzacții, datele nu sunt pierdute dacă sistemul cade.

În cadrul unui sistem format din microservicii, putem avea tranzacții de tipul acesta, în general atunci când microserviciul interacționează cu datele din baza de date proprie, iar microserviciul are control complet asupra acestora. Problemele apar atunci când modificările trebuie să fie făcute în cadrul mai multor microservicii în același timp iar asupra acestora să rămână proprietățile ce oferă consistența datelor. Acesta este un lucru mai dificil de implementat și e nevoie de mai mulți pași pentru a avea un comportament asemănător ca în cadrul unui sistem monolitic.

Cea mai simplă implementare a unei tranzacții distribuite este numită „Two-Phase Commit”, (Înregistrare în două faze) aceasta presupune o parte de inițializare și execuția propriu-zisă. În prima parte, un microserviciu ce vrea să inițieze o tranzacție va trimite către microserviciile implicate o înștiințare cu începerea tranzacției. Acestea verifică dacă pot să execute cerințele cerute și întorc un răspuns. Dacă toate răspunsurile sunt pozitive atunci se încep schimbările și se face commit-ul propriu zis. Din cauză că răspunsul acesta nu este livrat în același timp la toți cei implicați, înseamnă că putem să avem stări intermediare, deci se pierde izolarea datelor în timpul tranzacției.

O tranzacție implementată în acest stil are foarte multe probleme. În primul rând, garantarea faptului că un microserviciu poate să își execute partea sa din tranzacție se face de cele mai multe ori prin blocarea coloanei sau tabelului ce trebuie să fie modificat. Acesta introduce probleme atunci când avem un sistem destul de mare, întrucât trebuie să tratăm cazurile în care alte tranzacții vor să se execute dar nu pot. Dacă tranzacția aceasta este de lungă durată, atunci această blocare poate să afecteze pe termen lung. După apar probleme de tipul, ce se întâmplă dacă un microserviciu spune că poate să execute commit-ul, însă după nu primim răspuns dacă s-a executat sau nu, asta ar aduce multă complexitate pentru a anula tranzacția și e greu să garantăm că nu ar exista pierderi de date. Din cauza complexității și a multor probleme de acest tip, e bine să fi evitate. În general putem să facem asta grupând elementele care ar avea nevoie de o tranzacție distribuită și să le punem într-un singur microserviciu.

Chiar și în aplicațiile monolitice apar tranzacții care se desfășoară pe o perioadă mai lungă de timp, numite și long-lived transactions, acestea cauzează probleme întrucât pot bloca tabele pe o perioadă lungă de timp. În „Sagas”[6] se descrie un mod de prevenire a acestor tipuri de tranzacții prin separarea unei singure tranzacții în mai multe astfel încât acestea să poată să fie realizate separat pentru a evita blocările. Deși inițial a fost formulat pentru baze de date normale, se aplică foarte bine pentru microservici, însă se pierde atomicitatea generală a tranzacției, chiar dacă fiecare componentă este în continuare „acidă”.

Atunci când facem o tranzacție prin acest mod, se fac mai multe etape în funcție de câte microservicii sunt implicate, fiecare sub tranzacție fiind o etapă. Atunci când o operație eșuează într-o tranzacție unică, se anulează tranzacție. În acest mod de operare fiecare etapă care a rulat trebuie să fie anulată, acest lucru se face lansând operații de

anulare pentru toate etapele care au fost efectuată.

Acest mod de operație ne permite să tratăm erorile care apar din cauza erorilor care apar din motive condiționate de date și nu tehnic. Există două tipuri de anulare, cele care încearcă să anuleze tranzacția și cele care încearcă să compenseze tranzacția. De exemplu atunci când încercăm să anulăm tranzacția lansăm operații de anulare însă acestea nu pot fi identice ca cele care le-au creat pentru că aceste operații pot conține pași care nu pot fi anulați (de exemplu trimiterea unui mail), în cazul acesta operația de anulare ar trebui să fie o înștiințare că mail-ul a fost o greșeală, astfel acest tip de anulare este doar semantic. Compensarea unei tranzacții apare atunci când apare o eroare însă în loc să anulăm tranzacția, încercăm să continuăm cu ea însă în mod compensator pentru eroare care apare.

În general, anularea unei tranzacții de lungă durată este complex întrucât pot să existe cazuri în care nu se poate ajunge înapoi la starea inițială, deci pot exista tentative de anulare, de exemplu mutarea pașilor care sunt cei mai posibili să eșuează sau cei care ar fi mult mai greu de recuperat la început, astfel operațiile compensatoare sunt mai ușor de executat.

Implementarea unor astfel de tranzacții se poate face în mai multe moduri. Saga-urile orchestrare apar prin crearea unui microserviciu central care va trimite comenzile ce implementează etapele din cadrul tranzacției asupra căruia are control total. Acest mod de implementare duce la crearea unui cuplaj între aceste microservicii și are la bază comunicarea bazată pe cereri-răspunsuri ce în general face ca tranzacțiile să fie aproape secvențial. Saga-urile coregrafiate au la centru paralelismul arhitecturii, astfel nu mai există comenzi trimise de un microserviciu central ci acestea emit și reacționează la evenimentele pe care fiecare etapă le crează. Această abordare ne permite ca două microservicii să consume același eveniment însă problemele apar la partea de monitorizare, datorită naturii distribuite nu mai avem implicit o metodă să vedem ce se întâmplă, astfel putem crea un microserviciu ce ascultă toate evenimentele și asociem un ID de corelare folosit pentru a grupa acțiunile din cadrul unei tranzacții.

Ambele abordări ale saga-ului dar și tranzacțiile distribuite în două faze aduc complexitate sistemului nostru, cel mai bine e să fie evitate dar dacă este nevoie atunci ar trebui să o alegem pe cea care se pliază cel mai bine cu modul în care operația ar avea sens, iar dacă ar trebui să alegem modul de derulare atunci ar trebui să încercăm să fie cât mai asemănător cu stilul pe care vrem să îl implementăm și în cod.

3.4 Build

Partea de construire a executabilului este o parte importantă din dezvoltare. Pentru limbajele compilate este un moment în care putem să vedem dacă avem erori ce în general sunt destul de minore însă sunt prinse încă de la începutul rulării ceea ce ne permite să

le rezolvăm rapid, având feedback aproape instant de la compiler.

În aplicațiile monolitice, crearea executabilului este destul de ușoară întrucât există un singur produs ce se află într-un singur loc ce trebuie să fie lansat, astfel avem un număr de executabil minim. În contextul microserviciilor, fiecare microserviciu are „executabilul” propriu, însă lucrurile se complică atunci când luăm în considerare că fiecare microserviciu are și o bază de date locală pentru el, astfel aceasta trebuie să vină la pachet și să fie configurată și executată.

Un concept devenit standard este integrarea continuă („Continuous Integration”) - CI, aceasta presupune ca atunci când lucrăm într-o echipă, codul cu care lucrăm să fie mereu merge-uit într-un loc central pe care se rulează teste automate pentru a verifica că nu există probleme la integrare, iar dacă există atunci acestea ar trebui să fie principala prioritate. Acest concept își are baza în Agile și partea în care vrem ca mereu produsul să fie într-o stare în care poate fi livrat. CI-ul presupune să avem artefacte create din codul ce este merge-uit, astfel beneficiem de un istoric al build-urilor ce ne permite să testăm pe versiuni anterioare în cazul în care apar probleme.

Necesarul pentru un sistem de CI eficient este să integrăm codul frecvent, ideal zilnic, pentru a putea valida că codul rulează bine împreună, acest lucru făcându-se prin teste automate pentru a asigura că modificările aduse nu aduc schimbări de comportament, fiind validat prin teste dar și corectarea codului ce face ca testele să nu funcționează, acest lucru este necesar întrucât cu cât mai multe modificări aducem unui produs într-o stare stricată, va fi mai dificil de depanat.

Un mod de dezvoltare a caracteristicilor noi este de a crea branch-uri separate pentru fiecare. Dacă cerința pe care încercăm să o dezvoltăm necesită mult timp, asta înseamnă că nu o să integrăm codul decât la finalizare, ceea ce nu este în concordanță cu principiul integrării frecvente. Astfel, e mai indicat să fie un singur branch în care cei care dezvoltă adaugă schimbări frecvent ceea ce evită problemele de integrare, frecvent numit și „trunk-based development”, însă nu este perfect întrucât modificările făcute de alții sunt mai puțin vizibile, întrucât pull-request-urile sunt făcute ca o atenționare asupra modificărilor care vor apărea în cod, din acest motiv problema cea mai mare este să așteptăm primirea acceptului la pull-request-urile lansate.

Integrarea continuă se face prin intermediul pipeline-urilor care sunt configurate prin intermediul anumitor unelte care sunt făcute special ca să îndeplinească aceste cerințe, în cadrul lor se rulează mai multe etape în care se face o parte din procesul de construire al artefactului. Acestea sunt definite cu ajutorul aplicației de CI. Cel mai important pas este cel în care se rulează testele. În general este bine ca testele care rulează repede să fie executate la început întrucât cu cât feedback-ul pipeline-ului este mai rapid, cu atât fixarea lui va fi mai rapidă. În cadrul pipeline-ului se crează un artefact asupra căruia se vor executa restul pașilor.

Un alt aspect de considerat este modul de organizare al codului sursă, când avem o

aplicație monolitică aceasta nu are motiv să fie despărțită în mai multe repo-uri, însă în cadrul microserviciilor avem opțiuni.

Cea mai simplă alegere este să nu facem nimic special și doar să punem toate microserviciile într-un singur director, iar la final avem un singur build. Din punctul de vedere al unui operator, acesta trebuie să lanseze cod și să configureze un singur loc, iar un dezvoltator indiferent de cât de multe microservicii ar schimba, acesta face un singur commit. Însă o astfel de arhitectură face lucrurile simple la început dar complicate la final. Având un singur build, timpul pentru feedback crește întrucât se pot rula teste și pentru microservicii care nu sunt schimbate. De asemenea, nu știm ce microservicii ar trebui să fie lansate și care nu întrucât la final ce ne face să lansăm totul în același timp. Astfel, putem să începem cu o astfel de organizare însă dacă vrem să avem un sistem de succes, eventual va trebui să avem o altă abordare.

Într-o arhitectură distribuită formată din mai multe programe, probabil primul gând ar fi să avem o sursă pentru fiecare microserviciu, numit și multirepo este un mod de organizare în care fiecare dintre acesta are spațiul și pipeline-ul propriu, astfel se produce o separare foarte ușoară între organizare și se cunosc echipele ce lucrează la fiecare microserviciu. Însă simplitatea acestei separări aduce probleme în alte părți. Reutilizarea codului devine mai complicată, întrucât fiecare microserviciu va depinde de librării specializate, dacă microserviciul de la care pornește această librărie face o extensie atunci pentru a lansa caracteristicile noi trebuie să așteptăm ca microserviciile ce implementează librărie să aducă o versiune actualizată. Arhitecturile distribuite contribuie la un tot unitar, deci uneori o modificare trebuie să fie făcută în cadrul mai multor microservicii, în acest caz trebuie să avem acces la ambele repo-uri, să facem commit-urile în ambele și după să actualizăm și UI-ul ce consumă aceste microservicii. Dificultatea modificărilor crește mai ales dacă e nevoie să fie implicați mai mulți oameni, însă dacă avem astfel de probleme e posibil ca modelul de microservicii să fie prea cuplat. Folosirea unei astfel de organizări merge suficient de bine indiferent de nivelul la care ne aflăm însă dacă facem modificări des în mai multe locuri probabil ar trebui să facem câteva modificări la modul de desfășurare.

Un alt mod de organizare este de a avea tot codul sursă într-un singur loc. Acesta sună asemănător cu prima variantă prezentată însă aceasta este configurată și se cunoaște faptul că în cadrul ei sunt mai multe proiecte. În primul rând, serverele de integrare continuă ar începe să diferențeze între proiecte în funcție de poziția lor în director, astfel nu ar mai exista probleme cu build-urile de lungă durată întrucât s-ar face per proiect modificat. Un avantaj al acestui mod este faptul că reutilizarea codului este mult mai ușoară întrucât fișierele s-ar afla într-un loc foarte apropiat, singura dificultate ar fi să includem aceste fișiere în serverele ce crează build-urile. Totuși separarea între echipe nu este la fel de evidentă, pe GitHub există un tip special de fișier numit CODEOWNERS în care putem să specificăm pentru fiecare folder cine deține elementele, acest lucru ajută întrucât cei

menționați vor fi adăugați automat la recenzie atunci când se face pull request. Acest tip de pattern are dezavantajul că ar consuma extrem de multă memorie, în același sens ar fi destul de greu de descărcat mai ales când nu avem nevoie. În acest scop companiile mari își creează unelte proprii pentru a rezolva aceste probleme, precum Google ce folosește unelte de versionare a codului propriu, Piper, astfel acest tip arhitectural funcționează foarte bine pentru companiile mari ce pot investi în îmbunătățirea acestuia dar și pentru echipe foarte mici în care separarea între proiecte este ușor de făcut.

3.5 Testare

Testarea este o parte importantă în dezvoltarea produsului, avantajele creării de produse sustenabile ce nu conțin defecte sunt destul de evidente întrucât clienții pot aprecia faptul că aplicația nu se strică și pot construi o reputație în acest sens. Pe de altă parte, apariția unei probleme majore în produs poate fi extrem de costisitoare în funcție de momentul în care aceasta este găsită, astfel este important ca în procesul creării defectele să fie prinse și rezolvate din timp.

Testarea poate să aibă o reputație scăzută fiind asociată ca ceva repetitiv ce trebuie făcut manual însă aceasta a avansat și se face în diferite moduri inclusiv automat astfel poate să fie la același nivel cu dezvoltarea însă necesitând un alt set de abilități și un mod de gândire diferit.

Însă acest tip de testare se referă la oamenii dedicați pentru asigurarea calității produsului, deci echipa de QA. Însă ei nu sunt singurii cu responsabilitatea de a monitoriza calitatea sistemului, acesta pornește direct de la dezvoltator. În general, un tester poate să nu aibă acces sau să nu se uite la cod, numit și black-box testing, în timp ce primind acces se numește white-box testing iar atunci când avem acces parțial, gray-box testing.

Testele pot urmări lucruri diferite, astfel conform cărții „Agile Testing” de Lisa Crispin și Janet Gregory putem împărți testele în funcție de ce încearcă să descopere. Testele de acceptare testează dacă funcționalitatea este conform specificațiilor. Testarea exploratorie se face în general manual și încearcă să verifice dacă există probleme în implementare. Testele unitare sunt automate și verifică dacă nu s-a stricat funcționalitatea funcțiilor în urma unor modificări noi. Testele de proprietate verifică performanța sistemului și capacitatea acestuia, se face prin folosirea de unelte specializate și urmărește timpul de răspuns, scalabilitatea sau reziliența.

Putem grupa aceste categorii și să obținem mai multe clasificări, testele de afaceri ajută alte persoane să înțeleagă calitatea produsului (TU, TE) și testează o parte mai largă din produs, testele ce vizează tehnologia sunt mai rapide și au componente care sunt automatizate (TU, TP). Pentru a observa calitatea produselor se pot folosi teste care au ca scop asistența în dezvoltarea pentru a nu apărea regresii (TU, TA) dar și unele care critică produsul în funcție de performanță sau adaptibilitate (TE, TP). În general

testele se fac în timpul dezvoltării, însă există teste care se fac pe produsul care este în piață.

Testele de serviciu au ca scop testarea unui microserviciu, acestea evită testarea UI-ului ci doar a componentelor folosite de UI ca să aducă funcționalitate utilizatorului. Limitând testele la un singur microserviciu, testăm capacitatea acestuia iar orice comunicare cu exteriorul este mimată, astfel cauzele de eșec pot să fie strict legate de microserviciul testat. Aceste tipuri de teste, întrucât nu testează componenta grafică pot să fie la fel de rapide ca testele unitare însă întrucât pot intra în componență apeluri la baza de date sau la rețea. Astfel în implementarea acestora există necesitatea înlocuirii comunicării cu exteriorul și se face în general fie prin mimarea datelor sau prin înlocuirea acestora. Mimicarea se referă că la orice tip de apel al unei funcții, indiferent de datele de intrare, timpul la care a fost făcut sau numărul de ori de câte este apelată acesta va returna un singur set de date. Înlocuirea se face prin replicarea funcționalității astfel putem apela funcții care procesează datele de intrare și returnează date în funcție de acestea. Putem să dezvoltăm această idee și să încercăm să creăm servere de test care furnizează această înlocuire existând unelte pentru aceasta, de exemplu mountebank ce poate să mimeze date trimise prin HTTP/S, TCP sau SMTP însă nu poate să mimeze comunicarea prin evenimente.

Însă chiar dacă facem aceste teste, nu putem să garantăm că putem livra microserviciul. Testele end-to-end se fac prin mimarea utilizatorului și a diferitelor comportamente pe care acesta le-ar face, astfel verificând funcționalitatea sistemului prin interfața cu care el comunică. Acestea pot să implice multiple microservicii ce trebuie să nu aibă probleme, însă având o arie mai mare de acoperire acestea sunt mai greu de testat iar atunci când pică nu se poate observa din prima posibilă cauză. Una dintre problemele care apar în aceste teste este nevoia de a rula mai multe microservicii în același timp, astfel pentru fiecare din aceste microservicii s-ar rula aceleași teste. Acesta ar duce la o repetare de teste care nu s-au schimbat, astfel o metodă de a eficientiza procesul este de a combina testele end-to-end pentru mai multe microservicii în același timp.

Aceste tipuri de teste acoperă o arie extinsă de cod iar fiecare dintre acestea poate să aibă probleme, astfel pot exista în suita de teste unele teste care uneori pică deși uneori trec, cauza unor astfel de situații poate să fie probleme temporare de rețea ce cauzează un timeout. În astfel de cazuri ar trebui ca acest test să fie fixat, fie eliminat întrucât existența unui test neconcluziv poate duce la ignorarea acelui test, astfel este ca și cum acesta nu ar rula de la bun început.

Întrucât testele end-to-end acoperă mai multe microservicii apare problema asocierii responsabilității scrierii acestora. Întrucât în același timp mai multe echipe sunt implicate, acestea toate ar avea o parte de responsabilitate însă scrierea unui astfel de test include cunoașterea sistemului întreg și observarea locurilor ce pot fi mai sensibile, astfel companiile mari pot investi în crearea de echipe dedicate ce au ca rol scrierea, rularea

și rezolvarea (când este cazul) a testelor end-to-end. Însă crearea unor teste end-to-end complexe poate duce la timpuri mari de executare ale acestora, astfel feedback-ul compatibilității întârzie iar în acest timp se pot aduna mai multe caracteristici adăugate de dezvoltatori, astfel se încalcă regula menținerii o prioritate a rezolvarilor problemelor ce sunt asociate pipeline-urilor, în acest caz putem analiza suita de teste și să păstrăm doar esențialele, mai ales când un build poate să pornească suite de teste end-to-end diferite întrucât acesta este prezent în mai multe procese care se desfășoară în cadrul mai multor microservicii.

Problemele cu testele end-to-end duc la crearea unor probleme asociate cu arhitectura paralelă, întrucât acestea ar trebui să ruleze separat, atunci ar putea să fie și testabile independent, astfel putem rula teste end-to-end mai rar dacă avem diferite caracteristici care să ajute la interoperabilitatea precum interfețele explicite sau prin folosirea unor unelte precum Pact putem să creăm o schemă a comunicațiilor exterioare ce include așteptările pe care le avem de la aceștia. Acestea pot fi rulate după de dezvoltatorii microserviciului și să cunoască dacă apar probleme de compatibilitate. Acestea sunt folosite întrucât nu trebuie să mai rulăm mai multe microservicii în același timp ci rulează ca testele unitare sau cele de serviciu în cadrul unui singur microserviciu, astfel devin mult mai rapide.

Majoritatea testelor rulează în timpul perioadei de dezvoltare, însă putem rula teste și pe produsul final. Înainte de lansare se poate lansa o suită de teste numită Smoke, ce are ca scop verificarea funcțiilor de bază, de multe ori aceasta ar trebui să ruleze înainte de testarea manuală întrucât căderea acestuia ar însemna lipsa funcționalităților de care am avea nevoie în mod normal, însă ca o ultimă măsură putem să îl rulăm alături de un test de Sanity ce conține o gamă mai largă de teste înainte unei posibile lansări. De asemenea, alte teste în producție pot fi cele de A/B sau feature toggles ce includ lansarea unei funcționalități în două variante pentru a fi mai ușor de decis care dintre variante putem să o alegem (un caz recent a fost unul făcut de eMAG în care au testat performanța vânzărilor încercând să ascundă recenzile asupra acestora) sau Canary releases în care utilizatorii se pot înregistra pentru a primi varianta mai nouă a produsului, aceștia sunt limitați iar în cazul în care apar erori, doar o parte mică din utilizatori a fost afectată. O rulare în paralel se referă la lansarea unei funcționalități și păstrarea versiunii anterioare, orice cerere este trecută prin alte servicii și sunt evaluate anumite metrici. Ingineria haosului se referă la adăugarea de greșeli intenționat în produs și testarea acestuia dacă poate să își revină. Acestea pot fi considerate și metode de lansări de produs. Alte tipuri de teste pot fi cele care simulează un utilizator normal în timp ce produsul este în piață pentru a verifica dacă funcționalitățile rămân la fel și nu există probleme de configurări ale mediilor de lucru.

În cazul apariției a unei probleme în producție, unele companii rulează simulări și rapoarte a modului de lucru în cazul apariției unei probleme. În cazul acestora se pot defini

concepte precum timpul mediu de reparare(MTTR) sau timpul mediu între erori(MTBF). Prima dintre ele se poate rezolva ușor prin crearea unei restaurări a versiunii anterioare iar a doua prin îmbunătățirea capacității sistemului de a rezista la erori sau de a le trata. Aceste două metrice sunt la fel de importante ca și testarea întrucât ambele au același scop, asigurare calității unui produs.

Majoritatea testelor rulate vor avea ca scop testarea caracteristicilor și validarea că acestea funcționează în parametrii normali, însă datorită volatilității traficului pot apărea scenarii complexe în care pot apărea un influx de trafic, astfel există dorința de a testa după funcționalitate, adică aspecte precum performanța sistemului. În cadrul unei arhitecturi paralele aceasta poate cauza probleme întrucât într-o aplicație monolitică comunicarea se face într-un singur loc, însă datorită necesității a comunicării colective pot apărea momente în care sunt necesare crearea mai multor cereri, acestea pot cazuri în care pentru microservicii suprasolicitate să creze creșterea latenței sau chiar acestea să nu mai poată fi accesibile. Astfel pot exista mai multe teste de performanță, cele de duranță ce testează sistemul pe o perioadă lungă de trafic dar în parametrii normali, de stres în care testăm sistemul în afara parametrilor normali, ale spike-urilor ce sunt creșteri mari în trafic pe perioade scurte de timp și de scalabilitate ce observă capacitatea sistemului de a servi atunci când numărul de utilizatori crește. Toate acestea ar trebui să se facă cunoscând funcționalitatea normală a sistemului pe trafic real, în absența acestei condiții putem să ne concentrăm pe direcții pe care utilizatorul nu le ia. Într-o arhitectură paralelă, la fel ca în securitate, capacitatea acestuia de a rula eficient este dată de capacitatea celei mai încete piese astfel putem crea scenarii de test în care testăm lipsa unor componente ce au ca scop creșterea rezilienței în cazul erorilor.

3.6 Deployment

Ciclul de viață al unei aplicații monolitice continuă să fie simplă și în contextul lansării acesteia. În urma trecerii prin toate fazele pipe-ului de build și posibila validare manuală acesta poate fi trimis către client sau urcat din diverse platforme de hostare, posibil să existe o perioadă în care platforma nu ar putea să fie accesibilă datorită necesității de a reporni, dacă nu folosim un proces mai amplu de lansare care ar acoperi aceste goluri. Dacă este implicată și o bază de date atunci aceasta poate să fie lansate înaintea aplicației și totul ar funcționa, dacă este invers atunci probabil aplicația tratează cazurile în care aceasta nu este disponibilă.

În cazul microserviciilor, procesul este mai complex întrucât caracteristica specifică acestora este capacitatea de a fi lansate și a funcționa independent de celelalte, chiar dacă pot fi uneori limitate de absența celorlalte, astfel procesul menționat anterior se repetă pentru fiecare microserviciu. Totuși, fiecare microserviciu poate să fie de o tehnologie diferită, poate ca baza de date a fiecăruia să aibă un anumit tip și anumite cerințe de

lansare, astfel flexibilitatea oferită de acestea se poate resimți atunci când este timpul lansării.

Până acum, toate conexiunile între microservicii erau făcute doar logic, nu cunoșteam detalii despre modul cum acestea vor fi aranjate sau unde vor fi lansate sau sub ce formă. O caracteristică a microserviciilor este capacitatea de a scala un anumit serviciu ce simte o cerere mai mare, în acest context putem avea mai multe instanțe al aceluiași microserviciu, la fel cum am face și cu o aplicație monolitică în cazul în care am vrea să servim mai mulți clienți, mai lansăm o copie a aplicației. În acest caz, comunicarea nu se va mai putea face direct către microservicii întrucât scalarea se face automat, ci este nevoie să comunicăm cu un intermediar ce ar face rutarea către acestea care are și scop să echilibreze traficul între cele două, astfel folosim un load balancer și vom comunica direct cu acesta și nu cu microserviciul propriu-zis. În cazul comunicării cu evenimente, trebuie doar să ne asigurăm că evenimentul este consumat de un singur microserviciu iar în urma se va lansa evenimentul răspuns.

Un factor important însă care a fost ignorat în lansarea microserviciilor, este baza de date. Întrucât fiecare microserviciu trebuie să aibă propria bază de date pentru a-și ascunde starea internă și să poată să aibă control asupra tuturor operațiilor ce ar avea loc pe domeniul său, microserviciul trebuie să fie lansat împreună cu aceasta independent. Însă în momentul în care avem mai multe microservicii de același tip, adică mai multe replici, baza de date va fi defapt comună pentru toate microserviciile, adică toate replicile vor folosi ca sursă de date același loc. În acest context apar probleme cu scalarea, întrucât e posibil ca serviciul nostru să scaleze infinit, baza de date nu se va scala, astfel poate să devină un impediment în performanța sistemului, dacă nu cumva își cauzează sieși atacuri de tipul Denial of Service prin existența prea multor cereri. Însă scalarea bazelor de date este complicată, mai ales în cazul în care acestea sunt relaționale, întrucât necesită fragmentarea datelor ceea ce este greu de calculat, din acest motiv putem construi organizări precum existența unei singure baze de date pentru scriere și multiple baze de date pentru citire, în acest context putem separa traficul însă datele pe care le servim pot să nu fie actualizate. Modul de lansare al bazei de date poate să difere de cel al microserviciilor, mai ales pentru companiile existente ce au deja date, migrarea devine mult mai dificilă, din acest motiv aceștia pot preferă să folosească infrastructura existentă în loc să migreze către cloud unde ar putea avea o bază de date per microserviciu, astfel microserviciile pot avea baze de date logic diferite însă care rulează de pe aceleași server.

Atunci când se rulează un microserviciu se crează un mediu lucru pentru acesta. Când acesta este în dezvoltare, programatorul are drepturi depline la locul în care acesta rulează și cu ce alte microservicii comunică, însă atunci când ajunge în producție microserviciul are alt mediu ceea ce face creșterea importanței creării unui mediu de testare cât mai asemănător cu cel pe care va rula. Un exemplu simplu sunt datele de test, întrucât acestea pot avea o anumită distribuție atunci când sunt în producție și alta atunci când generăm

aleator date, din acest motiv timpul de răspuns al bazei de date și alegerea parametrilor de indexare nu poate fi ușor testată fără să interacționăm direct cu mediul de producție, iar existența anumitor legi nu ne permite să copiem datele utilizatorilor în mediul de test cu ușurință. Astfel lansarea microserviciului trebuie să fie ușor de adaptat pentru fiecare mediu, fie el prin anumite variabile ce ar lua valorile mediului sau comportamente diferite în funcție de numărul de instanțe ale microserviciului în fiecare loc.

Lansarea și mentenanța microserviciilor este o sarcină ce necesită coordonarea mai multor echipe, în general cei care se ocupă cu dezvoltarea nu o să se implice în configurarea infrastructurii, din acest motiv există câteva principii pe care le putem urmări atunci când lucrăm cu microservicii pentru a maximiza caracteristicile pe care acestea le oferă.

Întrucât microserviciile sunt implementate pentru a rula independent, ar fi o idee bună ca acestea să ruleze independent. În general, dificultățile în crearea infrastructurii apar atunci când sistemele de calcul trebuie să fie configurate, astfel o metodă ar fi lansarea microserviciilor pe aceeași mașină, în acest caz monitorizarea sistemului scade întrucât metricile asupra impactului peste componente nu ar putea fi atribuit unui anumit microserviciu, în același stil acestea nu ar putea să fie la fel de scalabile și ar fi constrânse la fel ca o aplicație monolitică. Recent, acesta a devenit mai ușor prin lansarea în container sau prin platforme precum Heroku ce izolează sistemele by default. Există mai multe tipuri de izolare în funcție de modul în care lansăm aplicația din acest motiv, cu cât izolarea este mai puternică cu atât cresc costurile asociate acelui microserviciu și dificultățile în administrare.

Toate activitățile încep manual însă în timp ce le facem putem să găsim metode de a le eficientiza, în general prin automatizare. Atunci când încercăm o arhitectură bazată pe microservicii, crește complexitatea lansării, astfel unele configurări care într-o aplicație monolitică era ușor de considerat că nu merită să se încerce să se automatizeze, în microservicii ar trebui ca tot ce se poate automatiza să fie automatizat. În general, problemele care apar pot să fie din greșeli umane care nu pot să fie evitate. Putem porni procesul de încercare a automatizării prin adaptarea de tehnologii și procese ce ne pot ajuta în acest scop.

Dezvoltarea cu ajutorul Git-ului ne permite să observăm schimbările din cod foarte ușor, putem să avem un istoric și persoanele implicate iar atunci când se crează o schimbare prin intermediul Pull Request-urilor putem fi anunțați că se dorește modificarea unei porțiuni de cod. În domeniul configurării infrastructurii aceasta se face manual, prin aducerea infrastructurii prin instalarea de computere, crearea unor mașini virtuale sau rularea unui container. Toate acestea implică munca unui operator ce ar trebui să seteze mașinile și ulterior să le configureze prin intrarea manuală prin SSH, comenzile făcute în interior, deși înregistrare nu sunt la fel de ușor vizibile ca și ceva pe Git, astfel a apărut conceptul de infrastructură ca și cod în care infrastructura este configurată prin intermediul unor limbaje ce pot fi interpretate de unelte ce vor configura pentru noi

computerele precum Puppet sau Ansible. Acestea ajută să scriem starea mașinii într-un mod declarativ iar acestea vor rula script-urile de care avem nevoie în interior. Terraform este o unealtă ce ajută în crearea resurselor ce provin din platforme cloud precum Azure sau AWS. Având infrastructură ca și cod aceasta poate fi urmarită de toți participanți și în același timp poate fi reaplicată pentru toate resursele ce trebuie să fie configurate în același stil.

Funcționarea independentă ne permite să lansăm produse fără ca alte microservicii să necesite o actualizare, în acest context putem să facem lansări mult mai dese accelerând dezvoltarea, acestea aduc avantajul că putem face o lansare pentru un număr restrâns de caracteristici, astfel dacă apar probleme în producție putem să fixăm mult mai ușor întrucât sunt puține funcționalități schimbate. În acest context trebuie să măsurăm contextul în care sunt afectate celelalte microservicii atunci când facem o lansare, în mod tradițional sistemul ar fi oprit și ar cauza o perturbare pentru consumatori și ulterior al utilizatorilor. Această problemă poate fi diminuată cu tehnici de lansare precum lansarea treptată în care serviciile existente se actualizează una câte una în timp ce consumatorii existenți sunt păstrați iar cei noi redirectionați către cele noi. În cazul acesta apar probleme atunci când între microserviciu și client se crează conexiuni de lungă durată precum un socket întrucât acestea nu pot fi întrerupte fără a cauza dificultăți.

O altă specificație importantă este capacitatea de a putea configura o stare dorită a întreg sistemului și să folosim uneltele potrivite pentru ca acesta să se întâmple. În acest scenariu, platforma folosită va monitoriza sistemul și va aplica comenzile necesare pentru a păstra forma dorită. Printre astfel de unelte se numără Kubernetes, iar unii distribuitori de infrastructură în cloud oferă aceleși servicii. În acest context, întrucât platforma este cea care administrează starea este important să automatizăm modul de lansare, astfel platforma primește doar imaginea și o aplică. O continuare al acestui mecanism este GitOps, la fel ca infrastructura ca și cod care prevedea configurarea sistemelor ca și cod, acest sistem prevede menținerea stării ca și cod și folosirea unor operatori automați care vor aplica comenzile din acel repository. Acest procedeu are aceleși beneficii ca și pentru infrastructură, adică capacitatea de a avea o privire mai bună per ansamblu, un istoric, anunțarea schimbărilor și interogatoriu. În acest mod de lucru configurarea manuală a infrastructurii pe care rulează sistemul nu mai este permisă, iar fiecare schimbare se va face în repo.

Locurile în care putem să lansăm microserviciile vor fi mai detaliate în chapter 4 însă la fel cum avem flexibilitate în implementare, avem flexibilitate și în lansare. Acestea pot include o lansare directă pe mașini direct sau într-o abstractizare precum un container sau ca Function as a Service (FaaS), astfel ar trebui să ne documentăm despre opțiunile pe care le avem, avantajele și dezavantajele fiecăreia și ce suntem confortabili să facem pentru a administra sistemul. Ar trebui să folosim metodele existente de deployment însă dacă acestea au nevoie de modernizare atunci putem face schimbări. Cu ajutorul platfor-

mele existente precum FaaS sau containerizarea aplicației putem să scădem dificultățile în configurarea mediului și să ne concentrăm doar pe rulare. În același mod se pierde controlul asupra modului de configurare întrucât acesta va fi administrat de platformă, astfel dacă folosim servicii precum Heroku, acesta ne va oferi anumiți parametri însă configurarea efectivă a sistemului de calcul va fi abstractizată ceea ce ne permite să ne concentrăm pe alte aspecte ale sistemului.

În cadrul implementării unei lansări eficiente s-au adunat doi termeni, livrare continuă (continuous delivery) și lansare continuă (continuous deployment). Diferența între cele două se crează la starea produsului. Cea din urmă poate fi considerată chiar și o continuare. Livrarea continuă se referă la menținerea produsului într-o stare în care poate fi lansat. Lansarea continuă se referă la automatizarea procesului din care produsul livrabil ajunge la utilizatori, astfel conceptul de intrare a unei caracteristici „în producție” poate fi considerat schimbat, astfel această stare poate fi considerată atunci când aceasta intră în produsul livrabil, deși ea nu a ajuns la consumatori. Lansarea continuă poate fi limitată de platformele pe care lucrăm dar și de modul de organizare, astfel acesta necesită crearea unei automatizări puternice atunci când dezvoltăm produsul.

3.7 Monitorizare

Necesitatea monitorizării nu apare atunci când dezvoltăm microserviciile ci atunci când le lansăm. Atunci când apare o problemă nereproductibilă, singurul lucru ce ne poate descrie modul în care am ajuns în această situație sunt urmele pe care le creăm intenționat, adică fișierele de jurnalizare sau tranzacțiile care s-au desfășurat. Într-o aplicație monilitică acesta are un număr mic de locuri în care poate să cedeze, astfel investigațiile pornesc mereu din același loc. Monitorizarea acestuia este mai simplă întrucât numărul de resurse pe care acesta le apelează este mai mic. În cazul unei arhitecturi distribuite atunci când apare o problemă, aceleași investigații pe care le-am face pe o platformă monolitică o să le facem asupra mai multor componente. Din acest motiv trebuie să găsim metode de monitorizare eficiente ce sunt ușor de accesat și folosit pentru cei implicați

Implementarea unui astfel de sistem în interiorul unui microserviciu depinde de modul în care acesta comunică. În cazurile simple în care avem o singură instanță al unui microserviciu atunci monitorizarea acestuia este simplă întrucât trebuie să colectăm date dintr-un singur loc. Atunci când acest microserviciu este scalat și există mai multe instanțe al acestuia care pot să apară sau să dispară în funcție de load, atunci trebuie implementat un sistem care colectează datele de la acestea și eventual le asociază datele asupra cererilor. Atunci când aceste microservicii comunică și cu exteriorul devine și mai complicat întrucât nu vom ști exact modul în care sistemul va acționa, în acest caz agregarea informațiilor nu este suficientă și avem nevoie de metode de prelucrare eficientă al acestora.

Observabilitatea este capacitatea de a înțelege starea unui sistem în funcție de datele pe

care acesta le comunică, asemănător cu bordul unei mașini ce ar lumina anumite becuțe în funcție de problemele pe care mașina ar putea să le întâmpine în interior. Aceste date furnizate pot ajuta în prevenirea unui eventuale probleme, fie prin scalarea automată a sistemului în cazul în care datele furnizate de acesta indică că este suprasolicitat fie prin identificarea problemelor înainte să apară prin crearea de algoritmi bazați pe inteligența artificială, iar toate acestea ar trebui să se vadă dintr-un loc central pentru a putea examina starea sistemului în același mod pentru toate componentele acestuia.

În acest scop, atunci când creăm un sistem bazat pe microservicii ar trebui să pornim de la construirea unei baze pentru observabilitate, astfel putem avea în vedere anumite caracteristici care trebuie să apară. Una dintre cele mai importante și simple de implementat acțiuni este agregarea log-urilor. Dacă putem crea asta atunci putem să construim un istoric al acțiunilor ce s-au declanșat în cadrul sistemului. Deși există unelte dedicate, toate ar lucra în același stil, microserviciile creează fișiere de jurnalizare local pe care alt componentă le va colecta și le va trimite către componenta de agregare. Este important ca acesta să nu există procesări mari ale acestor linii întrucât cantitatea de informație care apare poate să creeze creșteri mari în consumul de resurse. Fiecare intrare în aceste fișiere ar trebui să fie standardizată cu informațiile de care avem nevoie, de exemplu data, microserviciul, mașina, nivelul de importanță al log-ului și alte date de care am avea nevoie în depanare. Toate aceste date sunt folosite de agregator pentru a crea centralizări informațiilor sau pentru a cunoaște momentele în care ceva grav se întâmplă și să fie nevoie să alertăm personalul. Întrucât microserviciile funcționează pentru rezolvarea unei singuri cereri, este important să cunoaștem cererea pentru care se lucrează, în general se asociază un ID de corelare pe care fiecare microserviciu îl va comunica pentru ca tranzacțiile din cadrul cererii să fie ușor de interogat atunci când acestea vor fi agregate. Deși fiecare înregistrare are o dată, aceasta nu poate să fie sigură iar încrederea în acestea poate fi o capcană. În programare, cel puțin pentru mine, lucrul cu timpul este cel mai dificil, în interfață deoarece există fusuri orare ce trebuie tratate (desigur e ușor să salvăm datele în UTC și doar să le afișăm în mod diferit, însă se adaugă și schimbările create de economisirea de economisirea de lumină), iar în cadrul fișierelor de log ale microserviciilor timpul nu poate fi precis întrucât fiecare are timpul său intern, deși poate să fie sincronizat cu un server de sincronizare Network Timing Protocol (NTP) nu poate să fie exact. Putem folosi unelte precum Fluentd ce trimite date în Elasticsearch cu interfață în Kibana însă acest spațiu este populat cu soluții. Atunci când avem un sistem de acest tip, pot apărea probleme, pe lângă ce a fost menționat, apar costuri cu privire la stocarea datelor dar și necesitatea unei puteri computaționale în creștere pentru a menține index-urile coloanelor din bazele de date.

Un alt aspect ce comunică starea sistemului sunt metricile și telemetria. Datele colectate de acestea pot include procentul de resurse consumate, arhitectura sistemului, modul cum răspund la cereri sau chiar performanța acestora. Aceste metrici sunt în genera

simple întrucât urmăresc stocarea unei singure informații sau acțiuni, însă acompaniate de acestea apar informații suplimentare iar acestea sunt costisitor de depozitat întrucât va trebui să fie interogate. În acest scop putem folosi unelte precum Prometheus sau Graphite.

Colectarea datelor despre un sistem se face într-un singur scop, aprecierea dacă totul funcționează în parametri nominali. În cadrul unei aplicații monolitice, această apreciere era simplă întrucât lucrurile pot să meargă sau să nu meargă. Într-o arhitectură paralelă ce se bazează pe independență această apreciere devine mai dificilă. Dacă un microserviciu rulează, este un pas, însă nu este suficient. Acesta trebuie să fie accesibil pentru toți consumatorii acestuia. Însă dacă funcționarea acestui microserviciu depinde de alt microserviciu care momentan are probleme, putem să spunem că acest microserviciu este într-o stare bună? În aceste cazuri se creează anumiți termeni, primul concept fiind Termeni la nivel de serviciu, „Service level agreement” ce se creează între cei care construiesc un sistem și cei care folosesc acest serviciu, astfel aceștia din urmă pot să își creeze un anumit grad de încredere sau nivelul la care se așteaptă. De exemplu, providerii de infrastructură în cloud pot crea astfel de termeni pentru a garanta funcționalitatea serviciilor oferite de ei, pe perioadele în care acestea nu pot fi accesibile din cauza lor, probabil nu vor exista credite consumate însă probleme ale sistemului nostru tot pot apărea ce pot afecta clienții proprii. Obiectivele la nivel de serviciu („Service-Level Objective”) sunt create pentru a crea o imagine sau un posibil progres în atingerea SLA-ului. Acestea pot include lucruri precum timp stabil sau latență. Pentru a aprecia atingerea SLO-ului putem crea Indicatori la nivel de serviciu („Service-level Indicators”) ce depind de sistemul oferit. În construirea unui sistem putem să alocăm spațiu pentru erori, acesta este scopul SLA-urilor, întrucât nu putem să fim perfecți ar trebui să păstrăm loc și pentru apariția unor erori. Acestea ajută în crearea deciziilor care includ riscuri precum încercare de tehnologii noi.

Apariția erorilor este inevitabilă, din acest motiv trebuie să înțelegem modul în care acestea pot să fie raportate. Există erori al căror apariție poate fi ignorată pe moment, în general la aceste tipuri de erori ne putem aștepta și au o severitate redusă. În interiorul unui sistem trebuie să știm tipurile de erori care impactează funcționalitatea sistemului și trebuie remediate în cel mai scurt timp, în general pentru cazurile acestea un tehnician ar fi alertat prin diferite forme pentru a remedia problema. Întrucât astfel de situații creează un grad mare de disconfort, trebuie să alegem erorile care generează acest tip alerte cu grijă, întrucât dacă sunt prea multe e posibil ca acestea să fie ignorate. Studii create pentru a studia motivele pentru care unele accidente apar pot să concluzioneze că din cauza numărului mare de alerte raportate care nu au gradul corespunzător pot genera pentru cei care le urmăresc o oboseală, din acest motiv aceștia pot să nu urmeze procedurile corespunzătoare. Steven Shorrock menționează în articolul „Alarm Design: From Nuclear Power to WebOps” (Humanistic Systems (blog), October 16, 2015) că motivul alertelor este de a direcționa atenția utilizatorului către aspecte ale operației ce

au nevoie de atenție într-un timp scurt. („The purpose of [alerts] is to direct the user’s attention towards significant aspects of the operation or equipment that require timely attention.”). Din acest motiv ne putem inspira din reguli vizează comportamente ce necesită utilaje complexe. Asociația utilizatorilor de echipament și materiale (Engineering Equipment and Materials Users Association - EEMUA) a creat o descriere pentru crearea alertelor într-un mod corect. Acestea trebuie să fie relevante, să ilustreze o caracteristică importantă a sistemului, unice, pentru a evita repetiția, să ajungă într-un timp util în care acțiunile utilizatorilor mai pot crea un impact, să conțină suficiente informații pentru ca acesta să le poată prioritiza, informația ce provine din alerte să fie clară și ușor de înțeles, să conțină informațiile necesare pentru a începe o acțiune cu scopul de fixare, dacă este posibil să ofere sfaturi pentru rezolvare, și să fie concentrate pe problema pe care o raportează.

Monitorizarea unui sistem este dificilă, însă aceasta poate să fie introdusă treptat însă putem începe cu agregare fișierelor de jurnalizare dar și notarea parametrilor în care rulează mașinile de lucru și raportarea acestuia într-un loc central.

3.8 Securizare

Indiferent de platforma sau tipul dezvoltării pe care îl facem, front-end, back-end sau IoT, securitatea va fi una din problemele care vor fi ridicate atunci când dezvoltăm produsul. Însă aprofundarea acestei părți nu este ușoară întrucât necesită o mentalitate specială în care luăm în considerare un adversar. În general, în timpul dezvoltării construim elemente și le construim într-un mod în care să funcționeze eficient, însă există persoane care vor să abuzeze de vulnearibilitățile ce pot apărea la orice nivel al aplicației.

Din acest motiv, în cadrul unei companii există persoane speciale care se ocupă de securitate. Aceștia cercetează și studiază posibilele vulnerabilități care apar și crează anumite investigații asupra produsele companiilor. În același sens, se pot folosi și unelte speciale în interiorul pipe-ului care au ca scop scanarea fișierelor și raportarea posibilelor vulnerabilități, chiar dacă nu este amănunțit sau o scanare în detaliu, aceasta ar oferi feedback rapid ce poate fi soluționat în același mod.

Prin stilul lor, microserviciile ne oferă posibilitatea de a îmbunătăți securitatea sistemului întrucât fragmentarea duce la ascunderea informației către serviciile din exterior, astfel un atacator ar primi de fiecare dată un minim de informație, însă pe cealaltă parte, microserviciile sunt foarte complexe și necesită mult mai multe setări, pe lângă infrastructură și datele pe care acestea le păstrează, cresc numărul de rețele ce trebuie securizate, numărul de dependențe ce pot proveni de la terți (Eventbus, load-balancers, log tools) cresc și trebuie scanate regulat pentru vulnerabilități, astfel numărul de sisteme ce trebuie să fie luate în considerare atunci când construim metodele de protejare ale sistemului.

La baza securității pot sta anumite gânduri ce au ca scop restricționarea și controlarea

accesului. Anumite elemente le-am menționat până să ajungem aici, principiul ascunderii al informației se referea la limitarea numărului de date pe care le expunem, în cadrul securității ne putem gândii la limitarea privilegiilor, astfel, ar trebui să oferim drepturi fiecărui microserviciu cu care comunicăm numai cât are nevoie pentru a funcționa normal, dacă el folosește doar câteva funcționalități atunci acesta ar trebui să aibă vizibilitate numai asupra lor, motivul fiind, dacă acesta este compromis atunci atacatorul nu poate să acceseze mai mult decât microserviciul avea nevoie. Atunci când construim o soluție de securitate pentru sistemul nostru, este indicat să conțină mai multe nivele și moduri ce oferă protecție. Există atacuri ce s-au produs iar în timpul lor nu a existat vreo metodă de alertare. Astfel, putem să facem distincția între modurile preventive de securizare, acestea au scopul de a preveni apariția unui atac, iar dacă se întâmplă să se limiteze daunele ce pot fi produse, printre acestea includ acțiuni precum securizarea variabilelor secrete, crearea unui sistem de autentificare și criptarea datelor. Putem spori detecția prin setarea aplicațiilor de firewall sau alerte atunci când se accesează anumite date. Pentru a mitiga atacul putem folosi diferite procedee însă cel mai important este comunicarea eficientă între echipe iar ulterior metode de a recupera sistemul. Implementarea automatizării este o metodă de a crește securitatea întrucât forțează gândirea metodelor de acces dar și eliminarea greșelilor umane iar adăugarea unor metode de analizare a securității în pipeline ce pot investiga erorile simple la nivel de cod.

Tot de la bazele securității provine crearea unui model de amenințare. Procesul creării acestuia include mai multe etape ce au ca scop cunoașterea sistemului și identificarea atacatorilor. Prima etapă se referă la identificarea elementelor de valoare din componența sistemului, modul în care acestea ar putea fi abuzate și identificarea posibilor atacatori ce ar putea apărea pentru a cauza probleme în interiorul sistemului. În urma acestuia, putem să ne gândim la modul în care o să protejăm datele și accesul la acestea. Întrucât nu putem preveni inevitabilul de fiecare dată următorul pas este creșterea metodelor de detecție, în general prin folosirea de unelte dedicate la nivel de rețea și infrastructură. Capacitatea de răspuns în fața unui atac nu se referă doar la metodele de oprire ale atacului dar și alte elemente precum comunicarea problemelor către clienți și parteneri. Metodele de recuperare se referă la revenirea la normal și aplicarea elementelor descoperite în cadrul atacului pentru a preveni astfel de incidente.

Întrucât dezvoltarea pornește de la aplicație, de aici putem începe și cu crearea de practici bune pentru a crește securitatea sistemului. În general, o aplicație va avea cel puțin câteva date sensibile ce au de a face cu autentificarea, fie a microserviciului în sine către baze de date sau alte microservicii sau alte utilizatorilor. Întrucât numărul de conexiuni poate fi ridicat, am putea încerca crearea unor date generale și distribuirea acestora însă ar încălca unul din principiile menționate anterior. Microserviciile sunt create pentru a oferi un serviciu către utilizatori, în general aceștia pot să nu fie experimentați și să ignore sfaturi precum folosirea parolelor unice, lungi și administrare de programe speciale,

iar în cazul de față se poate ajunge ca dauna provocată să nu se producă doar în cadrul serviciului nostru ci prin re folosirea datelor chiar și către alte servicii pe care nu le administrăm. Pe lângă sfaturile simple precum securizarea parolele prin folosirea salt-ului și a algoritmilor siguri într-un timp ridicat (pentru a preveni atacurile de tip brute force) și folosirea de 2FA, putem extrage din articolul [8] al lui Troy Hunt în care a prelucrat recomandări de la NIST. Poate exista o limită la dimensiunea parolelor însă restricția ar trebui să fie mai mare de 64 de caractere suportând și caracterele UNICODE nerestricționate de reguli precum o anumită compoziție permițând folosirea de unelte de memorare a unor parole ce nu expiră ce sunt verificate pentru a preveni prezența acestora într-o bază de date de parolă existente și oferirea utilizatorului unei pagini în care își poate observa activitățile și dispozitivele. Pe lângă acestea, pot exista secrete, adică date senzitive care au ca scop funcționarea microserviciului într-un anumit mediu, acestea pot fi perechi de chei asimetrice publice/private dar și de API. Diferența dintre acestea și parolele sunt faptul că noi administrăm modul lor de viață, controlând crearea, distribuirea, stocarea, monitorizarea și schimbările. În acest context, în funcție de modul deployment-ului putem folosi caracteisticile acestora, Kubernetes prezintă o formă simplificată de secret management, acestea sunt memorate în memorie RAM și definite prin perechi chei-valoare iar în interiorul pod-ului sunt accesate sub forma de fișiere montate local, limitarea lor este că orice schimbare a secretului nu va avea efect decât atunci când pod-ul este restartat sau putem folosi platforme specializate precum Hashicorp Vault ce oferă o soluție complexă în acest domeniu. Implementând o metodă de rotație ne permite să generăm chei atunci când avem nevoie pentru o perioadă scurtă de timp, astfel dacă acesta este compromisă atunci putem limita pagubele ce pot fi produse, în același timp revocarea ne ajută în cazul conștientizării unei expunerii iar acestea ar trebui să ofere acces doar la resursele de care avem nevoie.

Securitatea aplicației nu se referă strict la ea ci și la modul cum interacționăm cu ea, adică prin infrastructură. Trebuie să fim la zi cu actualizările sistemului pe care aplicația rulează, dacă folosim infrastructură din cloud probabil acestea sunt făcute deja. Alte servicii pe care operatorii cloud le oferă sunt opțiunile de backup și encryptare a datelor la idle. Backup-urile ne permit să ne întoarcem la versiuni anterioare ale aplicației, astfel nu avem nevoie de backup-ul unei mașini propriu zise ci doar a modului de a ajunge în aceeași stare. Codul și imaginile proceselor sunt deja track-uite prin VCS, folosind infrastructură ca și cod sau GitOps atunci avem și configurarea sistemului, rămânând doar datele utilizatorilor ce în general se fac prin baze de date ce probabil vin cu versiunile proprii de backup în funcție de tehnologia aleasă. Pentru ca back-up-urile să fie eficiente, trebuie să avem încredere că acestea funcționează astfel putem să creăm un proces în care frecvent facem restaurare a backup-ului pentru a fi testat.

Microserviciile funcționează prin comunicarea la distanță prin diferite rețele, astfel se disting două cazuri, întrucât microserviciile sunt lansate într-un loc în care avem control

total și putem identifica orice neregularitate, putem avea încredere deplină că atunci când vine o cerere din interior, aceasta vine de la un serviciu de încredere. În partea opusă apare conceptul de zero încredere, în care toate serviciile care încearcă să ne acceseze sunt deja compromise, în acest caz trebuie să luăm precauții. Acesta este un mod de gândire specific securității, astfel cererile trebuie inspectate iar datele ce sunt comunicate trebuie să fie criptate. Atunci când construim un sistem, ar fi dicil ca acesta sa fie descris ca cel din ultimul caz, iar siguranța ar putea fi afectată dacă ar fi ca în cel din primul, astfel putem să le folosim pe amândouă în funcție de datele pe care le folosim.

Microserviciile vor face mereu schimb de date, fie ele către un singur alt microserviciu, fie către mai multe. Din acest motiv am vrea ca aceste date să nu fie interceptate de către un terț. Dacă comunicarea este făcută prin HTTP putem folosi TLS pentru a securiza traficul, iar dacă comunicarea se face printr-o aplicație de comunicare de evenimente atunci putem consulta documentația platformei și să găsim modalități de securizare. Atunci când dorim să securizăm o comunicare, sunt câteva puncte pe care le urmăim. Una din ele este stabilirea autenticității server-ului, astfel serviciul care face cererea vrea să știe dacă cel către care face cererea este chiar cel căruia vrea să îi facă cererea. Astfel de cazuri au făcut probleme pe Internet, astfel s-a dezvoltat HTTPS pe care îl putem folosi sub diferite forme. În același sens, serverul ar vrea să cunoască dacă clientul este chiar cel care încearcă să facă cererea. Astfel de cazuri nu prea există pe Internet-ul public întrucât serverele sunt accesibile, iar autentificarea clientului nu este o prioritară, mai ales când utilizatorii folosesc lucruri precum combinații de autentificare ce cuprind parole. În cazul microserviciilor putem folosi protocoale precum Mutual TLS pentru a stabili autenticitatea clientului. Implementarea acestuia ar fi fost dificilă în trecut fiind o problemă distribuirea certificatelor însă datorită uneltelor precum Hashicorp Vault acestea devin mai ușoare. După ce participanții la comunicare sunt autentificați aceștia își doresc ca mesajele lor să fie securizate. În acest sens ne interesează dacă mesajul poate să fie văzut de un terț, însă folosind HTTPS acesta va cripta datele folosind cheile publice iar cel care le decriptează va folosi cheia sa privată, în același mod ne interesează dacă datele au fost modificate, pentru asta putem folosi HMAC-ul prin calcularea un hash asupra datelor trimise pe care recipientul le poate calcula pentru și să vadă dacă acestea au fost modificate.

După ce securizăm comunicarea, a doua prioritate este securizarea stocării, pentru asta putem folosi o varietate de unelte, în general majoritatea distribuitorilor de infrastructură cloud oferă modalități de encriptare iar majoritatea bazelor de date oferă același lucru. Însă această criptare poate deveni costisitoare, și din privința procesorului ce ar trebui constant să lucreze dar și a spațiului de stocare necesare, din acest motiv ar trebui să ne limităm datele salvate și să encriptăm orice dată primim și să decriptăm doar atunci când fără aceasta nu am putea să continuăm operațiunile.

O altă problemă a microserviciilor este modul de autentificare și autorizare. Într-o

aplicație monolitică, aceasta nu era o problemă la fel de mare întrucât o dată autentificat toate cererile erau rezolvate de aceeași entitate, însă în cadrul unei arhitecturi decuplate acest lucru devine mai delicat. Prima necesitate de autentificare este cea a microserviciilor, pentru a valida componentele ce participă la comunicare, în acest sens putem folosi cheile pentru mutual TLS sau chei de API pentru a cripta cererea și server-ul ar valida că acel client chiar există. Însă dificultatea mai mare este autentificarea utilizatorilor. Întrucât aceștia folosesc mai multe microservicii, ar trebui să se autentifice către fiecare, iar acestea ar primi informația despre utilizator. Însă acest lucru este un inconvenient pe care vrem să îl evităm. Soluția ar fi crearea unui loc central în care utilizatorul se autentifică iar după cererea este redirecționată către cererea inițială căreia îi lipsea autorizația. Acest lucru este întâlnit la companiile ce oferă mai multe servicii precum Microsoft la care avem Outlook, OneDrive, Teams, acestea au o pagină comună de conectare, iar conectarea la una îți oferă posibilitatea de a te conecta la toate celelalte cu ușurință. Acest sistem ar funcționa prin intermediul unui intermediar care ar căuta dacă utilizatorul este autentificat și dacă cererea pe care încearcă să o facă necesită autentificare și este fie lăsat să își termine cererea fie către pagina de autentificare. O astfel de abordare ar avea probleme întrucât doar prima cererea ar primi informații despre utilizator, iar următoarele ar trebui să aibă încredere că microserviciul cu care comunică nu este compromis. Pentru a rezolva această problemă am putea încerca să facem autentificarea în cadrul fiecărui microserviciu, însă ar trebui să reutilizăm cod și deși am putea folosi librării nu ar fi sustenabil. O metodă de autentificare poate fi folosirea de JSON Web Tokens, aceștia pot fi semnați de către sistemul de autentificare central, pasați către utilizator care de fiecare dată când va face un request își va trimite token-ul, iar autentificarea de la un microserviciu la alt microserviciu s-ar putea face cu ușurință întrucât ar consta doar în trimiterea token-ului. În componența token-ului avem un header (ce conține informații despre algoritmul de semnare), datele și semnătura propriu zisă. Deși JWT-urile sunt ușor de folosit, acestea nu sunt perfecte. O primă problemă este setarea unui timp bun de expirare. Fiecare token are o dată de expirare setată de noi, după trecerea acestui timp, Token-ul ar trebui să fie considerat invalid, astfel pot apărea probleme în tranzacțiile de lungă durată, însă dacă timpul de expirare este mai ridicat atunci dacă este compromis se pot provoca mai multe daune întrucât nu există o metodă ușoară de a invalida un JWT. De asemenea, folosirea JWT-ului include administrarea informațiilor necesare pentru criptare, adică cheile și timpul de expirare ce poate fi greu de sincronizat între toate microserviciile care ar avea nevoie, însă unelte precum Vault ne poate ușura munca în acest context.

3.9 Evoluție

3.9.1 Stabilitate

Într-un proces clasic de dezvoltare a software-ului în stilul Waterfall, după ce aplicația este construită, testată și trece de audit-ul echipei de securitate aceasta va fi lansată și se va ajunge în starea de mentenanță. Însă modern, dezvoltarea aplicației nu se termină aici ci se va dori extinderea, consumul aplicației de către diferite interfețe, scăderea numărului de erori și tratarea acestora dar și eficientizarea sistemului.

Fiind implicată conexiunea prin Internet dar și globalizarea, utilizatorii se așteaptă ca serviciile pe care le oferim să fie disponibile 24/7. O arhitectură distribuită ne permite să creștem această disponibilitate întrucât poate să se întâmple ca un serviciu pe care îl creăm să fie căzut, celelalte fie nu ar fi afectate sau ar rula într-o variantă mai limitată.

Pentru a crește reziliența sistemului putem defini anumite arii în care aceasta acționează. Când aplicația poate să trateze erori pe care le anticipăm, precum tratarea erorilor sau reîncercarea cererilor căzute. Din această cauză putem să construim soluții ce devin mai complicate, de exemplu adoptarea Kubernetes pentru a beneficia de facilitățile acestuia precum menținerea constantă a stării, introducem complexitate și un efort ce poate să creeze mai multe probleme. Capacitatea de a absorbi atunci când se produce o anomalie sporește stabilitatea sistemului, concentrarea pe îmbunătățirea metodelor de prevenire al cazurilor de acest tip ne poate lăsa vulnerabili, astfel putem crea metode de recuperare sau metode alternative de funcționare atunci când apar probleme în sistem. Creșterea rezilienței ar trebui să fie făcută constant și să ne adaptăm la evoluția sistemului.

Deși putem încerca să reducem șansele de apariție ale unei probleme, acestea nu pot fi evitate, iar acestea pot apărea la orice nivel al aplicației inclusiv în afara ei la partea de infrastructură. Însă tratarea tuturor evenimentelor care ar apărea ar fi obositor și nu ar da randament, întrucât pot fi microservicii în sistemul nostru la care ne putem aștepta să cadă și să nu creeze perturbări la fel de mari. Astfel putem să testăm monitorizăm sistemul și să calculăm timpi de răspuns, dacă aceștia nu sunt în parametrii pe care îi dorim putem folosi sisteme de scalare, sau a timpului în care sistemul este căzut și stabilirea dacă aceasta este acceptabilă la fel și cât de multe date putem să pierdem.

Pentru a continua funcționalitatea sistemului în procent cât mai mare, uneori e nevoie să facem compromisuri, astfel dacă cunoaștem că un microserviciu de care suntem dependenți este căzut, putem să îl ascundem, astfel utilizatorul are o experiență mai proastă, însă sistemul încă este disponibil. Metode de menținere a stabilității sistemului pentru astfel de situații pot fi, introducerea de time-out-uri în cereri, astfel clienții nu așteaptă mai mult decât trebuie. Durata după care se introduce un timeout poate fi calculată după rularea testelor și ajustată conform telemetriei. Însă unele erori pot să fie doar temporare, iar o reîncercare probabil ar fixa problema, în acest caz putem introduce un sistem

de reîncercare în funcție de eroare pe care o primim. Aceste două metode de tratare a eșecului totuși sunt costisitoare din punctul de vedere al timpului. Întrucât conexiunile încete și un număr ridicat de reîncercări ar face ca clientul să trebuiască să aștepte foarte mult, chiar și pentru o eroare. În acest caz putem crea un fel de siguranță ce are scopul ca atunci când avem funcționalitatea normală aceasta să continue, iar atunci când avem probleme în sistem, clienții care ar încerca să acceseze resursele afectate ar primi un răspuns cât mai rapid, astfel nu ar cauza frustrări pentru aceștia. Erorile se pot propaga, astfel crearea unei izolări între microservicii ar cauza ca acestea își afectează starea reciproc. În acest sens putem crea separări logice în care comunicare între microservicii se face printr-un intermediar care în absența unuia dintre participanți va salva lucrurile ce trebuie să le trimită și le va trimite atunci când va fi disponibil, sau izolare fizică, astfel să ne asigurăm că bazele de date ale microserviciilor dar și aplicațiile în sine sunt separate între ele, astfel dacă un server cade, nu ar afecta mai multe servicii în același timp. Deși acest lucru este mai costisitor putem să folosim asta și să coordonăm metodele de eșec, astfel dacă știm că funcționalitate unui serviciu este puternic afectată de un altul, atunci acestea pot fi grupate și dacă inevitabilul se întâmplă cele două vor cădea împreună. În același stil putem să creștem numărul de apariții ale unui microserviciu, astfel dacă unul din ele este căzut, avem alte clone ce îi pot lua locul. Platforme precum Kubernetes ne permite să facem acest lucru cu ușurință. În același stil, putem să trimitem mai multe cereri de mai multe ori, acest lucru funcționează dacă cererile create și modul lor de prelucrare este idempotent, adică de oricât de multe ori este aplicat, acesta va da mereu același rezultat oricât de multe rulări am face.

În construirea unui sistem bazat pe microservicii stabil am discutat despre metode de eficientizare și de îmbunătățire a rezilienței în care folosim mai multe baze de date sau multiple instanțe ale unui microserviciu. Există o teoremă ce a fost demonstrată și matematic ce spune că putem avea consistență, disponibilitate (availability) și tolerare a separărilor (partition tolerance) însă dintre acestea putem să alegem în sistemul nostru doar două dintre ele, a treia neputând fi atinsă. Consistența reprezintă ca oricare din instanțele ale unui microserviciu, toate ar oferi același răspuns pentru aceeași cerere. Disponibilitatea reprezintă capacitatea unui microserviciu de a primi un răspuns iar toleranța separării reprezintă capacitatea sistemului de a manevra incapacitatea serviciilor de a comunica între ele.

Renunțând la consistență ar însemna ca atunci când avem mai multe microservicii de același tip ce comunică cu o bază de date, dacă nimerim la noduri diferite am primi răspuns diferit. Acest lucru se poate întâmpla atunci când baza de date este formată din mai multe baze de date, unele destinate doar citirii iar una doar pentru scriere. În acest context, sincronizarea dintre cele două tipuri de baze de date este pe moment întreruptă. În general putem sacrifica un pic din consistență în funcție de importanța datelor cu care lucrăm, însă ne așteptăm ca sincronizarea să se facă la un moment dat, cu cât

aceasta întârzie cu atât va fi mai dificil de executat, iar dacă refuzăm scrieri atunci când sincronizarea este lentă o să menținem consistența ridicată dar disponibilitatea scăzută.

Consistența este greu de atins întrucât necesită mulți pași pentru menținerea ei, iar interogarea datelor se face inițiind tranzacții pentru a ne asigura că datele pe care vrem să le accesăm sunt cu adevărat la fel, iar dacă aceste noduri nu sunt disponibile, atunci pentru a nu sacrifica consistența vom refuza cererea, astfel scăzând disponibilitatea. Crearea unui mod de partajare a datelor între mai multe noduri este dificilă, astfel ar fi bine să folosim soluții deja implementate, Consul oferă astfel de servicii pentru distribuirea elementelor de configurare.

Negarea tolerării separării este imposibilă în cadrul unui sistem distribuit, întrucât aceasta ar însemna să renunțăm la microservicii și să revenim la o arhitectură monolitică în care nu există separare. Astfel singurele tipuri de sisteme pot fi CP sau AP, adică cele care renunță la disponibilitate, respectiv la consistență, însă în lumea reală acesta nu este neapărat aplicabil. Putem în cadrul unui singur sistem să avem mai multe tipuri de alegeri, în funcție de datele și tehnologia cu care lucrăm.

Testarea stabilității unui sistem distribuit este destul de complicat și s-ar face la mai multe nivele. Un termen popular, original provenit de la Netflix, este „chaos engineering” (ingineria haosului) acesta are ca rol testarea rezilienței sistemului. În general, poate fi considerat ca doar rularea unor unelte și interpretarea acestora, ceea ce nu e greșit, însă aceasta are ca scop îmbunătățirea încrederii atunci când apar probleme ce nu pot fi prevăzute, ceea ce se poate întâmpla în orice moment. Netflix se bazează într-o proporție foarte mare de infrastructura ce provine de la AWS, însă la un moment dat aceștia au fost forțați să efectueze lucrări de mentenanță asupra unei întregi zone de serviciu, astfel se putea crea o gaură mare de down-time pentru clienții Netflix (dar și pentru ceilalți ce se bazează pe AWS). Ei au creat un serviciu numit Chaos Monkey ce are ca scop testarea sistemului la apariția unor probleme neprevăzute în cadrul infrastructurii. În cadrul serviciului putem întâlni mai multe stagii precum, Chaos Gorilla ce are ca scop căderea unei zone de disponibilitate, adică a unei regiuni mai mici (Europe Central), Chaos Kong, simulează căderea unei regiuni întregi, de exemplu Europa, Latency Monkey are ca scop creșterea latenței în cererile executate de clienți pentru a simula încetiniri în trafic, Conformity Monkey are ca scop închiderea instanțelor ce nu sunt în conformitate cu practicile cerute de sistem, Doctor Monkey monitorizează starea fiecărei instanțe și oprește sistemele ce nu sunt tratate în timp util, Janitor Monkey observă infrastructura ce nu este utilizată și o oprește după un anumit timp, Security Monkey are ca scop închiderea serverelor ce nu au toate practicile de securitate necesare.

3.9.2 Scalare

În timpul prezentării microserviciilor, un avantaj al acestora a fost mereu flexibilitatea, în același stil am menționat posibilitatea îmbunătățirii sistemului prin replicarea unor anumite funcții ale acestuia, fie pentru administrarea a mai mult trafic, fie pentru îmbunătățirea disponibilității și eliminarea timpilor în care serviciul este inaccesibil. Scalarea poate fi făcută în mai multe moduri, în funcție de necesitățile și de problemele pe care le-am rezolva.

Scalarea verticală se referă la îmbunătățirea sistemului de calcul prin aducerea mai multor resurse de calcul, fie prin îmbunătățirea sistemelor de I/O, adică folosirea unor SSD-uri rapide față de HDD-uri clasice, fie de calcul prin îmbunătățirea procesorului și a memoriei RAM. În cazurile clasice, când noi suntem cei care hostăm aplicația, aceasta poate fi mai de lungă durată întrucât sunt mai multe echipe implicate, însă având în vedere sistemele cloud, scalarea verticală se face foarte rapid și nu necesită configurări adiționale. Însă acest tip de scalare nu este o soluție ce va funcționa la nesfârșit, se observă că legea lui Moore nu se mai aplică, astfel îmbunătățirile tehnologice nu sunt la fel de drastice în fiecare an. Nu ar trebui să ne bazăm că putem schimba mașina de calcul, mai ales că folosirea tehnologiilor foarte puternice ar deveni costisitor. În acest scop, putem folosi algoritmi mai eficienți sau alte tipuri de scalare.

Scalarea orizontală se referă la multiplicarea instanțelor ale unui serviciu, în acest stil putem crește numărul de procesări ce se pot face în același timp. Implementarea cea mai simplă se poate face folosind un balansator de încărcătură (load balancer) ce va dirija traficul asigurând că traficul este egal pe fiecare dintre microservicii. Acest lucru este posibil atunci când putem să distribuim o acțiune către mai multe locuri, însă putem avea situația în care distribuim sarcinile din interiorul unei cereri, astfel o altă implementare ar fi separarea unor cerințe din interiorul unei cereri către mai multe microservicii ce au unic, de procesare, iar scalarea s-ar face asupra acestora. Am mai menționat și scalarea bazei de date prin folosirea de replici de citire ce sunt sincornizate cu baza de date destinată scrierilor. Limitările acestui tip de scalare se rezumă la dificultățile de implementare, astfel putem întâmpina mai multe probleme de rețelistică sau de sincronizare a datelor sau chiar recrearea unui nou mod de desfășurare destinat acestui stil distribuit de lucru.

Partiționarea datelor se referă la distribuirea cererilor către un anumit set de microservicii în funcție de natura cererii. Astfel, putem avea microservicii de același fel, cu funcționalitate identică însă care se ocupă de date diferite. Aceasta poate să fie o îmbunătățire întrucât am avea sisteme de procesare dedicate unor anumite cereri, astfel dacă am întâlnii mai mult trafic într-un anumit loc, am putea planifica o partiționare din nou, iar dacă nu avem atunci să regroupăm anumite date. Implementarea se poate face în funcție de tehnologie, astfel, dacă baza de date pe care noi o folosim oferă asistență în acest loc (de exemplu Cassandra) atunci putem să o facem la nivelul acesteia, sau să

implementăm servicii de proxy ce redirecționează traficul către microserviciile ce ne oferă funcționalitatea de care avem nevoie. Acest tip de scalare este eficient când avem suntem constrânși din punct de vedere al scrierilor, întrucât acestea nu pot fi eficient distribuite în alt mod. Limitările pot fi faptul că nu se aduce o îmbunătățire la fel de mare și în funcție de modul în care ne facem grupările ne poate duce să avem un grup de persoane afectat, astfel putem grupa acest tip de scalare cu cele precedente pentru efectivitate mai bună. În același stil, stabilirea modului de grupare și dimensiunile partițiilor pot fi dificil de stabilit.

Majoritatea formelor de scalare încercu să nu modifice structura aplicației prea mult, însă după ce epuizăm aceste forme, putem încerca extinderea aplicației prin decuplarea funcționalității în alte microsisteme. Beneficiul acestui tip ar fi posibila eficientizare a procesului întrucât s-ar putea parta încărcătura însă ar crește complexitatea sistemului și poate cauza degradări ale performanței întrucât s-ar introduce o cerere pe rețea nouă, astfel trebuie să ne asigurăm că implementarea acestui tip de funcționalitate ar avea sens.

Sunt multe cazuri în care soluțiile pe care le creăm sunt mult mai complicate decât ceea ce ne cere sistemul. Toate implementările de scalare sunt incluse în această frază, astfel nu ar trebui să adăugăm complicate fără să fie dovedit că am avea nevoie. În general, există multe caracteristici introduse de dezvoltatori însă care nu sunt folosite mai niciodată de către utilizator, însă atunci când aceste funcționalități au șansa să strice aplicația ar trebui să fie evitate și aplicate doar la nevoie. Iar în acest context, întrucât microserviciile ne oferă o flexibilitate ridicată putem să aplicăm la un anumit nivel toate tipurile de scalare în loc să alegem doar una din ele.

Există și conceptul de auto scalare, în general ne putem gândii ca sistemul să fie mai puternic accesat într-un anumit interval, astfel putem să creștem instanțele de microservicii în această perioadă. Ofertanții de infrastructură cloud ne oferă diferite metode de autoscalare deja în platformă și sub diferite forme, de exemplu atunci când ajungem la un anumită încărcătură pe un anumit nod pe o anumită perioadă putem să mai creăm o instanță automat. Astfel autoscalarea poate fi reactivă sau predictivă, ambele aducând îmbunătățiri însă trebuie să analizăm datele, ce poate fi făcut cu algoritmi complecși precum cei de Inteligența Artificială / ML.

3.9.3 Caching

Caching este o metodă de a salva informația cerută la un moment dat de către un utilizator și prevenirea recalculării modului de a ajunge la date și doar servirea aceluiasi răspuns. Atunci când putem să oferim același răspund întrucât l-am salvat putem spune că avem un „cache hit” și când nu, un „cache miss”

Există multe motive pentru care am adopta această tactică, întrucât sistemul de caching are ca rost prevenirea recalculărilor inutile, putem îmbunătății performanța sistemului

întrucât unele dintre aceste calcule pot include cereri pe Internet ce adaugă latență. În același stil putem eficientiza traficul dacă avem cachate anumite răspunsuri pe care doar le livrăm de mai multe ori, astfel întrucât nu aduc complexitate pot fi scalate cu ușurință. Deși ar scădea consistența, existența unui sistem de cache ne-ar putea permite să rulăm o anumită perioadă de timp fără ca sursa obținerii datelor să fie disponibilă, în acest stil putem ascunde problemele de puțină durată întrucât avem deja o sursă de date introdusă.

Această tehnică poate fi aplicată la mai multe nivele, la nivelul clientului, putem salva în interfață datele, de exemplu pe telefoane mobile putem face o dată o cerere către baza de date, salvăm rezultatul și acesta va fi disponibil chiar și atunci când nu avem Internet, acest lucru este vizibil în rețelele de socializare care atunci când deschidem aplicația mereu au anumite postări preîncărcate, această interacțiune ne permite să salvăm baterie și să nu creăm prea multe cereri, în același stil conținutul pe care îl afișăm utilizatorului este mai greu de controlat și poate diferi de la unul la altul. Cacharea la nivel de server ne permite să evităm anumite operațiuni de lungă durată pe care ne așteptăm să nu se întâmple foarte des, astfel putem să trimitem un răspuns salvat. Această operațiune ne permite să avem un control mai mare asupra datelor pe care le salvăm, însă clientul tot ar trebui să consume resurse făcând cererea. În acest stil putem să salvăm informații ale unei cereri specifice, iar dacă aceasta este repetată atunci se va trimite răspunsul vechi.

Atunci când vorbim despre salvarea unei informații prin caching, trebuie să luăm în vedere modul în care ștergem această salvare. Dacă nu am luat asta în considerare atunci clientul ar vedea informații destul de vechi ce nu ar fi actualizate. Cel mai simplu mod de implementare este specificarea unui timp la care informația din cache expiră. Intervalul poate fi dificil de ales iar în același timp pot exista cazuri în care datele se actualizează fix după ce clientul le înregistrează ca fiind cele mai recente iar în cazul în care TTL (Time To Live) este setat mare ar putea cauza probleme. În același fel putem eticheta fiecare cerere cu un ID, iar dacă ID-ul nu se schimbă atunci putem trimite un răspuns mult mai mic. Dacă avem o resursă ce este disponibilă la un anumit URI, dacă am face cerere pentru aceeași resursă putem trimite un identificator al datelor pe care le avem iar serverul ne poate trimite fie date noi, fie faptul că datele nu s-au schimbat. O altă metodă de invalidare a cache-ului poate fi folosirea de notificări, astfel serverul va trimite fiecărui client care a cache-uit un anumit mesaj, fie cu date noi fie doar faptul că au fost invalidate. Aceasta implementare ar aduce avantajul că nu ar exista date vechi pentru clienți însă implementarea unui astfel de mecanism este complexă și poate consuma destul de multe resurse. De asemenea, cache-ul poate fi modificat fix după ce resursa se schimbă, o astfel de abordare ar avea sens dacă am face cache-ing pe un server, sau să îl folosim ca un buffer în care scriem mai întâi în cache și după acesta se va sincroniza cu serverul.

Caching-ul este o metodă de optimizare a traficului, astfel ar trebui să fie folosit doar acolo unde am considera că ar crea o îmbunătățire a sistemului, la fel ca la scalare acesta poate fi introdus treptat, deși poate face parte încă din arhitectura originală însă putem

să investim resurse în ceva în care nu ar fi necesar și am aduce complexitate într-un loc în care poate nu e nevoie.

3.10 Organizare

3.10.1 Consum

Atunci când construim un sistem, ne așteptăm să îl dăm în folosință pentru cineva, fie clienți sau ca asistent al unui sistem existent. Însă în general acești clienți pot să nu consume chiar sistemul direct întrucât acesta poate fi complex, ci putem să îi oferim printr-o interfață construită tot de noi (în cazul aplicațiilor, putem considera front-end-ul o astfel de interfață), astfel putem avea ca responsabilitate și crearea unui astfel de lucru.

În modul tradițional de lucru, am avea dezvoltatori grupați în funcție de rolul și platforma pe care lucrează, astfel toți cei responsabili de o anumită componentă mai mare (de exemplu domeniu sau prezentare) ar fi în același loc. Acest tip de organizare ne permite să avem o experiență identică la același nivel, însă comunicările între nivele pot fi afectate, astfel s-ar putea introduce probleme de sincornizare.

O metodă mai eficientă de organizare este gruparea dezvoltatorilor în funcție de locul în care aceștia lucrează, astfel toți cei responsabili la toate nivelele de o anumită componentă a sistemului nostru vor lucra împreună. Acest mod ne permite un control mult mai bun, având metode de schimbări la orice nivel, astfel atunci când avem o inițiativă, aceasta e mai ușor de implementat și nu depindem de echipe exterioare pentru care trebuie să așteptăm o implementare.

Construirea unei interfețe grafice poate devenii dificilă. În cadrul unei companii ce oferă mai multe produse putem fi tentați să creăm o echipă dedicată pentru aplicațiile pe care utilizatorii le-ar folosi, astfel aceasta ar fi consistentă tuturor. Această abordare poate crea probleme atunci când vrem să lansăm caracteristici noi, de aceea o idee mai bună ar fi crearea unei echipe ce va lucra în cadrul mai multor echipe, având posibilitatea să schimbe proiectele.

Organizarea construirii interfețelor grafice poate fi făcută în mai multe moduri. Cea mai simplă variantă presupune existența unei aplicații monolitice, în cadrul acesteia consumăm API-urile propriu zise fără a face vreo optimizare modului cum sunt făcute. Dezavantajul unei astfel de abordări este decuplarea între backend și frontend dezvoltarea caracteristicilor acestora nu poate fi la fel de bine coordonată. În același stil, crearea multor cereri poate să cauzeze probleme legate de memorie, astfel acestea nu sunt efectuate eficient.

O versiune mai avansată bazată pe framework-urile noi de dezvoltare este microfront-end-uri, în cazul acesteia echipele furnizează un UI ce poate fi lansat independent de UI-urile existente. O astfel de abordare le permite echipelor să își creeze UI-ul independent de

celelalte și acestea rămânând să se lanseze aproape independent, dar uneori tot combinat.

Această abordare poate fi implementată în mai multe moduri. Una dintre acestea este decompoziția site-ului pe pagini, fiecare echipă fiind responsabilă de un anumit endpoint de pagini pe site-ul mare. Aceasta le permite să fie independente întrucât livrarea paginilor este independentă de paginile existente pe care server-ul trebuie să le arate.

O abordare mai dificilă de implementat, este separarea unei pagini în funcție de secțiunile pe care vrem ca acestea să le cuprindă, astfel fiecare echipă ar fi responsabilă de o componentă ce apare pe o pagină la un anumit moment. Implementarea unei astfel de soluții ar fi ca fiecare echipă să își construiască elementele ce vor să fie disponibile pe website și crearea unei metode de agregare ale tuturor elementelor create. Din acest motiv, deși elementele sunt separate acestea tot trebuie să fie integrate, iar crearea unui comportament împărțit în cadrul mai multor elemente poate fi dificilă.

Unul din dezavantajele a unui front-end monolitic era numărul mare de cereri ce ar fi trebuit să fie făcute către mai multe microservicii pentru o singură pagină. Pentru a rezolva problema aceasta putem să folosim un intermediar ce are ca rol această agregare a datelor, astfel de pe dispozitiv facem un singur call însă în spate sunt făcute mai multe. Dificultățile implementării ar consta în coordonarea echipelor, întrucât acest intermediar nu ar avea o echipă directă ci toți ar contribui la el. În același stil ar exista dificultăți de tratare al numărului de informații pe care tipuri de dispozitive le-ar folosi dar și de autentificare, întrucât totul devine agregat de un intermediar ce ar trebui să aibă acces la toate resursele pe care le oferim.

Oferirea de asistență pentru diferitele platforme ne-a dus la crearea conceptului de Backend for Frontend, astfel pentru fiecare tip de dispozitiv ce ar avea nevoie să ne acceseze sistemul, am crea un backend specializat ce s-ar ocupa doar de o anumită experiență. Această abordare poate avea sens dacă se fac multe apeluri peste rețea și dacă echipele de front end sunt decuplate de cele de backend. Implementarea poate fi făcută de exemplu folosind GraphQL.

3.10.2 Structură

Nu este ceva nou faptul că aplicațiile mari sunt create de mai multe persoane. Acestea pot fi organizate în mai multe feluri și în același timp, fac parte din mai multe echipe și dintr-o organizație ierarhică în cadrul companiei. Toate acestea au efect asupra modului în care software-ul este dezvoltat.

În „How Do Committees Invent” publicat în Datamation în Aprilie 1968, Melvin Conway afirmă că „Orice organizație ce crează un sistem, inevitabil va crea un design în care structura va copia structura de comunicare a companiei” („Any organization that designs a system (defined more broadly here than just information systems) will inevitably produce a design whose structure is a copy of the organization’s communication

structure.”). Această „lege” practic afirmă că dacă am avea mai multe echipe ce ar trebui să se ocupe de un sistem, atunci acestea își vor repartiza părțile astfel încât să lucreze fiecare, deci se va ajunge asemănător la structura companiei.

Modul de organizare în companie dar și al echipelor influențează evoluția sistemului. Astfel, unul din obiectivele pe care le putem crea este crearea independenței echipelor, acesta duce la crearea unor organizații decuplate ce poate duce la eficientizarea modului de lucru.

Există probleme al căror eficientizare se poate face prin creșterea numărului de persoane, în acest caz există multe exemple atunci când era prezentată regula de trei simplă, precum muncitorii pe șantier, aceștia au anumite cerințe ce pot să se desfășoare în paralel, astfel adăugarea de mai multe persoane ar apropia data finalizării. Însă unele probleme nu se rezolvă în același stil, în absența capacității de paralelizare a task-urilor s-ar putea să înrăutățim situația.

În acest context, alegerea unui număr potrivit de persoane într-o echipă poate fi dificilă. Aceasta trebuie să aibă suficiente persoane în cât să fie independente. În cadrul Amazon, s-a format conceptul de „Two-Pizza Team”, astfel în mod colocvial o echipă ar trebui să poată fi suficient de mare încât să nu existe probleme în împărțirea a două pizza în cadrul lor, și să fie suficient. Comunicarea ar trebui să fie eficientă în cadrul echipei iar între echipe să fie limitată pentru a se preveni co-dependența în cazuri mari pentru că scopul este crearea unor echipe autonome.

Crearea unor echipe autonome duce la crearea unor deținători asupra microserviciilor create de aceștia. Aceasta se numește o deținere puternică asupra acestuia, astfel doar echipa care deține microserviciul poate să dețină ce elemente intră în cod. Acest lucru se face prin intermediul pull request-urilor, avantajul fiind crearea unui microserviciu stabil întrucât singurele modificări aduse sunt evaluate de persoane ce cunosc foarte bine proiectul.

În mod opus, putem avea microservicii ce au scop comun, astfel nu sunt deținute de nimeni și oricine poate aduce modificări acestuia. Acest mod de organizare ne poate aduce posibilitatea să aducem modificări mult mai rapide, însă necesită ca echipele ce lucrează asupra acestuia să se coordoneze mai bine, mai ales când intră în discuție tehnologiile utilizate.

Scopul final al organizării este de a eficientiza procesul de dezvoltare. Pentru a face asta, pot exista situații ce pot necesita acțiuni hibride.

De exemplu, combinând cele două abordări, putem avea un microserviciu deținut de mai multe echipe. Acest lucru se poate întâmpla atunci când coeziunea este ridicată și decuplarea poate cauza mai multe probleme decât ne dorim. Însă având mai multe echipe se pot cauza probleme în modurile de lansare și introducerea schimbărilor ce pot strica anumite părți din alte microservicii, astfel aducem câteva din probleme unei arhitecturi monolitice însă la un nivel mai mic.

Pentru a crea o colaborare mai eficientă asupra codului comun, putem crea un mediu de lucru în stilul open-source intern, astfel toate echipele pot contribui însă există anumite persoane mai implicate în proiect care pot să aprobe schimbările. Acest mod poate avea sens și poate avea destul de eficient pe termen lung însă ar trebui ca proiectul să aibă un anumit nivel de maturitate, astfel încât să existe o anumită arhitectură ce trebuie să fie urmată.

Un alt concept ce poate apărea este posibilitatea creării unui microserviciu dintr-unul existent. Astfel putem avea o bază ce este liberă pentru toți și pentru funcționalitățile specifice se poate crea un synced fork al acestuia, asemănător moștenirii din OOP. Însă adoptarea unor astfel de schimbări trebuie să fie suportată de microserviciile cu care comunică.

Capitolul 4

Orchestrare și infrastructură

4.1 Concepte generale

Modurile în care un microserviciu poate fi lansat este destul de important întrucât decide o bună parte din arhitectura inițială a acestuia, performanța sau capacitățile pot fi eficientizate prin folosirea unei metode de deploy bună. Evoluția modului de a face rost de infrastructură a pornit de la achiziționarea și crearea echipelor special dedicate ce poate devenii extrem de costisitor și necesitând mai mult timp, iar acum se poate face adăugând cardul pe platforme specializate, adăugarea unor metode de configurare și de securitate și putem avea un server într-un timp foarte scurt a cărui mentenanță la nivel fizic devine responsabilitatea distribuitorului. Bill Baker, un inginer de la Microsoft afirmă că la început tratam serverele ca și animale de companie, „le numim, și atunci când devin bolnave, le îngrijești și le aduci înapoi. Acum ne tratăm serverele le tratăm ca o cîrădă. Le numeri iar atunci când se îmbolnăvesc, le împuști.” (You name them and when they get sick, you nurse them back to health. [Now] servers are [treated] like cattle. You number them and when they get sick, you shoot them.). Tehnologiile actuale ne permit să nu avem nevoie să încercăm să reconfigurăm ci dacă putem salva datele de care avem nevoie, atunci este mult mai ușor să reconstruim sistemul de la 0 și să îl aducem în starea în care avem nevoie de el.

Pentru a accelera crearea unui produs, nu este suficient să accelerăm dezvoltarea, ci găsirea unei metode de eficientizare a întregului proces. În trecut, lansarea unui produs era una dintre cele mai mari încetiniri întrucât includea crearea unui tichet și așteptarea ca un specialist să ne contacteze care ar primi programul și l-ar trimite în execuție. Procesul ar fi anevoios întrucât nu am putea să lansăm atunci când am vrea ci doar atunci când ni se acordă timp, iar în funcție de personal aceasta poate dura chiar luni, astfel lansările ar trebui să planificate din timp și să fie solide, să nu existe necesitatea de EFX-uri întrucât ar complica organizarea.

Cunoașterea metodelor de livrare și alegerea celei potrivite este esențială întrucât

în funcție de obiectivele pe care le avem, buget și alte considerente putem ajunge ca deployment-ul să fie automat sau controlat la anumite nivele, astfel livrarea ar putea ajunge nici să nu mai fie un punct dificil.

4.2 Tipuri de deployment

4.2.1 Mașină fizică

Cel mai simplu mod din punct de vedere conceptual, achiziționarea unui sistem de calcul, configurarea acestuia pentru a avea un sistem de operare, crearea unor conexiuni cu rolul de a ne face sistemul accesibil către locurile în care ar avea nevoie, descărcarea programului pe care vrem să îl rulăm și executarea acestuia.

Această metodă are foarte multe constrângeri pe care în lumea microserviciilor sunt greu de tolerat. Întrucât pe această mașină rulăm o singură aplicație, resursele consumate vor fi doar provenite de aici, dacă sistemul nu este utilizat atunci am putea considera că pierdem resurse.

În același stil nu avem metode ușoare de scalare, întrucât putem vedea că hardware-ul ales nu este potrivit pentru aplicație și am dori să le schimbăm, acest lucru este ceva complex ce necesită reconfigurări, migrări și complexități de up-time ce sunt sezibabile.

Întrucât noi suntem cei care furnizăm platforma de calcul, atunci tot noi trebuie să îi facem mentenanță, acestea incluzând actualizări ale sistemului de operare, back-up de date și audit-uri de securitatea, iar acestea pot consuma mult timp.

Deoarece sunt așa de multe limitări cu acest mod, în general preferăm să alegem unul din celelalte, însă acest mod de deployment va fi mereu valabil pentru cazurile în care este nevoie de el.

4.2.2 Mașină virtuală

Întrucât lansarea direct pe hardware poate fi considerată costisitoare, am putea încerca partajarea resurselor disponibile putând să rulăm mai multe microservicii pe aceeași mașină. Dacă am face acest lucru direct, atunci am încălca multe din bunele practici menționate anterior, printre care izolarea rulării și lansării, adăugând complexitate configurațiilor.

Virtualizarea ne permite să eficientizăm resursele prin partajarea acestora între mai mulți utilizatori. Aceasta este făcută prin crearea unor unelte ce încearcă să mimice elementele unui calculator pentru fiecare mașină virtuală, astfel aceasta nu face diferența între ce e fizic și ce e virtual și nu necesită o rescriere a unui sistem de operare pentru a-l face compatibil cu virtualizarea, astfel acestea sunt independente între ele dar și de sistemul gazdă putând să ruleze aproximativ orice.

Fiecare mașină virtuală poate să își configureze spațiul pentru microserviciul deținut, acestea pot rula concomitent, astfel cantitate de resurse care este irosită este minimizată. Aceste mașini virtuale pot fi limitate și le putem extinde în funcție de necesități. Însă crearea mai multor instanțe în cadrul aceleiași mașini nu va extinde posibilitățile fizice ale acestuia și putem să ajungem într-un caz în care pe aceeași mașină rulează prea multe mașini virtuale. În lumea microserviciilor, astfel de grupări ar trebui evitate întrucât dacă mașina fizică ar cădea, atunci tot ce e în interior ar fi căzut, însă dacă ar fi mai grupate am putea crește viteza de răspuns dintre ele.

O altă considerentă este modul în care facem rost de mașini virtuale și modul în care acestea sunt configurate. Furnizorii de infrastructură ne oferă API-uri detaliate, astfel sunt ușor de adoptat cerințelor noastre, de asemenea ne oferă facilități de auto-scalare, back-up și recuperare pe care în mod tradițional ar trebui să le configurăm. În mod clasic, companiile pot folosi platforme precum VM Ware care deși îndeplinesc nevoile din punctul de vedere al API-ului pe care îl oferă poate fi considerat limitat.

4.2.3 Container

Problema mașinilor virtualizate era că atunci când le cream, fiecare dintre acestea devenea o mașină independentă, asta ducea la duplicare de resurse, spațiu și procesare, necesare pentru rularea sistemului de operare și ulterior a aplicației. Containerele încearcă să rezolve această problemă prin partajarea mai multor lucruri, nu doar resursele fizice dar și a nucleului, astfel în cadrul unui container putem rula orice sistem de operare însă acesta rulează ca parte a ceva comun.

Procesele care rulează în container sunt suficient de bine izolate, acestea nu pot fi accesate din exterior și nu pot comunica cu acesta decât dacă le permitem, aceasta se aplică și între ele, sistemul de operare pe care rulează văzându-le ca niște simple procese.

Elementul care a continuat evoluția containerelor este Docker, acesta aduce beneficiile containerelor însă aduce contribuțiile proprii asupra modul acestora de organizare. Acesta ne oferă asistență în rețelistica între containere, aduce conceptul de imagine, astfel putem abstractiza implementarea microserviciului și să folosim ceva generic indiferent de tehnologia folosită și ne permite stocare și reutilizare la fel cum am face cu o librărie. De asemenea ne oferă un API pentru administrarea containerelor indiferent de ce e în interior.

4.2.4 Aplicație în container

Crearea unor multiple containere care funcționează ca un grup poate consuma destul de multe resurse. Construirea unui simplu proiect în Ktor Server, bazat pe JVM îmi ia aproximativ 2 minute, astfel dacă am avea o colecție de programe asemănătoare ce se bazează toate pe JVM ar include rularea a multiplor instanțe de JVM. În acest caz am

încerca să optimizăm această structură și să facem ca să avem un singur container cu scopul de a rula întreaga aplicație.

Deși limităm consumul de resurse, anulăm toate eficientizarea ce ar apărea în urma containerizării. Rulând mai multe aplicații ce ar putea fi rulate independent introducem probleme în ciclul de viață al aplicației, limităm scalarea și izolarea.

4.2.5 Platform as a Service

Pentru a ușura experiența de lansare există servicii ce ne oferă să ne ia codul, să îl compileze și să creeze un mediu în care să se execute. Acestea abstractizează procesul de deployment prin crearea unei legături între cod și platformă, pe Heroku aceasta se face prin conectarea repository-ului sau ca imagine de Docker.

Când nu există probleme ce țin de sistem, nu apare nicio problemă în folosirea unor astfel de soluții, însă dacă avem nevoie de investigarea la un nivel mai adânc atunci am putea să vedem cum suntem constrânși de aceasta.

Platformele acestea ne pot fi de folos pentru a lansa un serviciu rapid însă putem deveni constrânși de rigiditatea și de integrarea lor în sisteme complexe. De asemenea, există mai multe astfel de servicii precum Heroku, Netlify sau Vercel și deși acestea fac același lucru e posibil ca acestea să ofere facilități diferite.

4.2.6 Function as a Service

O tehnologie ce încearcă se revoluționeze modul de deployment este serverless. Unul din principalele dificultăți în lansarea codului este provizionarea cu un server pe care acesta rulează, însă cu tehnologia serverless concentrarea devine asupra scrierii codului iar platforma pe care o lansăm se va ocupa de rularea și scalarea acestora.

Diferențierea între aceste moduri de deployment este necesitatea de a avea un server ce lucrează constant în spate pentru a putea prelucra cereri. Cu serverless, codul rulează doar atunci când e invocat și acest lucru se poate face prin diferite evenimente, fie ele cereri în browser sau fișiere uploadate în funcție de facilitățile pe care ni le oferă platforma.

Limitările acestei adaptări sunt lipsa de control asupra sistemului, întrucât funcțiile acestea sunt la interior containere care sunt executate la cerere asta înseamnă ca nu putem pregăti un mediu pentru acestea. De asemenea acestea nu pot rula la infinit pe toate platformele existând o limită maximă, acest lucru însemnând că nu putem rula task-uri de lungă durată pe ele. În domeniul serverless există noțiunea de cold start și warm start. Întrucât funcțiile nu rulează în continuu, la prima rulare acestea au nevoie de timp de inițializare. O dată inițializate acestea pot rula mult mai rapid. De asemenea, nu există conceptul de stare, funcțiile de acest tip nu pot accesa date ale unor rulări anterioare ci doar ce e în instanța curentă.

Cu toate acestea, funcțiile de acest tip ne oferă o flexibilitate în rulare. Dacă în cazurile precedente, un microserviciu era asociat unei singure instanțe de mașină virtuală sau de container, la funcții nu e neapărat așa. Putem avea câte o funcție pentru fiecare serviciu sau un microserviciu format din mai multe microservicii însă pot apărea dificultăți în partajarea datelor din baza de date.

4.3 Infrastructure as a Service

4.3.1 General

Conceptul de închiriere de echipament nu ar trebui să surprindă pe nimeni. În general, echipamentele ce ar necesita o investiție semnificativă pentru o folosire extrem de rară ar permite ca împrumutarea acestora să nu fie o alegere proastă, de exemplu închirierea unei pereche de schiuri în contextul în care mergem o pentru prima dată sau o dată la o perioadă îndelungată. Compania ce închiriază schiurile trebuie să se asigure de buna funcționalitatea acestora iar noi plătim prețul pe care l-am negociat.

Însă atunci când vorbim de infrastructură, aceasta este posibil folosită în fiecare moment pe Internet, rentarea închirierii ar putea provenii strict de la serviciile pe care distribuitorul le-ar oferi, astfel unul din motivele pentru care ai opta pentru o astfel de opțiune nu este doar pentru componenta fizică ci și pentru bonusurile care provin cu acestea.

În lumea infrastructurii cloud, serviciile oferite pot fi variate începând de la resurse de calcul sub formă de platforme de procesare scalabile și actualizabile în orice moment, spațiu de stocare hostat cu servicii suplimentare precum criptare la rest sau redundanță, rețelistică ușor definită, sisteme de monitorizare și observare până la sisteme de Machine Learning deja antrenate alături de alte servicii precum Text-to-Speech sau servicii de traducere în timp real.

Prin implementarea unui serviciu de infrastructură cloud, complexitatea procurării de echipament este abstractizată prin adăugarea unor resurse noi virtuale, mentenanța la nivel fizic este asigurată de către provider iar în funcție de serviciul ales chiar mai mult din administrarea acestora, astfel ne putem concentra doar pe dezvoltarea și simplificând lansarea.

4.3.2 Extinderea limitelor

În comerțul obiectelor cu amănântul, conceptul de cumpărare în masă la un preț mai avantajos sună ușor nedrept însă furnizorilor să ofere servicii la prețuri foarte accesibile, din acest motiv uneori poate să fie mai rentabil să folosim această abordare.

Aceștia pot să creeze ferme de servere pe care să le ofere spre folosire, care deși nu sunt nelimitate, ar acoperi suficiente neajunsuri, astfel putem crea aplicații ce beneficiază de această resursă nelimitată în calcul.

În martie 2014, Titanfall a apărut pentru PC, Xbox One și Xbox 360. Consolele noi apărând în 2013, nu este neașteptat ca să existe jocuri disponibile pentru ambele, însă în general acestea sunt dezvoltate ca țintă consolele vechi. În acest caz, dezvoltatorul Respawn Entertainment lucra asupra versiunii de PC și Xbox One în timp ce un studio extern, Bluepoint, lucra asupra versiunii vechi.

Portarea unui joc ce vizează o platformă mult mai puternică către una mai slabă vine cu dificultățile ei, mai ales în cazul acestui joc ce este destul de complex (12 jucători iar fiecare dintre ei poate avea un titan pe care îl controlează, alături de aceștia existând și entități necontrolabile ce sunt de ambele părți).

În interviul dat pentru Digital Foundry [10] aceștia menționează dificultățile portării incluzând reducerea materialelor (asset) și îmbunătățirea modurilor de randare și procesare întrucât Xbox 360-ul este mult mai limitat decât Xbox One-ul.

Însă, un lucru special al acestui joc este faptul că este disponibil doar cu o conexiune activă la Internet, mai mult decât atât, acesta are doar o componentă ce conține mai mulți jucători (fără o poveste), lucru ce în absența anumitor cunoștințe ar face să privesc o astfel de abordare ca o lăcomie din partea distribuitului.

Chiar și cu o consolă mai puternică, la începutul unei generații dezvoltatorii nu pot să fructifice toată această putere, astfel aceștia au creat un parteneriat cu cei de la Microsoft ca pentru toate versiunile să se folosească servere ce provin din Microsoft Azure pentru a îmbunătății experiența.

Cel mai simplu mod prin îmbunătățirea performanței a fost trecerea la servere dedicate, în general jocurile pentru a economisi prețul infrastructurii foloseau o arhitectură de tip peer-to-peer însă în cazul Titanfall acesta nu ar fi fost articol, din interviul [10] am extras: "The server performance requirements for Titanfall are pretty high, so you can't run peer-to-peer. This also means you can't have a many virtual machines per physical server, meaning if you've got a popular game, then you're going to need a pretty big investment in server hardware."

În interviul [13] se menționează mai multe detalii în acest aspect, "We have all of this AI and things flying around in the world; that has obviously let us build a different game than we would have if we'd have gone with player-hosted", astfel dispozitivele se pot ocupa doar de procesarea informațiilor primite și să nu se ocupe și de obținerea acestora "Because Titanfall's advanced AI is handled by the Azure servers, your Xbox's or PC's innards can be used to achieve more detailed graphics and the game's silky-smooth frame rate. The Titan bodyguards, dropships and legions of AI-controlled combatants are essentially free from a processing-power standpoint.". În același articol se menționează cum serverele disponibile în mai multe regiuni, capacitatea de scalare au putut duce proiectul la bun sfârșit.

În articolul precedent [10] aceste informații despre avantajele cloud-ului se extind, întrucât pentru a face performanța suficient de bună a trebuit să calculeze anumite elemente

direct în cloud. "Respawn wrote a new PVS system that worked dynamically, likely because that's much better for iteration but it need additional CPU performance we haven't got on Xbox 360. We took a different approach and pre-calculated static visibility using a big farm of Xbox 360s; these ran multi-hour visibility calculations which reduced the amount of stuff the 360 had to draw without requiring runtime computation and allowed us to use mostly the same geometry as the Xbox One version. Same deal for shadows, pre-calculate. Also creating occlusion geometry to reduce pixels shaded, tech art to the rescue there getting those built!"

Exemplul precedent acoperă doar resursele de calcul, însă furnizorii de cloud nu se limitează doar la atât, aceștia oferă servicii mult mai complexe astfel putem să extindem capacitățile unui dispozitiv folosind resursele provenite de pe servere la distanță al căror disponibilitate este mai ridicată față de ce am putea aduce pe cont propriu.

4.3.3 Microsoft Azure

Principalul furnizor de elemente de cloud este Amazon prin Amazon Web Services, aceștia domină piața întrucât la început aceștia erau singurii care ofereau servicii în care companiile doreau să investească, însă recent, avantajul lor pe piață este în scădere datorită avansurilor competiției [9], mai ales în lumea furnizorilor cloud nu este suficient să oferi servicii de hostare întrucât acestea permit o ușoară migrare către alți competitori în funcție de preț. Din acest motiv, aceștia oferă multe servicii prin care trecerea devine mai dificilă, sau chiar imposibilă întrucât serviciul este disponibil doar pe acea platformă, însă în general aceștia oferă servicii asemănătoare, cel puțin la un nivel de bază.

Am ales să analizez Azure pentru că avem credite ca și studenți. Această platformă a pornit sub numele de Windows Azure anunțat în 2008. Pe parcursul timpului, acesta nu a prins atracția publicului larg întrucât era destul de limitată și tinde să se schimbe foarte des. În perioada aceea Microsoft se baza pe Windows și tehnologii proprii însă începând cu 2014 Microsoft a început să folosească și să ofere servicii ce provin din software open-source și își schimbă numele în Microsoft Azure pentru a reflecta acest lucru.

Elemente de bază

Atunci când vrem să folosim un serviciu de cloud, este necesară crearea unui cont, în general putem folosi un cont Microsoft, fie el personal sau unul administrat de o organizație.

După crearea unui cont acestuia trebuie să îi se asocieze o sursă de plată, întrucât acest tip de servicii o să creeze facturi doar pentru cât consumă, astfel administrarea serviciilor devin mult mai ușoară.

Însă Azure oferă mai multe tipuri de abonamente și modalități de organizare ale acestora incluzând acces pe bază de rol, destinat companiilor.

Toate elementele create în Azure sunt administrare printr-o interfață REST, Azure Resource Manager. Acesta poate fi accesat creând cereri HTTP alături de verbele frecvent utilizate, comunicând prin JSON. Însă pentru a spori eficiența, au fost dezvoltate SDK-uri de administrare folosite ca librării pentru diferite limbaje de programare precum Python sau Java, dar și executable-uri făcute pentru a fi rulate direct dintr-un terminal prin intermediul Azure PowerShell (mai vechi, cu o sintaxă asemănătoare comenzilor de Windows) sau Azure CLI (mai rapid și cross-platform)

Scopul folosirii serviciilor cloud este eliminarea necesității de organizare a serverelor, și administrarea acestora printr-un terț ce se oferă să le întrețină. Un avantaj al acestora este faptul că putem alege locul în care aceste servere sunt amplasate, nu suntem constrânși de locul fizic al acestora, astfel putem defini Azure Regions.

Acestea reprezintă locul în care cei de la Microsoft au fizic centrele de operare ale serverelor. În cadrul acestora există mai multe centre de disponibilitate, astfel aceeași zonă prezintă mai multe servere separate ce stau în spatele unui load balancer. Putem folosi aceste zone pentru a crea replici ca un back-up.

De multe ori avem regiuni pereche, acestea fiind situate într-un loc aproximativ aproape, de exemplu UK South - UK West, acestea pot fi folosite pentru backup.

Alegerea unei regiuni poate fi ușor dificilă, în primul rând contează locul în care o amplasăm pentru a scade latența către utilizatori, prezența zonelor de disponibilitate sau a unei regiuni pereche ca și backup însă datorită modului cum acestea apar, nu toate caracteristicile sunt disponibile în toate regiunile, mai mult, toate caracteristicile oferite de Azure pot să aibă preț diferit în funcție de regiune, astfel atunci când facem o alegere avem mulți factori ce ar trebui să îi luăm în considerare.

Stocare

În general, tipurile de date cu care interacționăm pot fi de mai multe feluri, nestructurate, în general obiecte de mari dimensiuni ce reprezintă executabile sau fișiere de variate tipuri, date semi-structurate, acestea nu au o formă impusă însă tot au anumite similrități sau moduri de indexare, și date structurate ce reprezintă date asemănătoare cu cele pe care le-am găsi în bazele de date relaționale.

Modul de interacționare cu stocare se face prin crearea unui cont de administrare a stocării ce poate fi de mai multe tipuri și are ca scop păstrarea unor anumite setări de configurare. Acesta poate fi de mai multe tipuri general v2, v1 sau blob și ne poate limita la tipurile de date pe care le putem folosi. De asemenea, putem alege modul de stocare al datelor, acestea putând să fie pe hard disk-uri ce sunt mai lente sau stocare mai scumpă însă mai rapidă prin SSD-uri. În final, putem selecta modalități de replicare, local (LRS), în care se crează 3 copii în interiorul unui centru de date din aceeași zonă, regional (ZRS), în care se crează 3 copii dar distribuite în mai multe centre de date din aceeași zonă, georegional (GRS) 3 copii în mai multe centre din aceeași zonă și altele 3 copii

într-o regiune secundară pereche sau georegional doar citire (RA-GRS) în care se primește acces doar pe citire în conturile secundare din stilul GRS.

Cel mai folosit mod de stocare al datelor este Azure Blob Storage, acesta este un tip de stocare de date nestructurate, astfel putem să îl folosim asupra oricărui tip de date, neavând optimizări specifice. Încărcarea de date prin acest serviciu se poate face grafic sau prin diferite API-uri. Atunci când folosim acest serviciu, costurile vin din mai multe categorii: volumul de date, numărul de tranzacții prin interfața REST și traficul de descărcare. Pentru optimizarea acestor prețuri putem alege între trei categorii, unul pentru date frecvent accesate, pentru date accesate mai puțin frecvent și stocate pentru o perioadă de cel puțin 30 de zile și cel de arhivare pentru date accesate foarte rar (mai mult de 180 de zile).

Pe lângă Blob Storage, avem File Storage ce poate să stocheze date sub forma unui drive montabil pe Internet, Table Service ce reprezintă o bază de date NoSQL de tipul cheie / valoare, fiind o variantă mai limitată a unei baze de date de acest tip întrucât nu are georeplicare sau scalare, și Queue Service ce reprezintă o platformă de transmitere de mesaje de tipul Event Bus în arhitectura microserviciilor.

De asemenea, putem crea Azure Virtual Hard Disks, acestea fiind necesare pentru a fi montate către mașinile virtuale create.

Mașini virtuale

Reprezintă cea mai simplă metodă de a face rost de o platformă de procesare a datelor. Fiecare mașină virtuală are repartizat o anumită putere de procesare și o anumită cantitate de RAM, asociat unei unități de stocare și conectat la o anumită rețea. Toate aceste elemente sunt configurate atunci când creăm mașina virtuală însă pot fi modificate și după ce este lansată.

Putem folosi diferitele metode de acces ale API-ului Azure pentru a modifica aceste date și putem folosi diferite strategii de migrare precum încărcarea unui .vhd pentru a urca o mașina virtuală deja existentă.

Modul de monetizare ale mașinilor virtuale este că plătim pentru fiecare secundă în care aceasta are resurse alocate. În funcție de nevoile noastre putem folosi diferite tipuri de procesoare în funcțiile de nevoile noastre, fiecare având optimizările proprii, în același timp putem selecta tipul de stocare de care avem nevoie, pornind de la HDD-uri standard către SSD-uri ultra rapide, opțiunile mai performante necesitând o investiție mai mare. Fiecare mașină virtuală necesită cel puțin o interfață de rețea virtuală și opțional adrese IP publice sau setări de firewall.

Un considerent asupra mașinilor virtuale este modul în care acestea sunt lansate. Cele mai importante aspecte sunt disponibilitatea și scalarea. Implicit, nu avem nicio astfel de garanție, din acest moment putem să încercăm fie seturi de disponibilitate în care mașinile virtuale sunt puse în același centru de date dar în zone fizice diferite sau o zonă

de disponibilitate în care mașinile virtuale sunt puse în regiuni diferite. Scalarea mașinilor virtuale este mai dificilă, însă există scalare orizontală prin intermediul Azure Virtual Machine Scale Sets, sunt dificil de implementat necesitând un load balancer, scumpe deoarece plătim pentru fiecare mașină și necesită configurări și sincronizări manuale.

Containere

Întrucât VM-urile au dezavantajele proprii ale consumului de resurse sau al modului de configurare, putem lansa containere în Azure prin Azure Container Instances. Acestea sunt configurate aproximativ la fel ca mașinile virtuale, singura diferență fiind necesitate specificării numele imaginii containerului și repository-ul de unde acestea să provină însoțit de detaliile de autentificare dacă acestea sunt necesare.

Pentru a facilita lucrul cu containere, avem la dispoziție Azure Container Registry ce ne permite să creăm propriul registru de imagini, putând să adăugăm imaginile proprii pe el ca după să putem să le consumăm atunci când vrem să lansăm o aplicație sub formă de container.

Pentru a abstractiza management-ul stării produsului, avem Azure Kubernetes Service prin care Microsoft abstractizează nodurile de management iar conținutul acestuia poate fi administrat printr-o interfață similară a lui kubectl, prin az aks.

Azure App Service

Reprezentând un PaaS, App Service este o modalitate de a lansa a codului fără întreținerea explicită a infrastructurii din spatele acesteia, astfel nu e necesar configurarea unei mașini virtuale și după încărcarea aplicației pe aceasta.

Avantajul renunțării la aceste configurări nu este doar ușurința modului de lansare ci și posibilitatea unor facilități care nu sunt disponibile mașinilor virtuale clasice, incluzând replicare și disponibilitate globală, autoscalare dinamică, integrare nativă în pipeline-uri.

Acest serviciu permite lansare prin intermediul încărcării codului unei aplicații în mod direct sau prin intermediul unui container. Configurarea se face asemănător cu o mașină virtuală

Prin intermediul interfeței putem să facem configurări minimale asupra aplicației, însă acestea sunt limitate și au ca scop mai mult moduri de rulare precum setări de mediu, porturi sau de SSL.

Monitorizarea unui astfel de serviciu se poate face prin Application Insights ce se face integrând SDK-ul în codul client, ulterior putem face rost de date precum timp de răspuns, performanțe sau telemetrie. Toate aceste date fiind disponibile în interiorul unui Dashboard central.

Serverless

Un concept ce este promovat de către furnizorii de cloud este capacitatea de a lansa cod fără să ne intereseze locul și modul în care acestea rulează. Deosebirea față de mașinile virtuale sau App Service este că putem plăti doar cât consumăm, adică doar timpul în care codul rulează.

Există două tipuri în care putem lansa cod prin această metodă, însă este important să facem diferențierea de rulare. Tehnologia serverless rulează doar atunci când este rulată, în spate acestea pot fi sub formă de containere Docker care sunt rulate în momentul în care e necesar, din acest motiv apare fenomenul de cold-start, atunci când containerul trebuie să fie inițializat, însemnând că prima rulare a funcției este mai înceată. În timpurile recente, furnizorii au optimizat acest timp iar rularea unui cold start este mai rară, acestea fiind păstrate într-o stare în care pot fi rulate aproape instantaneu.

Un alt concept important este faptul că nu avem acces la o stare implicit, fiecare rulare a funcției vine cu propriile reinițializări, astfel nu putem să accesăm datele unei rulări anterioare fără să o salvăm conștient într-un anumit loc, precum într-o bază de date externă.

Azure Functions rulează atunci când sunt invocate, și acesta poate fi configurat în funcție de necesități precum atunci când sunt invocate prin intermediul unui endpoint HTTP, la un anumit interval sau la evenimente precum încărcarea unui blob.

Acest tip de execuție, deși abstractizează aplicație, crearea unei funcții necesită scriere de cod în anumite limbaje compatibile. Acest proces poate fi simplificat și mai mult folosind Logic Apps, acestea sunt create aproape fără cod și printr-o interfață grafică. Componentele unui Logic App sunt predefinite prin diferiți conectori ce integrați împreună vor crea un flux de lucru ce vor fi executate la anumite evenimente.

Baze de date

Un alt serviciu pe care îl folosim în crearea unei platforme este o bază de date. Atunci când avem infrastructură în cloud putem avea mai multe opțiuni de abordare al acestei probleme.

Una dintre ele este crearea unei mașini virtuale și în stil clasic crearea serverului de bază de date și configurarea setărilor de securitate necesare conectării. În acest stil putem folosi orice versiune și tip de bază de date însă vom fi forțați să ne ocupăm de mașină, astfel va trebui să ne ocupăm de mentenanța software-ului ce funcționează pe aceasta.

O altă alternativă este folosirea unei opțiuni de tip PaaS pentru baze de date, în acest tip toată mentenanța pe care ar trebui să o facem este ascunsă, la fel și backup-urile și georeplicarea. Deși pierdem din configurări, avem o metodă de folosire a bazei de date mult mai ușoară.

Pentru bazele de date relaționale avem SQL Database ce are suport pentru MySQL,

MariaDB sau Postgres, al cărui tip de monetizare poate fi bazat pe Database Transaction Unit ce calculează folosința în funcție de CPU, I/O și memorie, vCore ce se face specificând un număr de instanțe de mașini virtuale iar numărul acestora va determina consumul și dacă avem mai multe baze de date avem modelul elastic ce vor cupla mai multe baze de date atunci când este înregistrat consumul.

De asemenea, avem opțiunea pentru baze de date nerelaționale prin Azure Cosmos DB ce are suport pentru multe variante precum MongoDB sau Cassandra ce vor putea avea avantajele unei baze de date de acest tip precum georeplicare și flexibilitatea în consistența acestora.

Monitorizare

Furnizorii de infrastructură cloud ne permit să creăm un număr nelimitat de echipamente pe care să le folosim în proiectele noastre, din acest motiv devine mai dificil să observăm capacitățile acestora, astfel în absența unor erori care să cauzeze ca serviciile să fie inaccesibile, nu le vom putea depista.

În orice caz, chiar și atunci când știm că avem o problemă și vrem să o investigăm, tot avem nevoie de o modalitate de a afla ce s-a întâmplat în trecut. Din acest motiv, Azure ne oferă câteva unelte pe care le putem folosi pentru a ne ușura procesul.

Azure Monitor este serviciul de bază de monitorizare ce colectează date fundamentale din resursele create, metrice ce sunt valori numerice ale resursei dar și fișiere de jurnalizare ce descriu evenimente ce se produc în cadrul sistemului.

Azure Monitor Activity Log memorează informații legate evenimente de control ce se declanșează în jurul resursei, acestea pot include informații despre orele de restart a unei mașini virtuale sau accesarea spațiilor de stocare.

Tot cu acesta putem să folosim metricile colectate pentru a crea grafice informative, de a crea alerte în cazuri specifice pe care le putem configura și putem colecta mai multe informații diferite în funcție de tipurile de resurse pe care le folosim, precum informații ce ar putea fi utile în diagnoza unei probleme cu o mașină virtuală, însă în acest caz trebuie instalată pe acea mașină extensia Azure Diagnostics.

Azure Application Insights este versiunea îmbunătățită a lui Azure Monitor însă aceasta necesită achiziție separată însă este o unealtă mult mai complexă ce ne permite să monitorizăm și să jurnalizăm starea unui sistem ci și performanța acestuia, astfel ne permite să colectăm date detaliate și să o prelucrăm într-un mod mult mai eficient.

Azure Log Analytics este o unealtă de centralizare a fișierelor de jurnalizare create pe Azure, acesta pot include resurse ce provin din mai multe subscripții dar și alte produse din afara infrastructurii precum DNS Resolver sau Office 365.

Log Analytics ne oferă o interfață în care toate aceste informații pot fi accesate prin intermediul unui limbaj specializat, KQL (Kusto Query Language) similar cu SQL.

Servicii de Inteligență Artificială

Nu numai asupra serviciilor de infrastructură se poate vedea o îmbunătățire prin folosirea furnizorilor de cloud ci și asupra serviciilor de inteligență artificială sau de machine learning.

Astfel, putem beneficia de o ușurare a muncii pe care o depunem atunci când vrem să adăugăm algoritmi creați cu scopul prelucrării de date.

La cel mai simplu nivel, putem folosi Azure Cognitive Services ce oferă unelte variate în mai multe domenii, precum viziune (clasificare de imagini, OCR, recunoașterea scrisului de mână), cunoaștere (extrage de întrebări de format QnA din text nestructurat), procesarea limbajului natural (extras de cuvânt cheie, traduceri de texte automate, corectarea scrisului) vorbit (transcriere din vorbit în scris, recunoaștere a vorbitorului) sau de căutare (Bing Web Search).

Mai mult decât atât, putem folosi platformele serverless în avantajul nostru, astfel putem folosi modele preantrenate pe care furnizorii ni le distribuie alături de algoritmi eficienți pentru a crea funcții ce rulează doar atunci când avem nevoie de acestea.

O altă unealtă reprezintă Microsoft Machine Learning Studio ce este bazat în browser și ne permite să creăm proiecte ce au ca scop explorarea de date, învățat supervizat și nesupravegheta dar și diferite forme de analiză fără a fi necesară scrierea de cod ci prin elemente puse la dispoziție de platformă.

4.4 Docker

4.4.1 General

Docker este una dintre tehnologiile ce a ajutat la revoluția DevOps de a eficientiza modul de lucru în dezvoltarea software. Acesta a apărut în 2013, și a început să fie viabil și cunoscut începând cu 2015 însă la momentul de față consider că este o unealtă necesară ce nu mai este o noutate și și-a găsit locul în piață.

Containerizarea nu a era o tehnologie nouă iar Docker nu este primul și nici singurul care a contribuit în acest spațiu, totuși aceștia sunt unii dintre cei mai populari fapt datorat de ușurința anumitor sisteme noi pe care aceștia l-au adus în favoarea competiției, aceștia începând trend-ul chiar încă din 2008 prin LXC.

Docker a apărut ca o alternativă la lansarea aplicațiilor în mașini virtuale. Deoarece acestea rulau efectiv pe hypervisor și necesitau anumite resurse alocate alături de rularea întregului sistem ceea ce le făceau greu de configurat. Docker a îmbunătățit acest proces, aplicațiile care rulează în container sunt mult mai eficiente din punctul de vedere al consumului de resurse și pot fi lansate mult mai rapid.

Docker permite abstractizarea naturii aplicației, astfel indiferent de ce rulează în ea, o altă persoană poate să interacționeze cu ea foarte rapid știind doar un număr limitat de

potențiale elemente de configurare. Acest lucru permite separarea responsabilităților în procesul de creare. Dezvoltatorii cunosc faptul că aplicația lor va rula într-un container, astfel aceștia au responsabilitatea de a containeriza aplicația. Operatorii cunosc modul de operare cu Docker, astfel aceștia pot administra containerele, fără a avea vreo cunoștințe ce e în ele ci doar de resursele la care acestea ar trebui să aibă acces.

Datorită modului în care funcționează, Docker permite modularizarea aplicațiilor, astfel fiecare container și-ar rula propria aplicație, astfel poate fi lansat independent, fiind susținută o arhitectură bazată pe servicii dar și microserviciile.

4.4.2 Componente

Docker face parte din proiectul Moby, și la mod general reprezintă o aplicație modelată de arhitectura client-server, în care un executabil Docker va reprezenta un client ce comunică cu un server ce reprezintă Docker Engine-ul sau Docker daemon-ul ce are responsabilitatea de a rula containerele în funcție de comenziile primite de la clienți.

Nu este necesar ca această aplicație să ruleze asupra aceleiași mașini, astfel putem rula clientul Docker independent și să folosim Docker Engine-ul pe o altă unitate pe care îl putem controla și prin API RESTful.

Componenta de bază reprezintă imaginea. Acestea reprezintă modul în care aplicația noastră este construită conținând instrucțiunile de formare. Acestea pornesc de la o imagine de bază iar fiecare instrucțiune crează imagini intermediare ce se suprapun sub diferite layer-uri până se creează complet o imagine.

Registriile reprezintă locul în care imaginile sunt stocate. De aici putem descărca imaginile create și să le rulăm. Aceste locuri pot fi publice, precum Docker Hub, dar și private putând să fie rulate independent. Atunci când le folosim acestea pot necesita autentificare, de exemplu pe Docker Hub putem să avem imagini publice ce pot fi văzute și accesate de orice utilizator dar și imagini private ale căror drepturi de citire sunt restricționate.

Containerele reprezintă o imagine construită ce este rulată, la fel cum un obiect este instanțierea unei clase. Acesta poate avea diferite stări externe precum dacă este creată, lansată sau distrusă, dar și interne, astfel acestea pot evolua în funcție de modul cum se interacționează cu ele. Acestea sunt izolate unul de celălalt, astfel chiar dacă o mașină are mai multe containere acestea nu pot accesa datele unui alt container fără să intervenim noi.

Extinderea conceptului de folosire de containere cu Docker reprezintă orchestrarea. Rularea a unui număr limitat poate fi făcută cu ușurință manual, însă atunci când avem mai multe aplicații, dependințe (precum baze de date) și alte configurări de rulare poate deveni destul de haotic, astfel s-ar pierde multe dintre beneficiile pe care Docker le aducea. Din acest motiv putem folosi aplicații specifice menite pentru extinderea Docker-ului

precum Docker Compose, Docker Swarm sau Kubernetes pentru a administra sisteme mai complexe.

4.4.3 Utilizare

Din punctul de vedere al unui dezvoltator, folosirea Docker-ului se rezumă doar la câteva instrucțiuni de bază ce pot fi cuprinse pe o listă foarte [scurtă](#), din acest motiv o prezentare în detaliu al fiecărei comenzi nu ar fi eficientă.

Însă, pot sublinia câteva utilizări generale după instalare.

- Pornirea oricărei aplicații ce are o imagine publicată pe Docker Hub, precum Ubuntu, `docker container run ubuntu` va încerca să lanseze un container ce rulează din imaginea ubuntu. Dacă nu găsește imaginea ubuntu acesta va încerca să găsească imaginea ubuntu cea mai recentă (latest) de pe repository-ul implicit (Docker Hub), dacă o va găsi acesta o descarcă și o să ruleze
- Atunci când o imagine este rulată, acesta este executată cât timp are un proces în execuție, din acest motiv, containerul anterior de Ubuntu se va închide imediat ce va porni. Putem accesa și prelungii viața containerului prin preluarea STDIN-ului și atașarea unui pseudo tty folosind `docker container run -it ubuntu bash`, acesta va rula alt container, și ne vom găsi în interiorul unui terminal cu `root@containerid`.
- Când lansăm o imagine, implicit acesta primește o rețea de tipul bridge, astfel aceasta poate fi conectată la rețeaua computerului principal. Putem schimba acest aspect prin directivele din `-network` ale comenzi `docker container run`.
- Containere diferite nu pot implicit să folosească datele altui container. Din acest motiv, Docker ne permite să configurăm Volumes, acestea au rolul de a salva anumite date și să fie accesibile sub diferite forme. Putem lăsa Docker să le administreze sau putem crea bind-mounts care pot să facă ca fișierele sau un director de pe mașina gazdă să fie sincronizate cu un director din interiorul mașinii. Acestea se fac folosind directivele `-v` ale comenzi `docker container run`.
- În general am vrea ca aplicația din interiorul containerului să fie accesibilă din exterior. Acest lucru se poate face sub diferite forme, însă în general vrem să configurăm un port de acces. Astfel aplicația din interior comunică pe un anumit port (de exemplu 3000), însă Docker nu o să deschidă implicit aceste porturi. Pentru a face asta, trebuie să folosim directiva `-p` care va asocia port-ul mașinii gazdă cu port-ul din interiorul containerului. De exemplu `docker container run -p 3001:3000`, port-ul 3001 pe gazdă va accesa aplicația din container care rulează pe portul 3000.

- Containerele sunt puse într-o rețea generală de către Docker, însă putem crea rețele proprii și să agregăm mai multe containere în el, astfel nu este nevoie să oferim acces către exterior tuturor serviciilor ci doar celor strict necesare.
- Construirea unei imagini se face, în general, folosind un Dockerfile, acesta pornește de la o imagine de bază asupra căreia se adaugă diferite configurări precum comenzi sau specificare de argumente de build sau environment. Comanda este `docker image build .` ce va construi imaginea din Dockerfile din directorul curent. Putem taga imaginea folosind directiva `-t`.
- Încărcarea unei imagini într-un repository se poate face folosind `docker image push`.

4.4.4 Simpla orchestrare

O metodă ușoară de a crea un sistem format din mai multe containere este de a folosi Docker Compose. Întrucât Docker nu a fost făcut pentru a administra mai multe containere în același timp și în același timp acesta susține ca unui singur serviciu să îi corespundă un container, putem avea dificultăți atunci când am vrea să avem un sistem mai complex, de exemplu un server web și baza de date asociată acestuia.

Docker Compose este o metodă de orchestrare, astfel acesta crează metode de rulare de containere, astfel dacă vrem să avem o aplicație în Compose, acesta trebuie să fie deja sub forma unei imagini.

Pentru a începe, avem nevoie de un fișier `.yaml`. În acesta putem defini diferite servicii ce reprezintă defapt containerele pe care vrem să le rulăm. Pentru fiecare acest serviciu putem defini în acest fișier toate proprietățile pe care le-am putea defini atunci când am rula `docker container run`, precum volume, rețele, porturi. De asemenea putem defini (pornind din imagine) health check-uri, comenzi care ar fi rulate în interiorul containerului și diferite proprietăți care ar defini starea unui serviciu, astfel în funcțiile de setările făcute, dacă un serviciu de care depindem nu este disponibil, am putea închide toate serviciile care depind de acesta, sau să încercăm resetarea serverului afectat.

Docker Compose ne permite să avem o imagine mai per ansamblu asupra sistemului pe care vrem să îl creăm. Însă acesta are limitările proprii ce îi oferă imaginea de o metodă simplă, una dintre acestea fiind faptul că nu avem asistență pentru replicare sau metode de actualizare ale versiunii unei imagini ce pot fi găsite în versiuni mai avansate ale sistemului.

O metodă de a face rost de replicare într-un mod simplu este folosirea unui serviciu de service discovery precum Hashicorp Consul, astfel putem avea un cluster pe care să rulăm diferite containere replicate, însă nu am cunoaște exact locația acestora. În același stil, locația acestora ar putea să se schimbe etc. O astfel de unealtă folosește metode asemănătoare cu un DNS pentru a crea comunicare în cadrul unui cluster, astfel am

putea să avem un load balancer și prin intermediul Consul să se ruteze către replica cea mai avantajoasă. Specific acestui serviciu putem să beneficiem de mutual TLS pentru a securiza rețeaua.

O altă metodă de orchestrare simplă reprezintă Docker Swarm, aceasta vine la pachet cu Docker Engine, având un mod special pentru acest tip de operații. În acest mod, un Docker Engine se poate alătura(sau crea) unui cluster de alte noduri care rulează în acest mod. Fiecare nod va fi fie Manager sau Worker, având minim un master(inițial, primul nod este nodul manager). Nodurile manager au rolul de a distribui task-urile primite către nodurile worker, însă acestea pot efectua task-uri la fel ca un nod obișnuit, însă aceasta nu este o prioritate pentru ele sau poate fi dezactivată în totalitate. Un avantaj al acestui serviciu este posibilitatea de a rula replici ale aceluiași serviciu, astfel putem ca un singur container (serviciu) să ruleze simultan în mai multe locuri, astfel dacă un loc este afectat le putem accesa pe celelalte, funcționând ca un load balancer. O limitare al acestui sistem este faptul că această scalare nu este automată și că sunt metode limitate de monitorizare implicit însă este extrem de ușor de configurat și de rulat, astfel pentru unele sisteme reprezintă o alternativă suficient de bună a unor soluții mai complicate.

4.5 Kubernetes

4.5.1 General

Containerele au reprezentat o soluție pentru problema comună privind modul de rulare și lansare pe diverse mașini diferite ale căror sisteme, dependențe, versiuni de rulare și librării pot fi diferite (frecventa „But it works on my machine”), astfel un dezvoltator poate să construiască și să lucreze la aplicație, acesta fiind atent să creeze imaginea, imediat după un tester nu e nevoie să interacționeze direct cu sursa ci exclusiv cu imaginea, putând să verifice că nu a apărut nicio problemă și pe mașinile de build să ruleze testele automate direct pe imagine. Ulterior, dacă imaginea reprezintă ceva ce trimitem în piață poate fi trimisă către operatori și planificat pentru lansare.

Utilitatea Docker-ului și a containerelor fără nimic suplimentar s-ar termina totuși în acest punct. Chiar dacă un operator nu ar trebui să instaleze dependențele aplicațiilor în fiecare loc unde ar rula, acesta fiind înlocuit cu instalarea runtime-ului de Docker, acesta trebuie să se preocupe cu dificultățile lansării, adică unde este lansat, în cât timp, replicare, dacă sistemele sunt în stare bună și alte detalii ce pot oglindii starea sistemului.

În acest scop, Docker nu ajută, la fel și itemii de care am discutat la simpla orchestrare a containerelor, fiind o soluție în regulă pentru sisteme mici, atunci când vorbim la scală avem nevoie de unele diferite. Aici continuăm folosind Kubernetes, o unealtă ce a apărut în 2014 devenită open-source care are ca scop orchestrarea unui cluster pe care rulează sub diferite forme și organizate în moduri diferite containere

În dezvoltarea și crearea modului de interacționare folosind Kubernetes au fost folosite experiențele dobândite prin lucrul manual cu acestea, astfel s-a urmărit diferite puncte ce contribuie la avantajele în folosirea acestei tehnologii, incluzând creșterea vitezii de deployment prin crearea unor sisteme imutabile bazate pe stări și nu pe modificări punctuale asupra sistemului activ (configurare declarativă), prin crearea capacității de auto-reglare creat pornind de la acest mod de configurare, astfel Kubernetes la cel mai simplu nivel, va aserta starea curentă și va încerca să o aducă la starea pe care este configurat să o aibă

În același timp, Kubernetes permite abstractizarea infrastructurii, astfel nu mai este necesar să cunoaștem mașinile virtuale sau sistemele ce rulează în interiorul clusterului, suntem interesați doar de modul de interacționare cu clusterul, astfel creăm aplicații care să ruleze pe acesta cu anumite configurări. Astfel, Kubernetes poate să mențină mașinile virtuale mai încărcate întrucât acesta va încerca să eficientizeze consumul de resurse.

Soluțiile de lansare ce includ Kubernetes ne permit să fim foarte flexibili în modul cum interacționăm cu acesta, și avem multe oportunități de rulare, însă „prețul” plătit pentru toate opțiunile pe care le avem este o creștere foarte mare de dificultate în crearea și administrarea unui astfel de cluster, însă recent această dificultate poate fi eliminată prin folosirea de infrastructură în cloud precum Azure ce oferă Kubernetes as a Service, astfel administrarea clusterului nu mai este o problemă ci doar creăm resursele de care avem nevoie.

4.5.2 Obiecte de bază

Atunci când folosim Kubernetes, deși putem folosi și comenzi manuale, ar trebui să abordăm o modalitate de gândire declarativă. Kubernetes folosește obiecte la diferite nivele ce administrează obiecte la nivel inferior pentru a-și obține starea pe care este setat să o păstreze. Toate acestea se fac prin intermediul unor fișiere, în general de format YAML în care stabilim lucrurile de care avem nevoie.

Etichete și adnotări

Întrucât administrarea unui cluster Kubernetes include administrarea de foarte multe obiecte, fiecare dintre acestea pot să îi se adauge etichete ce sunt o pereche cheie-valoare, în format de String-uri. După aplicarea unei etichete atunci asupra obiectului, acestea se pot vedea atunci când inspectăm elementul. Aplicarea unei etichete se poate face atunci când creăm un obiect, prin adăugarea lor asupra comenzii run,

```
kubectl run ... --labels="ver=1" sau separata folosind kubectl label [selector] "ver=1".
```

O dată aplicat un label, acesta poate fi folosit pentru a selecta resursele cu acel label în cadrul altor comenzi precum cele de inspectare sau stergere `kubectl get pods --selector "ver=1"` iar multiple etichete pot fi combinate cu operații logice, `kubectl get pods -l 'ver=1,canary'` -> Pods-urile cu label-ul versiunii 1 și care au definit label-ul canary. Etichetele pot fi

aplicate și de către alte obiecte pentru a identifica resursele pe care acestea le au în administrare. De exemplu, un ReplicaSet va aplica etichete specifice asupra pod-urilor pe care acestea le administrează pentru a cunoaște ce schimbări ar trebui să facă.

Întrucât etichetele au rol în modul de administrare al cluster-ului, adnotările au ca rol adăugarea de metadata asupra unui obiect, sub forma tot de o pereche cheie-valoare ce pot folosi caractere mai variate ca etichetele. Aceasta poate fi preluată de alte librării încercând să le oferim diferite informații precum versiunea sau să îmbunătățim vizibilitatea în cadrul unei interfețe de administrarea a cluster-ului.

Pods

Elementele Kubernetes pornesc de la Pods, acestea reprezintă corespondentul unui container din lumea Docker, însă extins, astfel dacă vrem să rulăm un container în Kubernetes trebuie să folosim un pod.

Reprezintă o colecție de containere ce rulează împreună, astfel acestea partajează metodele de acces (IP-uri, hostname-uri), astfel este garantat că rulează în interiorul aceluiași nod, putând să comunice local prin metode de comunicare interproces, însă două pod-uri identice pot să ajungă să ruleze pe noduri diferite prin replicare.

Reprezentând o colecție de containere, în interiorul unui pod ar trebui să avem programe ce comunică des împreună și care au nevoie să beneficieze de avantajele rulării în același mediu, de exemplu un server și o metodă de sincronizare cu un sistem central.

Crearea unui pod se poate face imperativ prin folosirea comenzii **run**, astfel putem crea un pod cu diferite proprietăți pe care obișnuim să le specificăm și la containere **kubectl run NAME --image=image**. Însă modul de creare recomandat pentru pod-uri este prin intermediul unor fișiere manifest, cel mai popular fiind scrise în YAML. Rulând comanda de aplicare a modificărilor **kubectl apply -f pod.yaml**, considerând că în fișierul indicat avem un template pentru un pod și este valid, acesta va fi rulat pe un nod din cadrul clusterului Kubernetes pe care îl administrăm.

Pentru a vizualiza detalii despre pod-urile în execuție folosim comenzi precum **kubectl get pods** ce ne dă o listă de pod-uri și detalii succinte despre acestea și **kubectl describe pods NAME** unde primim mai multe detalii despre pod însă nu neapărat complet. Pentru a șterge un pod putem folosi comanda **kubectl delete pods/NAME** sau prin fișierul de configurare **kubectl delete -f pod.yaml** iar după acesta, după o anumită perioadă pod-ul va fi șters după o perioadă de grație.

O dată creat, serviciile dintr-un pod pot fi accesate direct prin intermediul comenzii **kubectl port-forward NAME port:port**, însă aceasta reprezintă o metodă simplistă de accesare și nu poate fi recomandată. Pentru a prelua fișiere de logare din pod putem folosi comanda **kubectl logs NAME**, pentru a rula comenzi în interiorul unui pod folosim **kubectl exec -it NAME bash** și de asemenea putem copia fișiere prin **kubectl cp <pod-name>:/fișier ./fișier.txt**.

Pentru a facilita elementele de un nivel mai înalt, pod-urile trebuie să aibă definite o comandă ce indică starea acestora, de exemplu dacă un server web este activ. Astfel, clusterul Kubernetes sau responsabilul aceluia pod vor rula acele comenzi pentru a afla statusul pod-ului, dacă acesta nu este pozitiv atunci pod-ul poate să fie recreat. În același fel, putem defini o comandă ce indică dacă un pod a fost pornit corespunzător și poate servi trafic, un readiness check, ce sunt create asemănător unui healthiness check.

Prin fișierul de configurare al unui pod putem specifica un set de resurse minime și maxime pe care un pod le poate avea, astfel putem să ne asigurăm că clusterul va rula pod-ul într-un mod eficient și în același stil ca la containere stocarea pod-urilor nu este persistentă, astfel la fiecare rulare datele pe care le avem pe acel pod sunt scrise cele de pe imagine și nu cele ale unor rulări anterioare. Pentru a adăuga această caracteristică folosim volume ce sunt indicate în fișierul de template.

Services

Elementele administrate de Kubernetes pot apărea și dispărea în orice moment, fie datorită unor scalări sau restrângeri de pods-uri în funcție de traficul care se desfășoară în cluster. Din acest motiv, dacă avem replici a unor mai multe pods-uri, accesarea acestora devine dificilă atunci când punem în context scalarea. Pe Internet folosim DNS pentru a avea acces la servere fără a memora adrese IP exacte ce oricum se pot schimba, însă în Kubernetes aceasta activitate este chiar și mai dinamică. Din acest motiv, nu putem folosi un DNS tradițional întrucât acesta poate cache elementele mult mai mult decât ne dorim.

În Kubernetes putem crea servicii, astfel atunci când creăm niște pods ce rulează pe un port putem crea un serviciu pentru acel port, astfel nativ vom avea un singur IP și un port pe care alte pod-uri le pot folosi pentru a accesa asemănător unui DNS.

Însă acest trafic poate fi creat doar din cluster, pentru a permite elemente din afara pentru a accesa serviciul și implicit serviciul putem folosi NodePort în fișierul de configurare iar Kubernetes ne va oferi un IP prin care putem accesa serviciul.

Ingress

Definirea serviciilor se bazează doar pe IP și pe un port, astfel el rulează la Layer 4, din acest motiv nu putem să creăm redirectionare bazată pe ce conține calea căutată.

Pentru a rezolva această lipsă, putem adăuga un layer suplimentar care pentru aplicațiile ce se bazează pe HTTP poate funcționa ca și load balancer sau ca reverse proxy și în Kubernetes au numele de Ingress.

Deși avem acces la obiectul de tip Ingress direct din Kubernetes, nu există implementarea acestuia astfel aceasta trebuie să fie adăugată ca separat, însă modalitățile de configurare și diferitele fișiere de care am avea nevoie pentru a configura din punct de

vedere Kubernetes sunt standardizate.

După ce configurăm Ingress-ul, vom putea avea acces la mai multe servicii din cadrul clusterului și vom putea să redirecționăm către acestea prin intermediul controllerului în funcție de configurările noastre, de exemplu după hostname sau căi și să rescriem căile de acces sau să folosim TLS.

Replica Sets

Am observat că există multe avantaje pe care le putem avea rulând același server de mai multe ori creând redundanță, astfel dacă un nod ar fi afectat ar exista un alt nod deja pregătit pentru a servi trafic. În același timp putem face ca acest nod să fie mai apropiat de anumiți clienți în funcție de zona serverului și în același mod folosind metodele de rutare din Ingress să partajăm traficul în funcție de conținutul pe care vrem să îl servim (sharding).

Modalitatea principală de a interacționa Kubernetes este declarativ stabilind o anumită stare ideală pe care vrem ca sistemul să o aibă, astfel acesta s-ar ocupa de a menține această stare, fie prin crearea mai multor resurse pentru a ajunge la acea stare inițială sau să scadă din resursele deja create.

În acest mod de operare, Replica Sets reprezintă fix acest lucru, acesta funcționează la un nivel mai înalt de Pod-uri, astfel specificând numărul de replici pe care ni-l dorim din template-ul unui specific Pod, Replica Set-ul va avea responsabilitatea de a menține acel număr. Pentru a face asta, el va aplica anumite etichete astfel va cunoaște asupra căror Pod-uri poate să acționeze.

Crearea de Replica Sets se face creând un fișier de configurare specific și aplicat folosind comanda `kubectl apply -f NAME.yaml`

Deployments

O altă problemă frecvent apărută este modalitatea de aducere a modificărilor în piață. Atunci când avem un server sau o colecție de servere pornite atunci când vrem să adăugăm anumite schimbări asupra funcționalităților oferite inevitabil vom avea un timp în care nu vom putea servi clienți, o perioadă necesară de mentenanță, în același fel putem întâlni problema în care există două servere în același timp ce ar face același lucru însă datorită existenței a două versiuni diferite, practic clienții pot apela oricare dintre versiuni.

Pentru a rezolva aceste probleme în Kubernetes avem obiectul de Deployments ce are ca rol să administreze un set de Replica Sets și poate să aducă actualizări în diferite moduri, recrearea tuturor pod-urile curente prin strategia de Recreate, aceasta crează o perioadă de down-time și prin RollingUpdate în care pod-urile sunt create pas cu pas, iar numărul de câte pod-uri necesare să fie disponibile sau alți parametrii pot fi configurați. Kubernetes face asta schimbând parametrii obiectelor de Replica Sets pe care acesta le

crează și terminând activitatea pod-urilor pe care acestea le au în supervizare.

Pentru a păstra stabilitatea sistemului, la fel ca în cazul Replica Sets, se vor urmări parametrii de healthiness și se vor aștepta ca acestea să fie ready și healthy.

DaemonSets

Atunci când configurăm un Deployment o facem pentru a avea un număr specific de replici ale unui serviciu, scalate în funcție de setări și ar fi preferabil ca poziționarea acestora să fie pe cât mai multe moduri distincte.

Însă existând necesitatea de a avea nevoie să avem servicii care rulează pe toate nodurile pe care le avem apare conceptul de DaemonSets, acestea au ca scop rularea unui container pe toate nodurile însă funcționalitatea acestuia are strict legătură cu rularea o singură dată pe acel nod iar replicarea acestuia nu ar aduce vreun beneficiu.

Crearea unui astfel de obiect se face tot prin intermediul unui fișier de configurare pe care îl aplicăm folosind comanda **apply**, și putem specifica anumite etichete pentru a specifica doar anumite noduri pe care vrem să ruleze serviciul.

Jobs

Kubernetes este creat pentru a orchestra rularea unor pod-uri care rulează pe o durată nedeterminată, însă folosind obiecte de tip Job putem orchestra și rularea unor pod-uri a căror durată de execuție este egală cu durata de execuție a programului din interior.

Astfel, prin API-ul disponibil avem acces la trei tipuri, *One Shot*, tipul de Job care este rulat o singură dată până când este executat de către un singur pod, iar în funcție de modul în care executăm comanda se poate restarta atunci când eșuează până când se execută cu succes. *Parallel* în care avem amii multe pod-uri care fiecare se execută independent iar rezultatul total este agregat și formează mai multe răspunsuri și prin pattern-ul de coadă, putem distribui un singur task să fie completat de către mai multe pod-uri.

De asemenea, avem acces la CronJobs ce sunt un tip special de Job care se execută la un interval de timp, acesta se declară printr-un template și va fi organizat de Kubernetes.

Config Maps & Secrets

Crearea unor imagini care pot fi rulate oriunde cu ușurință vine din standarul Docker, ce ofera suport pentru argumente de construire dar și de mediu într-un mod convenabil.

Astfel, în Kubernetes Config Map-urile sunt folosite ca un obiect ce poate să fie creat fie dintr-un template YAML sau imperativ prin comanda **kubectl create configmap** și ulterior poate fi folosit fie prin montarea acestuia ca pe un fișier în configurarea unui Pod, ca argument de mediu sau ca argument de comandă adăugând aceste detalii în fișierul

template (fie montarea volumului sau specificarea de unde să își ia valorile specificând că acestea trebuie luate din config map-ul creat anterior)

Întrucât aceste date sunt vizibile în procesul de build, pentru detaliile mai confidențiale putem folosi `Secrete`, astfel acestea vor fi adăugate către Pod-uri printr-o declarație specifică și nu sunt la fel de expuse. `Secretele` pot fi create strict prin comanda `kubect1 create secret` și vor fi expuse Pod-urilor care au acces la acestea ca la un fișier însă acestea sunt stocate defapt în memoria RAM pentru a evita scrierea acestora pe disc.

Pentru aceste tipuri de obiecte avem operațiunile uzuale de create, read și delete iar în cazul operațiunilor de actualizare, pod-urile vor fi notificate de schimbarea acestora însă rămâne la latitudinea aplicațiilor ce rulează în interior dacă beneficiază de această actualizare.

Capitolul 5

Îmbunătățirea metodelor de dezvoltare software

5.1 Schimbari in livrarea produselor

5.2 DevOps și aspecte ale comunitații

5.3 Ce înseamnă să aplici DevOps

5.4 Unelte

Capitolul 6

Aplicație

Capitolul 7

Concluzii

Bibliografie

- [1] John Biggs și Vicente Herrera García, *Building intelligent cloud applications: Develop scalable models using serverless architectures with Azure*, O'Reilly Media, Incorporated, 2019.
- [2] Brendan Burns, Kelsey Hightower și Joe Beda, *Kubernetes: Up and running: Dive into the future of Infrastructure*, O'Reilly Media, 2019.
- [3] Bahrim Dragoș, „Temă „Principii SOLID și CI/CD în Microservicii” pentru cursul Metode de dezvoltare software”, în (Iunie, 2022).
- [4] Albert Endres și H. Dieter Rombach, *A Handbook of Software and Systems Engineering: Empirical observations, laws, and theories*, Pearson/Addison Wesley, 2003.
- [5] Eric Evans, *Domain-Driven Design: Tackling Complexity in the Heart of Software*, Addison-Wesley Professional, August 20, 2003.
- [6] Hector Garcia-Molina și Kenneth Salem, „Sagas”, în *Proceedings of the Association for Computing Machinery Special Interest Group on Management of Data 1987 Annual Conference, San Francisco, CA, USA, May 27-29, 1987*, ed. de Umeshwar Dayal și Irving L. Traiger, ACM Press, 1987, pp. 249–259, DOI: [10.1145/38713.38742](https://doi.org/10.1145/38713.38742), URL: <https://doi.org/10.1145/38713.38742>.
- [7] Thomas Hamilton, *What is software testing? definition*, Aug. 2022, URL: <https://www.guru99.com/software-testing-introduction-importance.html>.
- [8] Troy Hunt, *Passwords evolved: Authentication guidance for the modern era*, Aug. 2017, URL: <https://www.troyhunt.com/passwords-evolved-authentication-guidance-for-the-modern-era/>.
- [9] Rahul Kumar, *AWS Market Share 2022: How far it rules the cloud industry?*, Sept. 2022, URL: <https://www.wpoven.com/blog/aws-market-share/>.
- [10] Richard Leadbetter, *Tech interview: Titanfall on xbox 360*, Apr. 2014, URL: <https://www.eurogamer.net/digitalfoundry-2014-titanfall-xbox-360-tech-interview>.
- [11] Sam Newman, *Building microservices: Designing fine-grained systems*, O'Reilly, 2021.

- [12] David Lorge Parnas, „Information Distribution Aspects of Design Methodology”, în *Information Processing, Proceedings of IFIP Congress 1971, Volume 1 - Foundations and Systems, Ljubljana, Yugoslavia, August 23-28, 1971*, ed. de Charles V. Freiman, John E. Griffith și Jack L. Rosenfeld, North-Holland, 1971.
- [13] Timothy J. Seppala, *A closer look at Titanfall's not-so-secret weapon: Microsoft's cloud*, Mar. 2014, URL: <https://www.engadget.com/2014-03-10-titanfall-cloud-explained.html>.
- [14] James Turnbull, *The docker book*, s.n., 2014.
- [15] Timothy L. Warner, *Microsoft Azure, For Dummies*, a Wiley Brand, 2020.