

České vysoké učení technické v Praze
Fakulta elektrotechnická
Otvorená informatika



Diplomová práce

Veľkoobjemové úložisko emailov

Bc. Patrik Lenárt

Vedúci práce: Ing. Jan Šedivý, CSc.

Študijný program: Otvorená informatika, Magisterský

Odbor: Softwarové inžinierstvo

11. mája 2011

Pod'akovanie

Rád by som poďakoval vedúcemu práce pánovi Ing. Janovi Šedivému, CSc. za konzultácie, cenné rady, pripomienky a návrhy, ktoré mi ochotne poskytol počas vypracovávania tejto práce. Tak isto sa chcem poďakovať svojim najbližším, bez ktorých podpory by táto práca nevznikla.

Prehlásenie

Prehlasujem, že som svoju diplomovú prácu vypracoval samostatne a použil som iba podklady uvedené v priloženom zozname.

Nemám závažný dôvod proti užitiu tohto školského diela v zmysle §60 Zákona č. 121/2000 Sb., o autorskom práve, o právach súvisiacich s právom autorským a o zmene niektorých zákonov (autorský zákon).

V Prahe dňa 1.3.2011

.....

Abstract

Abstrakt

Obsah

1	Úvod	1
2	Databázové systémy	3
2.1	História	3
2.2	ACID	4
2.3	Škálovanie databázového systému	5
2.3.1	Replikácia	6
2.3.2	Rozdeľovanie dát	7
2.4	BASE	8
2.5	CAP	9
2.5.1	Konzistencia verzus dostupnosť	9
2.6	Čiastočná konzistencia	10
2.6.1	Konzistencia z pohľadu klienta	11
2.6.2	Systémová konzistencia	12
2.7	MapReduce	12
2.7.1	Architektúra	13
2.7.2	Použitie	13
3	Definícia problému	15
3.1	Archivácia elektronickej pošty	15
3.2	Požiadavky na systém	16
3.2.1	Funkčné požiadavky	16
3.2.2	Nefunkčné požiadavky	18
4	NoSQL	21
4.1	Typy NoSQL databázových systémov	21
4.1.1	Kľúč-hodnota	21
4.1.2	Stĺpcovo orientovaný model	22
4.1.3	Dokumentový model	23
4.1.4	Grafový model	23
4.2	Porovnanie NoSQL systémov	23
4.2.1	Dátový a dotazovací model	24
4.2.2	Škálovateľnosť a schopnosť odolávať chybám	24
4.2.3	Elasticnosť	24
4.2.4	Konzistencia dát	25

4.2.5	Prostredie behu	25
4.2.6	Bezpečnosť	26
4.3	Výber NoSQL systémov	26
5	Cassandra	29
5.1	Dátový model	29
5.2	Rozdeľovanie dát	30
5.3	Replikácia	31
5.4	Členstvo uzlov v systéme	31
5.5	Zápis dát	32
5.6	Čítanie dát	32
5.7	Zmazanie dát	32
5.8	Konzistencia	32
5.9	Perzistentné úložisko	33
5.10	Bezpečnosť	33
6	HBase	35
6.1	Dátový model	36
6.2	Architektúra systému	37
6.3	Rozdeľovanie dát	37
6.4	Replikácia	37
6.5	Perzistentné úložisko	38
6.6	Konzistencia	38
6.7	Zápis dát a čítanie dát	38
6.8	Zmazanie dát	38
6.9	Bezpečnosť	39
7	Testovanie výkonnosti	41
7.1	Testovacie prostredie	41
7.2	Popis testovacej metodológie	42
7.2.1	Testovací klient	42
7.2.2	Testovací prípad pre zápis dát	42
7.2.3	Testovací prípad pre čítanie dát	42
7.2.4	Zaťažovací test	43
7.3	HDFS	43
7.4	HBase	44
7.5	Cassandra	45
7.6	Voľba databázového systému	48
8	Návrh systému	49
8.1	Zdroj dát	49
8.2	Analýza dát	50
8.3	Databázová schéma	50
8.4	Fultextové vyhľadávanie	51
8.5	Implementácia	52
8.5.1	Klient	53

8.5.2	Výpočet štatistík	54
8.5.3	Webové rozhranie	54
8.6	Overenie návrhu	55
8.6.1	Doporučenie najvhodnejšieho systému	56
9	Záver	57
	Literatúra	59
A	Zoznam použitých skratiek	63
B	Inštalčná a užívateľská príručka	65
B.0.2	Inštalácia simulátoru OMNeT++ pre platformu Linux	65
B.0.3	Inštalácia mnou modifikovaného Mobility Frameworku	65
B.0.4	Práca s modelom IEEE 802.15.4	66
C	Obsah priloženého CD	67

Zoznam obrázkov

3.1	Približná distribúcia emailových správ. MEAN = 301 kB, std = 1.3 MB . . .	17
4.1	Pozícia dátového modelu z pohľadu škálovania podľa veľkosti a komplexnosti. Zdroj: Neo4J a NOSQL overview and the benefits of graph databases, Emil Eifrem, prezentácia, TODO link.	25
4.2	Rozdelenie databázových systémov podľa CAP	26
8.1	Obsah obálky z programu qmail-scanner	50
8.2	Databázová schéma	52
8.3	JSON schéma pre fultextové vyhľadávanie	53
8.4	Architektúra Celery, Zdroj: [online], http://ask.github.com/celery/getting-started/introduction.h	
8.5	Spracovanie emailu	55
8.6	Programová ukážka v jazyku Pig	55

Zoznam tabuliek

4.1	Stručný prehľad vlasností stĺpcovo orientovaných systémov NoSQL	27
7.1	Priepustnosť pri zápise dát na HDFS	43
7.2	Hbase: zápis riadkov o veľkosti 1000 B	44
7.3	HBase: čítanie riadkov o veľkosti 1000 B	45
7.4	Hbase: maximálna priepustnosť klastru v MB/s	45
7.5	Zápis riadkov o veľkosti 1000 B	47
7.6	Čítanie riadkov o veľkosti 1000 B	47
7.7	Cassandra: maximálna priepustnosť klastru v MB/s	47

Kapitola 1

Úvod

S neustálym rozvojom informačných technológií súčasne narastá objem informácií, ktoré je potrebné spracúvať. Tento fakt podnietil vznik databázových systémov, ktoré slúžia na organizáciu, uchovávanie a prácu s veľkým objemom dát. V dnešnej dobe existuje množstvo databázových systémov, ktoré sa navzájom líšia svojou architektúrou, dátovým modelom, výrobcom atď.

Od začiatku sedemdesiatych rokov 20. storočia sú v tejto oblasti dominantou relačné databázové systémy (angl. Relational Database Management Systems). Z dôvodu pretrvávajúceho rozvoja internetových technológií a rapídneho rastu dát v digitálnom univerze [20] začínajú byť tieto systémy nepostačujúce. Medzi hlavné faktory pre výber relačného databázového systému doposiaľ patrili výrobca, cena a pod. V dnešnej dobe so vznikom moderných webových aplikácií ako sociálne siete a neustále rastúcim objemom dát, ktorý spracúvajú dátové sklady alebo analytické aplikácie požadujeme od týchto systémov vlastnosti ako vysoká dostupnosť, horizontálna rozširiteľnosť a schopnosť pracovať s obrovským objemom dát (PB). Novo vznikajúce databázové systémy, spĺňajúce tieto požiadavky sa spoločne označujú pod názvom NoSQL (Not Only SQL). Pri ich výbere je v tomto prípade dôležité porozumenie architektúry, dátového modelu a dát, s ktoré tieto systémy spracúvajú.

Táto práca si kladie za cieľ viacero úloh, ktorými sú pochopenie a popis základných konceptov, ktoré tieto systémy využívajú, určenie kritérií vďaka ktorým môžeme tieto systémy navzájom porovnávať. Ďalej je úlohou analyzovať a popísať požiadavky pre systém veľkoobjemového úložiska elektronickej pošty, ktorý bude schopný spracovávať milióny emailov. Poslednou úlohou je na základe našich požiadavkov vybrať, čo najlepšie odpovedajúci NoSQL systém a s jeho použitím implementovať prototyp aplikácie.

Osnova

...

Kapitola 2

Databázové systémy

V tejto časti stručne popíšeme históriu vzniku databázových systémov, identifikujeme problémy, ktoré nastávajú pri tvorbe distribuovaných databázových systémov a popíšeme možné spôsoby ich riešenia. Ďalej uvedieme základné koncepty využívané pri tvorbe distribuovaných databázových systémov a techniku MapReduce, ktorá slúži na spracovanie veľkého objemu dát (PB) uloženým v systémoch NoSQL.

2.1 História

V polovici šesťdesiatych rokov 20. storočia bol spoločnosťou IBM vytvorený informačný systém IMS (Information Management System), využívajúci hierarchický databázový model. IMS je po rokoch vývoja využívaný dodnes. Po krátkej dobe, v roku 1970, publikoval zamestnanec IBM, Dr. Edgar F. Codd článok pod názvom „A Relational Model of Data for Large Shared Data Banks“ [10], ktorým uviedol relačný databázový model. Prvým databázovým systémom implementujúcim tento model bol System R od IBM. Systém používal jazyk pod názvom SEQUEL, ktorý je predchodca dnešného SQL (Structured Query Language) slúžiaceho na manipuláciu a definíciu dát v relačných databázových systémoch. Tento koncept sa stal základom pre relačné databázové systémy, ktoré vďaka širokej škále vlastností (ako napríklad podpora transakcií a dotazovací jazyk SQL) patria v dnešnej dobe medzi najpoužívanejšie riešenia na trhu.

V minulosti boli objem dát, s ktorým tieto systémy pracovali menší a výkon hardvéru mnohonásobne nižší. Dnes napriek tomu, že výkon procesorov a veľkosť pamäťových zariadení rapídne stúpa, je najväčšou slabinou počítačových systémov rýchlosť prenosu dát medzi pevným diskom a operačnou pamäťou. Tento fakt je kritický pre novovznikajúce webové aplikácie ako napríklad sociálne siete alebo systémy využívajúce *cloud computing*, ktoré majú neustále vyššie nároky na spracovávaný objem dát v reálnom čase a vyžadujú podporu škálovania, ktorá zabezpečuje vysokú dostupnosť a spoľahlivosť. Tieto požiadavky sa snažia efektívne riešiť novovznikajúce distribuované systémy pod spoločným názvom NoSQL, ktoré sú popísané štvrtej kapitole tejto práce 4.

2.2 ACID

Relačné databázové systémy poskytujú veľkú množinu operácií, ktoré je možné vykonávať nad dátami v nich uloženými. Transakcie [24][25] sú zodpovedné za korektné vykonanie operácií v prípade, že spĺňajú množinu vlastností ACID. Význam jednotlivých vlastností akronymu ACID je nasledovný:

- *Atomicita* (Atomicity) - zaisťuje, že sa vykonajú všetky operácie reprezentujúce transakciu, čo spôsobí korektný prechod systému do nového stavu. V prípade zlyhania transakcie nemá daná operácia žiaden vplyv na výsledný stav systému a prechod do nového stavu sa nevykoná.
- *Konzistencia* (Consistency) - každá transakcia po svojom úspešnom ukončení garantuje korektnosť svojho výsledku a zabezpečí, že systém prejde z jedného konzistentného stavu do druhého. Konzistentný stav zaručuje, že dáta v systéme odpovedajú požadovanej hodnote. Systém sa musí nachádzať v konzistentnom stave aj v prípade zlyhania transakcie.
- *Izolácia* (Isolation) - operácie, ktoré prebiehajú počas vykonávania jednej transakcie nie sú viditeľné ostatným. Operácie tvoriace transakciu musia mať konzistentný prístup k dátam a to aj v prípade, že u inej transakcie dôjde k jej zlyhaniu.
- *Trvácnosť* (Durability) - v prípade, že bola transakcia úspešne ukončená, systém musí garantovať trvácnosť jej výsledku aj v prípade svojho zlyhania.

Implementácia vlastností ACID, ktoré zaručujú konzistenciu, zvyčajne využíva u relačných databázových systémov metódu zamykania. Transakcia uzamkne dáta pred ich spracovaním a spôsobí ich nedostupnosť až do jej úspešného ukončenia, poprípade zlyhania. Transakcie sú vykonávané sekvenčne. Pre databázový systém, od ktorého požadujeme vysokú dostupnosť tento model nie je vyhovujúci. Zámky spôsobujú stavy, kedy ostatné transakcie musia čakať na ich uvoľnenie. Náhradou je mechanizmus pod názvom „súčasná správa viacerých verzií“ (Multiversion concurrency control), ktorý umožňuje paralelné vykonávanie operácií nad dátami, jeho popis obsahuje práca od P. Bernsteina a N. Goodmana [5]. Tento mechanizmus je zároveň využívaný systémami NoSQL.

Transakcie spĺňajúce vlastnosti ACID využívajú v distribuovaných databázových systémoch¹ dvojfázový potvrdzovací protokol (Two-phase commit protocol [6]). Systém využívajúci tento protokol, zaručuje konzistentnosť a je schopný odolávať čiastočným poruchám na sieti alebo v systéme. Vlastnosti ACID nekladú žiadnu záruku na dostupnosť systému, naopak nedostupnosť je uprednostnená v prípade operácie, ktorá by mohla spôsobiť nekonzistenciu dát. Takéto systémy sú vhodné pre aplikácie, v ktorých sa vykonávajú platobné operácie a pod. Existuje množstvo aplikácií, u ktorých sa uprednostňuje dostupnosť pred konzistenciou. Pri tvorbe distribuovaných databázových systémov je preto potrebné upustiť z niektorých ACID vlastností, čo spôsobilo vznik nového modelu pod názvom BASE 2.4.

¹Distribuované databázové systémy sú tvorené pomocou viacerých samostatne operujúcich databázových systémov, ktoré nazývame uzly a môžu komunikovať pomocou siete. Užívateľovi alebo aplikácii sa javia ako jeden celok [ref].

2.3 Škálovanie databázového systému

Obecná definícia pojmu *škálovateľnosť* [7] je náročná bez vymedzenia kontextu, ku ktorému sa vzťahuje. V tejto kapitole budeme škálovateľnosť chápať v kontexte webových aplikácií, ktorých dynamický vývoj kladie na databázové systémy viacero požiadaviek. Medzi hlavné z nich patrí neustála potreba zvyšovania diskového priestoru a teda zvyšovanie veľkosti databáze alebo schopnosť obslúžiť čoraz vyšší počet užívateľov aplikácie (zvýšenie počtu operácií pre čítanie a zápis do databázového systému). V tomto prípade pod pojmom škálovateľnosť databázového systému rozumieme vlastnosť, vďaka ktorej je systém schopný spracúvať narastajúce požiadavky webovej aplikácie v definovanom časovom intervale. Typicky sa táto vlastnosť realizuje pridaním nových uzlov do aktuálneho systému, čo má za následok potrebu distribuovaného databázového systému.

Škálovateľnosť delíme na vertikálnu a horizontálnu. Táto metóda dodáva systému nasledujúce vlastnosti [27]:

- umožňuje zväčšiť veľkosť celkovej kapacity databáze a táto zmena by mala byť transparentná z pohľadu aplikácie na dáta
- zvyšuje celkové množstvo operácií, pre čítanie a zápis dát, ktoré je systém schopný vykonať v danú časovú jednotku
- v určitých prípadoch môže zaručiť, že systém neobsahuje jednotku, ktorá by v prípade zlyhania spôsobila nedostupnosť celého systému (Single point of failure)

Vertikálna škálovateľnosť je metóda, ktorá sa aplikuje pomocou zvyšovania výkonnosti hardvéru, tj. do systému sa pridáva operačná pamäť, rýchlejšie viacjádrové procesory, zvyšuje sa kapacita diskov. Jednou z nevýhod tohoto riešenia je jeho vysoká cena a možná nedostupnosť systému v prípade jeho zlyhania. Proces vertikálneho škálovania relačnej databáze obsahuje nasledujúce kroky:

- zámena hardvéru za výkonnejší
- úprava súborového systému (napr. zrušenie žurnálu a uchovávanie informácie o poslednom prístupe k súborom)
- optimalizácia databázových dotazov, indexovanie
- pridanie vrstvy pre kešovanie (memcached, EHCACHE, atď.)
- denormalizácia dát v databáze, porušenie normalizácie

V tomto prípade je možné naraziť na výkonnostné hranice bežne dostupného hardvéru a na rad nastupuje horizontálna škálovateľnosť, ktorá je omnoho komplexnejšia. Horizontálnu škálovateľnosť je možné realizovať pomocou replikácie alebo metódou „rozdeľovania dát“ (sharding).

2.3.1 Replikácia

V distribuovaných systémoch sa pod pojmom replikácia rozumie vlastnosť, ktorá má za následok, že sa daná informácia nachádza na viacerých uzloch² tohto systému. Táto vlastnosť zvyšuje dostupnosť, spoľahlivosť a odolnosť systému voči chybám. Replikácia nie je určená pre zálohu dát.

V prípade distribuovaného databázového systému sa časť informácií uložených v databáze nachádza na viacerých uzloch. Toto usporiadanie môže napríklad zvýšiť výkonnosť operácií, ktoré pristupujú k dátam a to tak, že dochádza k čítaniu dát z databázy paralelne z viacerých uzlov. V systéme obsahujúcom repliku dát nedochádza k strate informácií v prípade poruchy uzlu. Replikácia a propagácia zmien (pridanie alebo odstránenie uzlu s replikou) v systéme sú z pohľadu aplikácie transparentné. Metóda replikácie nezvyšuje pridávaním nových uzlov celkovú kapacitu databázy. Problémom tejto techniky je konzistencia dát. Dáta sa zapisujú na viacero fyzicky oddelených uzlov a zmena sa nemusí prejavíť okamžite vo všetkých replikách. Medzi metódy pomoci, ktorých sa realizuje replikácia patria napríklad:

- *Read one - Write all*, u tejto metódy sa čítanie dát prevedie z ľubovôležného uzlu obsahujúceho repliku. Zápis dát sa vykoná na všetky uzly s replikou a až v prípade, že každý z nich potvrdí úspech tejto operácie je výsledok považovaný za korektný. Táto metóda nie je schopná pracovať v prípade ak dôjde k prerušeniu sieťového spojenia medzi uzlami (network partitioning) alebo v prípade poruchy jedného z uzlov.
- *Read all - Write one* metóda zabezpečí zápis dát na jednu repliku a tieto dáta sa na ostatné uzly propagujú asynchrónne. Tento spôsob zanáša do distribuovaného systému nekonzistenciu.

Výber metódy replikácie určuje dvojicu vlastností distribuovaného databázového systému a to dostupnosť a konzistenciu. Podľa teórie s názvom CAP 2.5 nie je možné, aby systém súčasne poskytoval obe vlastnosti v prípade, že medzi jeho uzlami môže dôjsť k chybe v sieťovej komunikácii.

V relačných databázových systémoch sa replikácia rieši pomocou architektúry *master - slave*. Uzol pod názvom master slúži ako jediný databázový stroj, na ktorom sa vykonáva zápis dát a replika týchto dát je následne distribuovaná na zvyšné uzly pod názvom slave. Táto metóda umožňuje mnohonásobne zvýšiť počet operácií, ktoré slúžia pre čítanie dát z databázového systému a v prípade zlyhania niektorého zo systémov máme neustále k dispozícii kópiu dát. Slabinou v tomto systéme je uzol v roli master, ktorý nezvyšuje výkonnosť v prípade operácií vykonávaných zápis a zároveň jeho porucha môže spôsobiť celkovú nedostupnosť systému.

Druhým možným riešením je technika *multi - master*, kde každý uzol obsahujúci repliku je schopný zápisu dát a následne tieto zmeny preposiela ostatným. Tento mechanizmus predpokladá distribuovanú správu zamykania a vyžaduje algoritmy pre riešenie konfliktov spôsobujúcich nekonzistenciu dát.

²Pod pojmom uzol v tomto prípade myslíme samostatný počítačový systém, ktorý je súčasťou distribuovaného systému

2.3.2 Rozdeľovanie dát

Rozdeľovanie dát (sharding) je metóda založená na princípe, kde dáta obsiahnuté v databáze rozdeľujeme podľa stanovených pravidiel do menších celkov. Tieto celky môžeme následne umiestniť na navzájom rôzne uzly distribuovaného databázového systému. Táto metóda umožňuje zvýšiť výkonnosť operácií pre zápis a čítanie dát a zároveň pridávaním nových uzlov do systému zvyšuje celkovú kapacitu databázy. V prípade, že architektúra distribuovaného databázového systému využíva túto techniku, zvýšenie výkonu jeho operácií a objemu uložených dát sa realizuje automaticky bez nutnosti zásahu do aplikácie.

Techniku rozdeľovania dát môžeme považovať za aplikáciu architektúry známej pod názvom „zdieľanie ničoho“ (shared nothing) [34]. Táto architektúra sa používa pre návrh systémov využívajúcich multiprocesory. V takomto prípade sa medzi procesormi nezdieľa operačná ani disková pamäť. Architektúra zabezpečuje takmer neobmedzenú škálovateľnosť systému a využíva ju mnoho NoSQL systémov ako napríklad Google Bigtable [9], Amazon Dynamo [14] alebo technológia MapReduce [12].

Pri návrhu distribuovaných databázových systémov s využitím tejto techniky patrí medzi kľúčový problém implementácia funkcie spojenia (JOIN) nad dátami, ktorá sa preto neimplementuje. V prípade, že sa, dáta nad ktorými by sme chceli túto operáciu vykonať, nachádzajú na dvoch rozdielnych uzloch prepojených sieťou, takéto spojenie by značne znížilo celkovú výkonnosť systému a viedlo by k zvýšeniu sieťového toku, záťažou systémových zdrojov a možným nekonzistenciám.

Keďže sa dáta nachádzajú na viacerých uzloch systému, hrozí zvýšená pravdepodobnosť hardverového zlyhania, poprípade prerušenie sieťového spojenia a preto sa táto technika často kombinuje s pomocou využitia replikácie.

V prípade použitia tejto techniky v relačných databázach, je nutný zásah do logiky aplikácie. Dáta uložené v tabuľkách relačnej databázy zachytávajú vzájomné relácie. Týmto spôsobom dochádza k celkovému narušeniu tohto konceptu. Príkladom môže byť tabuľka obsahujúca zoznam zamestnancov, ktorú rozdelíme na samostatné celky. Každá tabuľka bude reprezentovať mená zamestnancov, ktorých priezvisko začína rovnakým písmenom abecedy a zároveň sa bude nachádzať na samostatnom databázovom systéme. Táto technika so sebou prináša problém, v ktorom je potrebné nájsť vhodný kľúč, podľa ktorého budeme dáta rozsekať a zabezpečíme tak rovnomerné zaťaženie uzlov daného systému. Existuje viacero metód [27]:

- segmentácia dát podľa funkcionality - dáta, ktoré je možné popísať spoločnou vlastnosťou ukladáme do samostatných databáz a tieto umiestňujeme na rozdielne uzly systému. Príkladom môže byť samostatný uzol spravujúci databázu pre užívateľov a iný uzol s databázou pre produkty. Túto metódu spracoval Randy Shoup³ [33], architekt spoločnosti eBay.
- rozdeľovanie dát podľa kľúča - v dátach identifikujeme kľúč, pomocou ktorého je možné ich rovnomerne rozdeliť. Následne sa na tento kľúč aplikuje hašovacia funkcia a na základe jej výsledku sa tieto dáta umiestňujú na jednotlivé uzly.

³“If you can't split, you cant scale it.” – Randy Shoup, Distinguished architect Ebay

- vyhľadávacia tabuľka - jeden uzol v systéme slúži ako katalóg, ktorý určuje, na ktorom uzle sa nachádzajú dané dáta. Tento uzol zároveň spôsobuje zníženie výkonu a v prípade jeho havárie spôsobuje nedostupnosť celého systému (SPOF).

Replikácia a rozdeľovanie dát patria medzi kľúčové vlastnosti využívané v NoSQL systémoch, ktoré popisuje kapitola 4.

2.4 BASE

Akronym BASE [15] bol prvýkrát použitý v roku 1997 na sympóziu SOSP (ACM Symposium on Operating Systems Principles). BASE tvoria nasledujúce slovné spojenia:

- „*bežne dostupný*“ (Basically Available) - systém je schopný zvládať svoje čiastočné zlyhanie za cenu nižšej komplexity.
- „*zmiernený stav*“ (Soft State) - systém nezaručuje trvácnosť dát s cieľom zvýšenia výkonu.
- „*čiasťočne konzistentný*“ (Eventually Consistent) - je možné na určitú dobu tolerovať nekorektnosť dát, ktoré musia byť po určitom časovom intervale konzistentné.

Tento model poľavil na požiadavku zodpovednom za konzistenciu dát, jeho aplikáciou v distribuovanom databázovom systéme sa dosahuje vyššia dostupnosť aj v prípade čiastočného zlyhania alebo sieťového prerušenia. Každý distribuovaný databázový systém môžeme klasifikovať ako systém spĺňajúci vlastnosti ACID, BASE alebo oboje.

BASE umožňuje horizontálne škálovanie relačných databázových systémov bez nutnosti použitia distribuovaných transakcií. Pre implementáciu tejto techniky môžeme použiť rozdeľovanie dát metódou segmentácie dát podľa funkcionality [31].

Bankomatový systém je príkladom systému obsahujúceho čiastočnú konzistenciu dát. Po vybratí určitej čiastky z účtu, sa korektná informácia o aktuálnom zostatku môže zobrazíť až za niekoľko dní, kdežto transakcia ktorá túto zmenu vykonala musí spĺňať vlastnosti ACID. Medzi webové aplikácie, u ktorých sa nepožadujú všetky vlastnosti ACID patria napríklad nákupný košík spoločnosti Amazon, zobrazovanie časovej osi aplikácie Twitter, poprípade systémy spoločnosti Google indexujúce web. Ich nedostupnosť by znamenala obrovské finančné straty (napríklad zlyhanie vyhľadávania pomocou systému Google⁴ by znamenalo zobrazenie nižšieho počtu reklám, nedostupnosť nákupného košíka Amazon⁵ by spôsobila pokles predaja atp).

Aplikácia vyššie popísaných techník na relačné databázové systémy môže byť netriviálnou úlohou. Relačný model, je spôsob reprezentácie dát, ktorý umožňuje efektívne riešiť určité typy úloh, preto snaha prispôbiť tento model každému problému je nezmyselná. V tomto prípade, môžeme uvažovať alternatívne riešenia, medzi ktoré patria systémy NoSQL.

⁴<http://www.google.com>

⁵<http://www.amazon.com>

2.5 CAP

Moderné webové aplikácie kladú na systémy požiadavky, medzi ktoré patrí vysoká dostupnosť, konzistencia dát a schopnosť odolávať chybám. Dr. Brewer v roku 2000 nastolil myšlienku dnes známu pod názvom teória CAP [8], ktorá tvrdí že je možné súčasne dosiahnuť len dvojicu z týchto vlastností. V roku 2002 platnosť tejto teórie pre asynchrónnu sieť matematicky dokázali Lynch a Gilbert [23]. Modelu asynchrónnej siete odpovedá svojimi vlastnosťami sieť Internet. Akronym CAP tvoria nasledujúce vlastnosti:

- *Konzistencia* (Consistency) - distribuovaný systém je v konzistentnom stave, ak každý jeho uzol v prípade požiadavku dát vracia tú istú odpoveď.
- *Tolerancia chýb* (Partition Tolerance) - uzly distribuovaného systému navzájom komunikujú pomocou siete, v ktorej hrozí strata správ. V prípade vzniku sieťového prerušenia dané uzly medzi sebou navzájom nedokážu komunikovať. Táto vlastnosť podľa definície (viď. Gilbert a Lynch) tvrdí, že v prípade vzniku zlyhania sieťovej komunikácie medzi niektorými uzlami, musí byť systém schopný naďalej pracovať korektne. V reálnych podmienkach neexistuje distribuovaný systém, ktorého uzly na vzájomnú komunikáciu využívajú sieť a nedochádza pri tom k strate správ, teda k poruchám sieťovej komunikácie.
- *Dostupnosť* (Availability) - distribuovaný systém je dostupný, ak každý jeho uzol, ktorý pracuje korektne, je schopný pri prijatí požiadavku zaslať odpoveď. V spojení s toleranciou chýb, tato vlastnosť hovorí, že v prípade ak nastane sieťový problém⁶, každá požiadavka bude vykonaná.

Pravdepodobnosť, že dôjde k zlyhaniu ľubovoľného uzla v distribuovanom systéme, exponenciálne narastá s počtom pribúdajúcich uzlov.

$$P(A) = 1 - P(B)^{\text{počet uzlov}}$$

$P(A)$ - pravdepodobnosť zlyhania ľubovoľného uzlu

$P(B)$ - pravdepodobnosť, že individuálny uzol nezlyhá

2.5.1 Konzistencia verus dostupnosť

V distribuovanom systéme nie je možné súčasne zaručiť vlastnosť konzistencie a dostupnosti. Ako príklad si predstavme distribuovaný systém obsahujúci tri uzly A, B, C, ktorý zaručuje obe vlastnosti aj v prípade sieťového prerušenia. Na všetkých uzloch sa nachádzajú identické (replikované) dáta. Ďalej uvažujme, že došlo k sieťovému prerušeniu, ktoré rozdelilo uzly na dva samostatné celky {A,B} a {C}. V prípade, že uzol C obdrží požiadavku pre zmenu dát má na výber dve možnosti:

1. vykonať zmenu dát čo spôsobí, že sa uzly A a B o tejto zmene dozvedia až vo chvíli ak bude sieťové prerušenie odstránené

⁶týmto sa nemyslí porucha uzla

2. zamietnuť požiadavok na zmenu dát, z dôvodu že uzly A a B sa o tejto zmene nedozvedia

V prípade výberu možnosti číslo 1 zabezpečíme neustálu dostupnosť systému naopak v prípade možnosti číslo 2 jeho konzistenciu. Nie je možný súčasný výber oboch riešení.

CP

Ak od daného systému tolerujúceho sieťové prerušenia požadujeme konzistenciu na úkor dostupnosti jedná sa o alternatívu CP. Takýto systém zabezpečí konzistentnosť operácií pre zápis a čítanie dát a zároveň sa môže stať, že na určité požiadavky nebude schopný odpovedať (možnosť číslo 2). Medzi takéto systémy môžeme zaradiť distribuovaný databázový systém využívajúci dvojfázový potvrdzovací protokol (2PC).

AP

V prípade, že poľavíme na požiadavku konzistencie tak takýto systém bude vždy dostupný aj napriek sieťovým prerušeniam. V tomto prípade sa jedná o model AP. Je možné, že v takomto systéme bude dochádzať ku konfliktným zápisom alebo operácie čítania budú po určitú dobu vracaať nekonzistentné výsledky. Tieto problémy s konzistenciou sa v distribuovaných databázových systémoch riešia napríklad pomocou metódy „vektorové hodiny“ (Vector clock) [?] alebo na aplikačnej úrovni na strane klienta. Príkladom systému patriaceho do tejto kategórie je Amazon Dynamo.

CA

AK systém nebude schopný zvládať sieťové prerušenia, tak bude spĺňať požiadavok konzistencie a dostupnosti, varianta CA. Jedná sa o nedistribuované systémy pracujúce na jednom fyzickom hardvéri využívajúce databázové transakcie.

Vyššie popísané vlastnosti nám umožnia vhodný výber distribuovaného databázového systému podľa požiadavkov našej aplikácie.

2.6 Čiastočná konzistencia

V ideálnom svete sa pod pojmom konzistencia v distribuovaných systémoch rozumie fakt, že ak sa v systéme vykoná zmena (zápis alebo aktualizácia dát), na všetkých uzloch sa táto zmena prejaví súčasne s rovnakým výsledkom. Konzistencia je zároveň úzko spojená s replikáciou. Väčšina NoSQL systémov poskytuje čiastočnú konzistenciu, poprípade dáva možnosť výberu medzi CP a AP (napríklad systém Cassandra⁷). V nasledujúcej časti popíšeme rôzne typy konzistencie.

V predchádzajúcom texte sme už spomínali, že v dnešnej dobe existuje mnoho aplikácií, u ktorých je možné poľaviť na požiadavku konzistencie. Ak sa určitá zmena prejaví s miernym oneskorením funkčnosť systému nebude v tomto prípade ohrozená. Táto konzistencia nie je

⁷<http://cassandra.apache.org>

totožná s konzistenciou definovanou u vlastností ACID, kde ukončenie transakcie zaručuje, že sa systém nachádza v konzistentom stave. Na konzistenciu sa môžeme pozeráť z dôch pohľadov. Prvým je klientský pohľad na strane zadávateľa problému resp. programátora, ktorý rozhodne aká je závažnosť zmien, ktoré sa budú vykonávať v systéme. Druhý pohľad je systemový, zabezpečuje technické riešenie a implementáciu techník zodpovedných za správu konzistencie v distribuovaných databázových systémoch.

2.6.1 Konzistencia z pohľadu klienta

Pre potrebu nasledujúcich definíc uvažujme distribuovaný databázový systém, ktorý tvorí úložisko dát a tri nezávislé procesy $\{A, B, C\}$, ktoré môžu v danom systéme zmeniť (vykonať zápis) a načítať hodnotu dátovej jednotky. Na základe toho ako jednotlivé procesy pozorujú nezávisle zmeny systému delíme konzistenciu [37] na:

Silná konzistencia (Strong consistency) - proces A vykoná zápis. Po jeho ukončení je nová hodnota dátovej jednotky dostupná všetkým procesom $\{A, B, C\}$, ktoré k nej následne pristúpia (vykonajú operáciu čítania). Túto konzistenciu zabezpečujú transakcie s vlastnosťami ACID.

Slabá konzistencia (Weak consistency) - proces A vykoná zápis novej hodnoty do dátovej jednotky. V takomto prípade systém negarantuje, že následne pristupujúce procesy $\{A, B, C\}$ k tejto jednotke vrátia hodnotu zapísanú procesom A. Definujeme pojem „nekonzistentné okno“, ktoré zabezpečí, že po uplynutí stanovenej časovej doby sa táto nová hodnota dátovej jednotky prejaví vo všetkých procesoch, ktoré k nej pristúpia.

Čiastočná konzistencia (Eventual consistency) - je to špecifická forma slabej konzistencie. V tomto prípade systém garantuje, že ak sa nevykoná žiadna nová zmena hodnoty dátovej jednotky, po určitom čase budú všetky procesy pristupujúce k tejto jednotke schopné vrátiť jej korektnú hodnotu. Tento model má viacero variácií, niektoré z nich popíšeme v nasledujúcej časti textu.

Read-your-write consistency - v prípade, že proces A zapíše novú hodnotu do dátovej jednotky, žiadny z jeho následujúcich prístupov k tejto jednotke nevráti staršiu hodnotu ako naposledy zapísaná.

Session consistency - v tomto prípade pristupuje proces k systému v kontexte relácií. Po dobu trvania relácie platí predchádzajúci typ konzistencie. V prípade zlyhania relácie sa vytvorí nová, v ktorej môže systém vrátiť hodnotu dátovej jednotky, zapísanú pred vznikom predchádzajúcej relácie.

Monotonic read consistency - v prípade, že proces načítal hodnotu dátovej jednotky, tak pri každom následujúcom prístupe nemôže vrátiť predchádzajúcu hodnotu dátovej jednotky.

Tieto typy konzistencie je možné navzájom kombinovať a ich hlavným cieľom je zvýšiť dostupnosť distribuovaného systému na úkor toho, že poľavíme na požiadavkách konzistencie.

Príkladom môže byť asynchrónna replikácia v relačnom databázovom systéme využívajúca architektúru master-slave, ktorá spôsobí, že systém bude čiastočne konzistentný.

2.6.2 Systémová konzistencia

Techniky založené na protokoloch kvóra (Quorum-based protocols [22]) je možné použiť pre zvýšenie dostupnosti a výkonu v distribuovaných databázových systémoch s garanciou silnej alebo čiastočnej konzistencie.

Definujme nasledujúcu terminológiu:

- N - počet uzlov, ktoré obsahujú repliku dát
- W - počet uzlov obsahujúcich repliku, na ktorých sa musí vykonať zápis, aby bola zmena úspešne potvrdená
- R - počet uzlov s replikou, ktoré musia vrátiť hodnotu dátového objektu v prípade operácie čítanie

Platí, že ak $W + R > N$, operácie pre zápis a čítanie sa stále prekrývajú minimálne na jednom uzle, ktorý bude obsahovať aktuálnu hodnotu danej operácie. Tento prípad, zabezpečuje silnú konzistenciu v systéme. V prípade $W + R \leq N$, môže nastať situácia keď predchádzajúca podmienka neplatí a teda daná operácia je čiastočne konzistentná. Rôzna konfigurácia týchto parametrov zabezpečí rozdielnu dostupnosť a výkonnosť distribuovaného systému. Uvažujme nasledujúce príklady, kde $N = 3$.

1. $R = 1$ a $W = N$, v takomto prípade zabezpečíme že systém bude optimalizovaný pre operácie čítania dát. Operácie budú konzistentné, pretože uzol z ktorého dáta čítame sa prekrýva s uzlami na ktorých vykonávame zápis. Nevýhodou tohoto modelu je, že v prípade nedostupnosti jednej repliky nebude možné do systému zapisovať. V prípade systémov, u ktorých požadujeme aby obsluhovali veľký počet požiadavkov pre čítanie sa môže hodnota N pohybovať v stovkách až tisícoch, závisí to od počtu uzlov v systéme.
2. $W = 1$ a $R = N$, tento prípad je vhodný pre systémy u ktorých požadujeme rýchly zápis. Tento model môže spôsobiť stratu dát v prípade, že systém s replikou na ktorú sa vykoná zápis zlyhá.

2.7 MapReduce

Nárast diskových kapacít a množstva dát, ktoré na nich ukladáme spôsobuje jeden z ďalších problémov, ktorým je analýza a spracovanie dát. Kapacita pevných diskov sa za posledné roky mnohonásobne zvýšila v porovnaní s dobou prístupu a prenosových rýchlostí pre čítanie a zápis dát na tieto zariadenia.

Pre jednoduchosť uvažujme nasledujúci príklad, v ktorom chceme spracovať pomocou jedného počítačového systému 1 TB dát uložených na lokálnom súborovom systéme, pri priemernej prenosovej rýchlosti diskových zariadení 100 MB/s. Za ideálnych podmienok by čas na prečítanie týchto dát presahoval dve a pol hodiny. V prípade, že by sme 1 TB dát

rovnomerne rozdelili na sto počítačov a tieto úseky paralelne spracovali, celková doba by sa znížila za ideálnych podmienok na necelé dve minúty.

Spoločnosť Google v roku 2004 zverejnila programovací model pod názvom MapReduce [12], ktorý slúži na paralelne spracovanie obrovského objemu dát (PB). Model využíva vlastnosti paralelných a distribuovaných systémov, je optimalizovaný pre beh na klastri tvorenom vysokým počtom (tisícky) spotrebných počítačov. Pre programátora sa snaží zastreť všetky problémy, ktoré prináša paralelizácia výpočtov, poruchovosť systémov, distribúcia dát vzhľadom na ich lokalitu a rovnomerne rozvhanie záťaže medzi systémami.

Pre použitie tohoto nástroja musí programátor zadať dve funkcie pod názvom *map* a *reduce*. Funkcia *map* na jednotlivých uzloch systému, transformuje vstupné data na základe zadaného kľúča *k1* na medzivýsledok, ktorý obsahuje nové kľúče *g1* až *gn* a k nim odpovedajúce hodnoty. Tieto hodnoty sa zoradia podľa ich príslušnosti ku kľúčom *g1* až *gn* a následne sa odošlú na uzly s funkciou *reduce*. Užívateľom definovaná funkcia *reduce* vykoná nad hodnotami priradenými ku kľúčom *g1* až *gn* danú operáciu, ktorej typickým výsledkom je jedna výsledná hodnota.

```
map(k1, X) -> list(kľúč(z g1,...,gn), Y)
reduce(kľúč(z g1,...,gn), list(Y)) -> list(Z)
```

2.7.1 Architektúra

Obrázok

2.7.2 Použitie

MapReduce nie je vhodný na spracovanie dát v reálnom čase. Je optimalizovaný na dávkový beh. Jeho implementácia spoločnosťou Google, ktorá zároveň využíva distribuovaný súborový systém Google file system (GFS) nie je k dispozícii. V rámci hnutia NoSQL vzniklo open source riešenie pod názvom Hadoop⁸, ktoré implementuje tento model na vlastnom distribuovanom súborovom systéme Hadoop distributed file system (HDFS). K dispozícii sú frameworky HIVE alebo PIG, ktoré sú nadstavbou modelu MapReduce a poskytujú vyššiu abstrakciu vo forme jazyka podobného SQL.

⁸<http://hadoop.apache.org>

Kapitola 3

Definícia problému

Množstvo digitálnych informácií, každým rokom prudko narastá. Podľa štatistík spoločnosti IDC[20] v roku 2006 dosahovala kapacita digitálneho univerza veľkosť 161 exabytov¹(EB). Podiel elektronickej pošty (emailov) bez spamu, tvoril 3% z tohoto objemu. Predpoveď na rok 2011[19] predpokláda celkovú kapacitu 1800 EB, čo je viac ako desaťnásobok nárastu pôvodnej kapacity v období piatich rokov. V rozmedzí rokov 1998 až 2006 sa mal počet schránok elektronickej pošty zvýšiť z 253 miliónov na 1.6 miliardy. Predpoveď IDC ďalej uvádzala, že po ukončení roku 2010 tento počet presiahne hodnotu dvoch miliard. Počas obdobia medzi rokmi 1998 až 2006 celkový počet odoslaných správ elektronickej pošty rástol trikrát rýchlejšie ako počet jej užívateľov, dôvodom tohoto prudkého nárastu bola nevyžiadaná elektronická pošta. Odhaduje sa, že až 85% dát z celkového predpokladaného objemu 1800 EB budú spracovávať, prenášať alebo zabezpečovať organizácie. Napriek tejto explózii digitálnych informácií je potrebné správne porozumieť hodnote týchto dát, nájsť vhodné metódy pre ich ukladanie do pamäti počítačových systémov, ich archiváciu a to tak, aby sme ich mohli ďalej spracovávať a efektívne využiť. Táto kapitola práce si kladie za cieľ analyzovať potreby pre archiváciu elektronickej pošty a definovať požiadavky pre systém slúžiaci k archivácii emailov.

3.1 Archivácia elektronickej pošty

S neustálym nárastom informácií obsiahnutých v digitálnom univerze, sa zároveň zvyšuje objem dát, ktorý reprezentuje elektronickú poшту. Je preto potrebné porozumieť štruktúre emailových správ a následne ich vhodne spracovať. Tieto dáta je potrebné uložiť tak aby sme dosiahli úsporu diskového priestoru, boli sme nad nimi schopný vykonávať operácie ako fultextové vyhľadávanie, zber údajov pre tvorbu štatistík alebo umožnili ich opätovné sprístupnenie. Emaily obsahujú čoraz viac obchodných informácií a iný dôležitý obsah, z tohto dôvodu musia byť organizácie všetkých rozmerov schopné uchovávať tento obsah pomocou vhodných archivačných nástrojov. S problémom archivácie zároveň úzko súvisí problém bezpečnosti. Pod pojmom bezpečnosti v tejto oblasti máme na mysli hlavne ochranu proti nevyžiadanej pošte tj. spam, spyware, malware a phishingu. Na boj proti týmto hrozbám

¹1EB = 10¹⁸B

využívajú organizácie anti-spamové a anti-vírusové systémy. Možné dôvody prečo archivovať elektronickú poštu sú nasledovné[29]:

- záloha dát a ich obnova v prípade havárie systému
- vysoká dostupnosť dát
- sprístupnenie dát koncovému užívateľovi
- splňanie regulačných noriem a zákonov
- ochrana súkromia a e-Discovery
- vyťažovanie dát (angl. data mining)
- efektívne využitie úložného priestoru (deduplikácia príloh)

3.2 Požiadavky na systém

V nasledujúcej časti popíšeme požadované vlastnosti systému, ktorý bude slúžiť na archiváciu veľkého objemu emailových správ. Primárnou požiadavkou na systém je jeho neustála dostupnosť, rozšíriteľnosť a nízkonákladová administrácia. Predpokladané množstvo uložených dát v tomto systéme bude dosahovať desiatky až stovky terabajtov² (TB). Takúto kapacitu dát nie je možné uchovať na bežne dostupnom hardvéri. Dáta uložené v systéme musia byť replikované v prípade vzniku havárie niektorej z jeho časti. Nad uloženými emailovými správami je potrebné vykonávať výpočtovo náročné operácie ako generovanie štatistík a fultextové vyhľadávanie v reálnom čase. Tieto požiadavky prirodzene implikujú využitie distribuovaného databázového systému. Medzi hlavných kandidátov, vďaka ktorým sme tieto požiadavky schopní vyriešiť patria NoSQL databázové systémy, ktorých porovnanie popíšeme v nasledujúcej kapitole.

3.2.1 Funkčné požiadavky

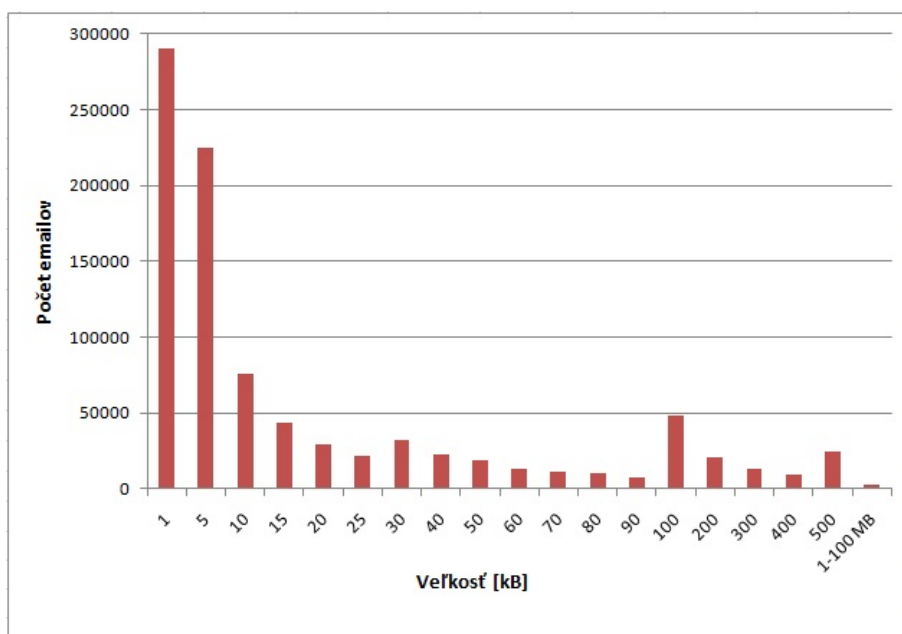
Ukladanie emailov

Základnou jednotkou, ktorú budeme do systému ukladať je emailová správa. Graf na obrázku 3.1 znázorňuje uporiadanie emailov podľa ich veľkosti nad vzorkom približne 1,000,000 emailových správ z reálneho prostredia³. Z daných dát vyplýva, že veľkosť cca 80% emailov je do 50 kB. Tieto údaje sú hrubou aproximáciou a závisia na konkrétnych používateľoch.

Systém musí umožňovať uloženie emailu bez porušenia jeho integrity. Kľúčovým požiadavkom je ukladanie príloh emailov, kde požadujeme aby každá príloha bola jednoznačne identifikovaná a v prípade jej duplicity nebola opakovane uložená v systéme. Dôvodom je vysoká úspora diskového priestoru. Ďalším požiadavkom je automatické zmazanie emailov patriacich do danej domény po uplynutí predom špecifikovanej doby.

²1TB = 10¹²B

³Vzorok emailov pre analýzu bol sprístupnený spoločnosťou Excello.



Obr. 3.1: Približná distribúcia emailových správ. MEAN = 301 kB, std = 1.3 MB

Export emailov

Systém musí umožňovať prístup k ľubovoľnému uloženému emailu v jeho pôvodnej podobe poprípadne skupine všetkých emailov (angl. inbox) patriacej danému užívateľovi.

Vyhľadávanie emailov

Vyhľadávanie je potrebné realizovať nad všetkými emailovými správami uloženými v systéme a jednotlivo nad správami podľa danej domény a nad správami, ktoré prináležia danému užívateľovi. Požadujeme fultextové vyhľadávanie emailov podľa nasledujúcich údajov:

- príjemca emailovej správy
- odosielateľ emailovej správy
- predmet správy
- dátum obsiahnutý v emailovej správe
- identifikátor emailu (MessageID)
- názvy príloh a ich veľkosti
- veľkosť emailu
- vyhľadávanie v tele emailu

Pre administrátorské účely požadujeme vyhľadávanie údajov podľa:

- originálny odosielateľ a príjemca
- IP adresa odosielateľa
- dátum a čas spracovania správy emailovým serverom

Štatistické údaje

Nad uloženými dátami požadujeme výpočet štatistík pomocou využitia MapReduce. Pre emaily patriace do danej domény je potrebné spracovať nasledujúce štatistické ukazatele:

- počet emailov označených príznakom spam
- počet emailov bez príznaku spam
- celková veľkosť emailov pre danú doménu
- veľkosť najväčšieho emailu v doméne
- celková dĺžka filtrácie emailov v danej doméne

Nad celým úložiskom je ďalej potrebné spracovať tieto štatistiky:

- počet všetkých emailov
- počet unikátnych domén
- počet unikátnych príloh

3.2.2 Nefunkčné požiadavky

Dostupnosť

Systém musí byť neustále dostupný (99,9%) a schopný odolávať poruchám v sieťovej komunikácii (angl. network partitions), krátkodobej nedostupnosti uzlov, úplným zlyháním jednotlivých uzlov a umožňovať spracúvať tok pre zápis dát v rozmedzí 10 Mbit až 1 Gbit. Ďalším požiadavkom je aby sa dáta replikovali vo vnútri datacentra na dva uzly a tretia replika bola umiestnená v datacentre, ktoré sa bude nachádzať na geograficky odlišnom mieste. Vyžadujeme aby systém neobsahoval bod, ktorého porucha by spôsobila celkovú nedostupnosť systému (ang. single point of failure). Tento problém rieši vlasnosť decentralizácie, ktorá zabezpečuje, že každá jednotka systému vykonáva rovnakú funkciu a je kedykoľvek plne nahraditeľná.

Rozšíriteľnosť

Predpokladáme použitie bežne dostupného spotrebného hardvéru (angl. commodity hardware), namiesto superpočítačov. Z dôvodu neustalého nárastu objemu elektronickej pošty, musí systém podporovať horizontálne škálovanie, ktoré bude schopné umožňovať zvýšenie celkovej kapacity dátového úložiska (desiatky petabajtov⁴). Pridávanie nových uzlov do systému umožní zvýšiť celkový vypočetný výkon, ktorý sa využije na spracovanie dát pomocou techniky MapReduce. U distribuovaného databázového systému je nutná podpora replikácie, ktorá zvýši výkonnosť operácií pre čítanie a zápis do systému a vďaka nej nebude potrebné riešiť zalohovanie pomocou externých systémov.

Nízkonákladová administrácia

Prevádzkovanie systému a jeho administrácia by mali byť čo najmenej závislé na zásahu ľudských zdrojov. Detekcia nefunkčných uzlov a automatické rozdeľovanie záťaže sa musí vykonávať automaticky. Pridanie poprípadе odobranie nového uzla nesmie ovplyvniť beh celkového systému.

Bezpečnosť

Osoby s oprávnením pre prístup k systému budú schopné operovať s jeho celým obsahom. Predpokladáme beh systému v bezpečnom prostredí a nekladíme žiadne požiadavky na užívateľské role v kontexte prístupu k datam.

Implementačné požiadavky

Cieľom je implementácia systému s využitím dostupných open source technológií.

Z analýzy obecných vlastností relačných databázových systémov v prechádzajúcej kapitole vyplýva, že použitie týchto systémov nie je vhodné pre riešenie zadaného problému. Medzi základné problémy týchto systémov patrí náročné horizontálne škálovanie a možnosť zabezpečenia vysokej dostupnosti. V nasledujúcej kapitole sa budeme zaoberať popisom NoSQL systémov a po ich analýze vyberieme vhodného kandidáta, ktorého použijeme k implementácii prototypu, z dôvodu vysokej komplexnosti riešeného problému.

⁴1PB = 10¹⁵B

Kapitola 4

NoSQL

Názov NoSQL bol prvýkrát použitý v roku 1998 ako názov relačnej databázy [30], ktorej implementácia bola v interpretovaných programovacích jazykoch a neobsahovala dotazovací jazyk SQL. V druhej polovici roku 2009 sa názov NoSQL začal používať v spojení s databázovými systémami, ktoré nepoužívajú tradičný relačný model, dotazovací jazyk SQL, sú schopné horizontálneho škálovania, pracujú na spotrebných počítačových systémoch, vyznačujú sa vysokou dostupnosťou a používajú bezschémový dátový model.

Pôvodným cieľom hnutia NoSQL bolo vytvoriť koncept, pre tvorbu moderných databáz, ktoré by boli schopné spracúvať požiadavky neustále sa rozvíjajúcich webových aplikácií. Filozofiou tých systémov je nesnažiť sa za každú cenu prispôbovať dáta modelu relačnej databázy. Cieľom je vybrať systém, ktorý bude čo najvhodnejšie zodpovedať požiadavkom na uloženie a spracovanie našich dát. NoSQL nepopisuje žiaden konkrétny databázový systém, namiesto toho je to obecný názov pre nerelačné (non-relational) databázové systémy, ktoré disponujú odlišnými vlastnosťami a umožňujú prácu s rôznymi dátovými modelmi. Medzi ich ďalšie znaky patrí napríklad čiastočná konzistencia, schopnosť spracúvať obrovské objemy dát (PB) a jednoduché programové rozhranie (API). Tieto systémy nepodporujú operáciu databázového spojenia a medzi dátami, ktoré do nich ukladáme je možné vytvárať vzájomné závislosti až na aplikačnej vrstve. Pre tieto databázové systémy ďalej platí, že sú distribuované, automaticky poskytujú replikáciu a rozdeľovanie dát. Jedná sa o relatívne mladé systémy, ktoré sú neustále vo vývoji. Ideou tohto konceptu je riešiť spomínané novo vznikajúce problémy a zároveň koexistovať s relačnými databázovými systémami.

4.1 Typy NoSQL databázových systémov

Medzi nerelačné databázové systémy patria dokumentové, grafové, stĺpcové a systémy s dátovým modelom typu „kľúč-hodnota“ (Key-value). V nasledujúcej časti práce ich stručne popíšeme.

4.1.1 Kľúč-hodnota

Relačný databázový model reprezentuje dáta pomocou tabuliek, pre ktoré definujeme ich štruktúru a ktoré sú normalizované aby sme predišli duplikácii a inkonzistencii dát. Pre

zabezpečenie referenčnej integrity jednotlivých entít využívame cudzie kľúče. Tabuľky s popisom názvov stĺpcov a vzťahy medzi nimi nazývame databázovou schémou.

Tento model využíva pre ukladanie dát jednoduchý princíp. Blok dát s ľubovoľnou štruktúrou je v databáze uložený pod názvom kľúča, ktorý môže byť reprezentovaný ako textový reťazec. Databázové systémy využívajúce tento dátový model majú jednoduché programové rozhranie, znázornené na obrázku XY.

```
void Put(string kluc, byte[] data);  
byte[] Get(string key);  
void Remove(string key);
```

Výhodou tohto modelu je, že databázový systém je možné ľahko škálovať. Bohužiaľ v tomto prípade sa o prácu so štruktúrou uložených dát musí starať klient, čo umožňuje dosahovať vysokú výkonnosť na strane databázového systému.

Najväčšou nevýhodou je, že databázový systém nie je schopný medzi uloženými dátami zachytiť ich vzájomné vzťahy (relácie), čo patrí medzi základné požiadavky pri tvorbe komplexných modelov. Úložisko typu kľúč-hodnota nevyužíva normalizáciu dát, dáta sú často duplikované, vzťahy a integrita medzi nimi sa riešia až na aplikačnej úrovni. Pre vkladané dáta a k nim asociované kľúče sa nedefinujú žiadne obmedzenia.

Medzi databázové systémy využívajúce model kľúč-hodnota patria Amazon Dynamo, Tokyo Cabinet, Voldemort, Redis a iné.

4.1.2 Stĺpcovo orientovaný model

Množstvo databázových systémov využíva pre reprezentáciu dát tabuľky, ktoré sú tvorené stĺpcami a popísané schémou (tj. názvy stĺpcov, tabuliek). Každý riadok tabuľky reprezentuje dáta, ktoré sa nazývajú záznamy a tieto sú následne sekvenčne ukladané na pevný disk. Tento model, nazývaný riadkový, je vhodný pre systémy, u ktorých sú dominantné operácie vykonávajúce zápis. Relačné databázové systémy, využívajúce tento model sú teda optimalizované pre zápis. Pre efektívny prístup k dátam je možné použiť indexy.

V dnešnej dobe existuje veľký počet aplikácií, u ktorých prevládajú operácie čítania nad zápisom. Patria sem dátové sklady, systémy pre vyťažovanie dát alebo analytické aplikácie pracujúce s obrovským objemom dát. Pre potreby týchto aplikácií a ich reprezentáciu dát je vhodné použiť stĺpcový model [3][4], ktorý je optimalizovaný pre operácie čítanie dát. Data reprezentujúce stĺpce sú uložené na pevnom disku v samostatných a súvislých blokoch. Načítanie dát do pamäti a následná práca s nimi je efektívnejšia ako u riadkových databáz, kde je potrebné načítať celý záznam obsahujúci aj hodnoty stĺpcov, ktoré sú pre nás v daný moment irelevantné.

Riadkový model obsahuje v jednom zázname dáta z rôznych domén, čo spôsobuje vyššiu entropiu v porovnaní so stĺpcovým modelom, kde sa predpokláda, že dáta v danom stĺpci pochádzajú z totožnej domény a môžu si byť podobné. Táto vlastnosť umožňuje efektívnu komprimáciu dát, ktorá znižuje počet diskových operácií. Nevýdou tohto modelu je zápis dát, ktorý spôsobuje vyššiu záťaž diskových operácií. Optimalizáciou môže byť dávkový zápis.

Tento model sa používa v databázových systémoch ako Google Bigtable, HBase, Hypertable alebo Cassandra.

4.1.3 Dokumentový model

Dokumentové databáze sú založené na modeli typu kľúč-hodnota. Požiadavkom na ukládané dáta je, že musia byť v tvare ktorému rozumie databázový systém. Štruktúra vkládaných dát môže byť popísaná napríklad pomocou XML, JSON, YAML. Táto štruktúra nám následne umožňuje okrem jednoduchého vyhľadávania pomocou kľúč-hodnota vytvárať s využitím indexov zložitejšie dotazy nad dátami na strane databázového systému.

Medzi databáze reprezentujúce tento typ úložiska patrí napríklad CouchDB a MongoDB.

4.1.4 Grafový model

Tento typ databáz využíva pre prácu s dátami matematickú štruktúru graf. Dáta sú reprezentované pomocou uzlov, hran a ich atribútov. Základným objektom je uzol. Pomocou hrán medzi uzlami modelujeme závislosti, ktoré sú popísané pomocou atribútov. Nad uzlami a hranami sa využíva model kľúč-hodnota. Medzi hlavné výhody patrí možnosť prechádzania týchto štruktúr s využitím známych grafových algoritmov.

Tento model sa napríklad využíva v aplikáciach sociálnych sietí alebo pre sémantický web. Patria sem databázové systémy Neo4j alebo FlockDB.

4.2 Porovnanie NoSQL systémov

V dnešnej dobe existuje veľké množstvo NoSQL databázových systémov, ktoré majú odlišné vlastnosti, komplexitu a vďaka tomu ich môžeme použiť pre rôzne účely. Snaha porovnať tieto systémy na globálnej úrovni je nerealizovateľná a často vedie k chybám. Cieľom tejto sekcie je definovať základné body vďaka ktorým je možné tieto systémy kategorizovať a vrámci danej kategórie porovnávať. Tieto zistenia nám následne môžu pomôcť pri výbere vhodného systému, ktorý bude vhodný na riešenie našich požiadavok.

Následujúce body patria medzi hlavné kritéria pri porovnávaní týchto systémov:

1. dátový model
2. dotazovací model
3. škálovateľnosť
4. schopnosť odolávať chybám (angl. failure handling)
5. elasticita
6. konzistencia dát
7. prostredie behu
8. bezpečnosť

4.2.1 Dátový a dotazovací model

Dátový model definuje štruktúru, ktorá slúži na ukladanie dát v databázovom systéme. Dotazovací model následne definuje obmedzenia a operácie, ktoré je možné nad uloženými dátami vykonávať. V predchádzajúcej sekcii sme popísali základnú kategorizáciu NoSQL systémov podľa dátového modelu. Dátový a dotazovací model do určitej miery popisuje výkonnosť, komplexnosť a vyjadrovaciu silu databázového systému. Dotazovací model určuje programové rozhranie (API).

Pri výbere vhodného dátového modelu pre aplikáciu je dôležité porozumieť štruktúre dát a definovať operácie, ktoré nad týmito dátami budeme vykonávať.

4.2.2 Škálovateľnosť a schopnosť odolávať chybám

Tieto vlastnosti kladú na systém požiadavky ako podpora replikácie a rozdeľovanie dát. Distribuované databázové systémy implementujú tieto techniky na systémovej úrovni. V prípade ich podpory, môžeme od systému požadovať:

- podporu replikácie medzi geograficky oddelenými dátovými centrami
- možnosť pridania nového uzlu do distribuovaného databázového systému, bez nutnosti zásahu do aplikácie

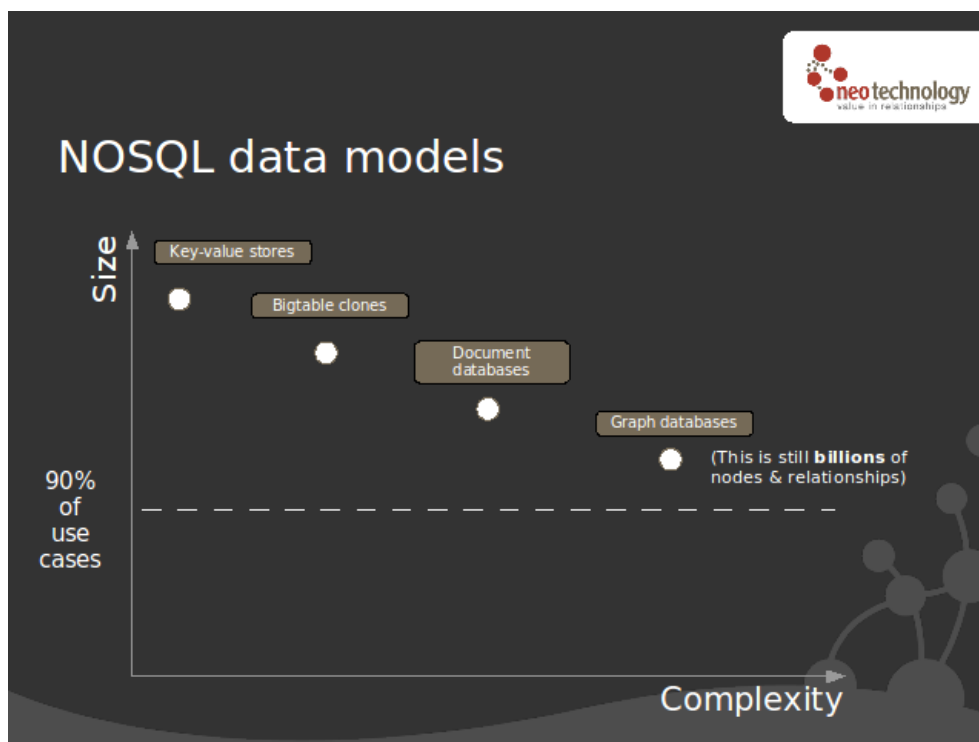
Počet uzlov, na ktorých sa vykonáva replika dát a konfigurácia databázového systému, ktorá podporuje geograficky oddelené dátové centrá určujú stupeň odolnosti systému voči chybám, medzi ktoré patria porucha uzla alebo sieťové prerušenie.

Častým požiadavkom webových aplikácií, z dôvodu neustáleho nárastu dát, na databázový systém je podpora škálovania s cieľom zvýšenia veľkosti databáze. S neustálym vývojom nových aplikácií musíme zároveň uvažovať potrebu škálovania z pohľadu komplexnosti dát [32]. Výber dátového modelu môže byť ovplyvnený na základe komplexnosti dát. Obrázok 4.2 zachytáva pozíciu dátových modelov z pohľadu škálovania komplexnosti a veľkosti dát.

Dátový model typu kľúč-hodnota a stĺpcovo orientovaný model (Bigtable clones) majú jednoduchú štruktúru, ktorá kladie minimálnu náročnosť na horizontálne škálovanie. Nevýhodou tohto prístupu je naopak to, že všetká práca s dátami a ich štruktúrou sa prenáša do vyšších vrstiev, o ktoré sa musí starať programátor. Naopak dokumentový a grafový model poskytuje bohatšiu štruktúru na prácu s dátami, ktorá spôsobuje komplikovanejšie škálovanie vzhľadom k veľkosti dát. Podľa odhadov spoločnosti Neotechnology až 90% aplikácií, v prípade že sa nejedná o projekty spoločností Google, Amazon atď., spadá do rozmedzia kde sa veľkosť záznamov pohybuje rádovo v miliardách. Za zmienku stojí fakt, že aj napriek tomu, že tieto dátové modely sú si navzájom izomorfné, vhodnosť ich použitia závisí na konkrétnom príklade a požiadavkách na aplikáciu.

4.2.3 Elastickosť

Vďaka horizontálnemu škálovaniu sa snažíme o zvýšenie kapacity a dostupnosti dátového úložiska. Elastickosť škálovania určuje ako sa daný systém dokáže vysporiadať s pridaním alebo odobraním uzla, či je v tomto prípade potrebný manuálny zásah pre rovnomerné



Obr. 4.1: Pozícia dátového modelu z pohľadu škálovania podľa veľkosti a komplexnosti. Zdroj: Neo4J a NOSQL overview and the benefits of graph databases, Emil Eifrem, prezentácia, [TODO link](#).

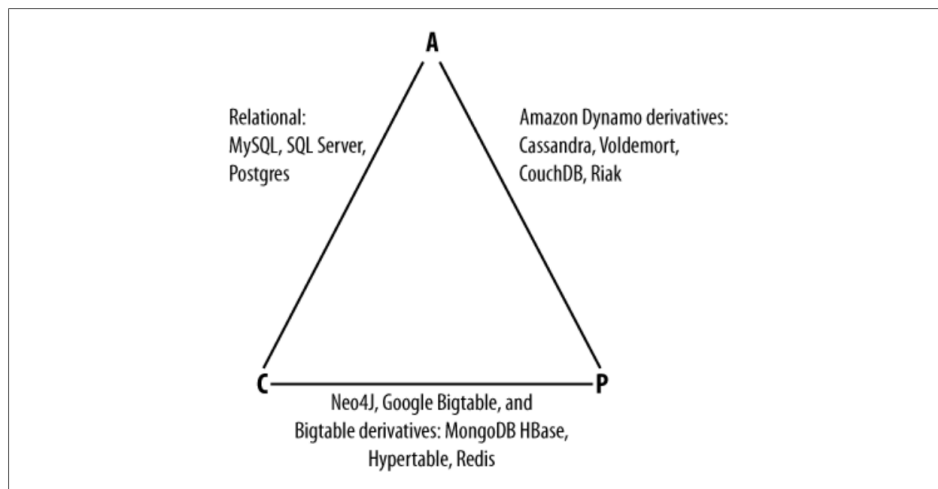
rozloženie záťaže, reštartovanie systému alebo zmena v užívateľskej aplikácii. V prípade pridávania nových uzlov do systému by táto vlastnosť mala zabezpečiť lineárne zvyšovanie výkonu u operácií ako je čítanie alebo zapisovanie dát.

4.2.4 Konzistencia dát

Podľa teórie CAP platí, že v prípade výskytu sieťových prerušení, ktorých prítomnosť v prostredí distribuovaných databázových systémov je prirodzená, nie je možné súčasne zaručiť vlasnosť konzistencie a dostupnosti. NoSQL systémy preto môžeme rozdeliť podľa tohto modelu do troch skupín, ktoré znázorňuje obrázok 4.2. Umiestnenie niektorých dátázových systémov sa môže meniť podľa na základe ich konfigurácie.

4.2.5 Prostredie behu

NoSQL systémy ďalej delíme podľa prostredia, v ktorom pracujú. Väčšina týchto open source systémov umožňuje ich nasadenie do vlastného prostredia teda privátnych cloudov alebo do verejného cloudového riešenia EC2 od spoločnosti Amazon. Medzi tieto systémy patria napríklad Cassandra, HBase, Riak, Voldemort. Databázové systémy Amazon SimpleDB, Microsoft Azure SQL, Yahoo! YQL alebo Google App Engine sú poskytované ako kom-



Obr. 4.2: Rozdelenie databázových systémov podľa CAP

plexné cloudové riešenie. Rozhranie pre prácu s dátami a funkčnosť perzistentného úložiska zabezpečujú poskytovatelia týchto systémov.

4.2.6 Bezpečnosť

Distribučovaný databázový systém je možné nasadiť do claudu. V prípade, použitia verejných claudov môže hroziť nebezpečenstvo zneužitia dát treťou stranou a preto je potrebné aplikovať bezpečnostné mechanizmy. Pri nasadení systému do privátneho klaudu zasa môžeme požadovať viacero úrovní ochrany pre prístup k dátam. Systémy NoSQL disponujú minimálnymi prvkami pre zaručenie autentizácie a autorizácie. Mnoho systémov tieto mechanizmy vôbec neimplementuje.

4.3 Výber NoSQL systémov

V predchádzajúcej sekcii sme tieto distribuované databázové systémy rozdelili do štyroch hlavných kategórií na základe ich dátového modelu, ktorý je jedným z kľúčových faktorov pri výbere vhodného kandidáta podľa požiadavkov cieľovej aplikácie. Detailný popis a výkonnostné porovnanie NoSQL systémov, ktoré reprezentujú jednotlivé kategórie by boli nad rámec tejto práce. Paralelne s touto prácou vzniká diplomová práca, ktorá rieši podobný problém s využitím dokumentových databázových systémov ??, preto túto kategóriu vynecháme.

Podľa analýzy požiadavkov na našu aplikáciu, ktoré sme popísali v predchádzajúcej kapitole nie je vhodné použitie databázových systémov s grafovým modelom a modelom typu kľúč-hodnota. Systémy s grafovým modelom sú určené na odlišnú úlohu problémov. V prípade použitia systémov s dátovým modelom kľúč-hodnota by sme komplexnosť riešenej úlohy preniesli na úroveň klienta. Naším požiadavkom najlepšie vyhovuje stĺpcovo orientovaný model, ktorý využijeme v návrhu našej aplikácie. Pre potreby nášho riešenia použijeme stĺpcovo orientovaný NoSQL systém.

Stĺpcovo orientované NoSQL systémy

V tejto časti sa zameriame na stručný prehľad a vzájomné porovnanie systémov, ktoré poskytujú stĺpcovo orientovaný model. Medzi tieto open source systémy patria HBase, Cassandra a Hypertable. Napriek totožnému dátovému modelu sú tieto systémy založené na rôznych architektonických princípoch a disponujú čiastočne odlišnými vlastnosťami. Tabuľka 4.1 zobrazuje stručný prehľad vlastností, na ktoré sme sa zamerali pri výbere víťaznej dvojice.

Vlastnosť	Databázový systém		
	HBase	Cassandra	Hypertable
Distribučný systém	áno	áno	áno
Architektúra	Bigtable	Dynamo	Bigtable
Klient	Thrift, REST	Thrift, Avro	Thrift, C++
Perzistentné úložisko	HDFS, AS3, KFS	LSS ¹	HDFS, KFS, LSS
SPOF ²	áno	nie	- ³
Podpora viacerých datacentier	áno	áno	áno
Automatické rozdeľovanie dát	áno	áno	áno
Replikácia	áno	áno	áno
Konzistencia	CP	voliteľná	CP
Kompresia dát	LZO, GZIP	nie	LZO, ZLIB
Programovací jazyk	Java	Java	C++
MapReduce	áno	áno	áno
Dokumentácia	+	+	-
Komunita	+	+	-

Tabuľka 4.1: Stručný prehľad vlastností stĺpcovo orientovaných systémov NoSQL

Zo stručného prehľadu v tabuľke je vidieť, že systémy obsahujú množstvo spoločných vlastností. Pri výbere systémov sme zohľadnili aj ich praktické využitie spoločnosťami globálne pôsobiacimi na IT trhu. Napríklad systém Cassandra je používaný spoločnosťou Facebook v aplikácii pre vyhľadávanie a uchovávanie súkromnej pošty. Ďalšími dôležitými požiadavkami pre výber boli podpora komunity, dokumentácia a vývojový cyklus týchto systémov. Zo zvolených kandidátov sme vybrali systém HBase a Cassandra. Dôvodom prečo sme zavrhlí systém Hypertable je nedostačujúca dokumentácia, málo aktívna komunita a pomalý vývojový cyklus. V nasledujúcich dvoch kapitolách detailne popíšeme oba systémy.

Kapitola 5

Cassandra

Distribúovaný databázový systém Cassandra bol vytvorený pre interné účely spoločnosti Facebook v roku 2007. Cassandra slúžila na vyhľadávanie v súkromnej pošte a poskytovala úložisko pre indexy vytvárané nad týmito dátami. Od systému sa predpokladalo, že bude zvládať miliardy zápisov denne, podporovať škálovanie podľa narastajúceho počtu používateľov, umožňovať beh na spotrebných počítačoch a podporovať replikáciu medzi geograficky oddelenými dátovými centrami. Ďalším požiadavkom bola vysoká dostupnosť, aby chyba žiadneho uzlu nespôsobila celkovú nedostupnosť systému. Cassandra je decentralizovaný systém, kde každý uzol vykonáva tie isté operácie. V roku 2008 bola zverejnená ako open source projekt a je neustále vyvíjaná mnohými spoločnosťami a vývojármi. Tento systém využíva architektonické princípy distribuovaného databázového systému Dynamo od spoločnosti Amazon a zároveň ich kombinuje s dátovým modelom distribuovaného databázového systému Bigtable vytvoreného spoločnosťou Google. V nasledujúcom texte popíšeme hlavne princípy, na ktorých je tento systém založený.

5.1 Dátový model

Cassandra k dátovému modelu systému Bigtable pridala štruktúru pod názvom „*super stĺpec*“ (Super column). Základnou jednotkou dátového modelu je stĺpec. Stĺpec je tvorený názvom, hodnotou a časovým odtlačkom, ktorý využíva Cassandra pri riešení konfliktov. Skupina stĺpcov je identifikovaná pomocou unikátneho kľúča a predstavuje riadok, počet a názvy stĺpcov nie je potrebné vopred definovať. Zoradené riadky podľa hodnoty kľúčov a v nich zoradené stĺpce obaľuje štruktúra pod názvom „rodina stĺpcov“ (Column family). Kľúče interne reprezentované ako reťazec znakov sú zároveň zotriedené. Názvy stĺpcov môžu byť viacerých typov ako napríklad Ascii, UTF-8 a bajtové pole, podľa ktorých sú zotriedené. Je možné implementovať vlastnú metódu pre triedenie. Riadky obsiahnuté v jednej rodine stĺpcov sú na pevnom disku fyzicky umiestnené v jednom súbore typu *SSTable*. Je vhodné do rodiny stĺpcov ukladať relevantné záznamy, ku ktorým budeme pristupovať spoločne, čím sa vyhneme nadbytočným diskovým operáciám. V rámci jednej repliky sú operácie nad stĺpcami atomické. Operácie nad riadkom nevyužívajú zamykanie. Voliteľným príznakom, ktorý môžeme u stĺpca nastaviť je parameter TTL (Time to live), ktorý po uplynutí časového intervalu označí dáta príznakom pre zmazanie.

Štruktúra super stĺpec je špeciálny typ stĺpca, ktorý je tvorený obyčajnými stĺpcami. Stĺpec typu super má názov a jeho hodnota je tvorená zoznamom názvov obyčajných stĺpcov. Super stĺpce obaľuje štruktúra pod názvom „*super-rodina stĺpcov*“ (Super column family). *Keyspace* združuje rodiny a super-rodiny stĺpcov. Definuje faktor a metódu replikácie, ktorá môže byť závislá popri prípade nezávislá na sieťovej topológii. Na keyspace sa môžeme pozerať ako na databázu a rodiny stĺpcov môžeme prirovnať k tabuľkám v relačných databázových systémoch.

Aktuálna verzia Cassandra definuje maximálnu veľkosť dát 2 GB, ktoré je možné uložiť do jedného stĺpca a stanovuje limit dve miliardy pre maximálny počet stĺpcov tvoriacich riadok.

5.2 Rozdeľovanie dát

Kľúčovým požiadavkom systému Cassandra je jeho schopnosť horizontálneho škálovania, čo vyžaduje pridávanie nových uzlov. Tento požiadavok vyžaduje mechanizmus, ktorý zabezpečí automatické rozdeľovanie dát medzi uzlami systému. Uvažujme príklad, kde máme k dispozícii jeden server obsahujúci veľké množstvo objektov, ku ktorým pristupujú klienti. Medzi server a klientov vložíme vrstvu kešovacích systémov, kde každý z týchto systémov bude zodpovedný za prístup k časti objektov nachádzajúcich sa na serveroch. Klient musí byť schopný určiť na základe hľadaného objektu, ku ktorému kešovaciemu systému. Predpokladajme, že klientom zabezpečíme výber jednotlivých kešovacích systémom pomocou lineárnej hašovacej funkcie $y = ax + b \pmod{p}$, kde p je počet kešovacích systémov. Pridanie nového kešovacieho systému alebo jeho zlyhanie bude mať katastrofálny dopad na funkčnosť systému. V prípade zmeny hodnoty p , bude každá položka bude odpovedať novej a zároveň chybné lokácii. Tento problém rieši elegantne technika pod názvom „*úplné hašovanie*“ (Consistent hashing) [28], ktorá sa využíva v distribuovaných systémoch pre určenie pozície pre prístup k distribuovaným hašovacím tabuľkami a využíva ju systém Cassandra.

Výstup hašovacej funkcie MD5 reprezentujeme pomocou „kruhu“, kde v smere hodinových ručičiek postupujeme od minimálnej hodnoty hašovacej funkcie (0) k maximálnej (2^{127}). Každý uzol v systéme má pridelenú hodnotu z tohoto rozsahu, ktorá určí jeho jednoznačnú pozíciu. Identifikácia uzlu v systéme, na ktorý sa uložia dáta reprezentované hodnotou kľúča sa vykoná aplikáciou hašovacej funkcie na dáta reprezentujúce kľúč. Na základe tejto hodnoty je jednoznačne určená pozícia v kruhu a v smere hodinových ručičiek je vyhľadaný najbližší uzol, ktorému dáta prináležia. Výhodou tejto metódy je, že každý uzol je zodpovedný za hodnotu kľúčov, ktorých poloha sa nachádza medzi ním a jeho predchodcom. V prípade pridania nového uzlu alebo jeho odobratia, sa zmena mapovania kľúčov v kruhu prejaví len u jeho susedov. Táto technika prináša výhody, medzi ktoré patrí rovnomerná distribúcia dát a vyváženie záťaže. Dynamo tento problém rieši spôsobom kde každý uzol je zodpovedný za viacero pozícií na kruhu, takzvané virtuálne pozície. Cassandra využíva vlastné mechanizmy na monitorovanie záťaže uzlov a automaticky presúva ich pozície v kruhu. Zároveň je možné explicitne stanoviť polohu každého uzla v kruhu pomocou zadania jeho identifikátora. Tento spôsob je vhodný v prípade, ak vieme predom určiť koľko uzlov bude systém obsahovať. V prípade zvyšovania počtu uzlov je možné tieto identifikátory zmeniť za behu systému. Identifikátor polohy uzlov je možné určiť pomocou nasledujúceho programu v jazyku Python, kde K je počet uzlov v systéme.

```
def tokens(n):
    r = []
    for x in xrange(n):
        r.append(2**127 / n * x)
    return r

print tokens(K)
```

5.3 Replikácia

S úplným hašovaním úzko súvisí replikácia, ktorá zabezpečuje vysokú dostupnosť a odolnosť dát proti ich strate (Durability). Každá dátová jednotka vložená do systému je replikovaná na N uzlov, kde počet N je voliteľne nastaviteľný pre daný keyspace. Každý uzol sa v prípade replikácie $N > 1$ stáva koordinátorom, ktorý je zodpovedný za replikáciu dát, ktorých kľúč spadá do jeho rozsahu. Počas operácie zápisu koordinátor replikuje dáta na ďalších $N - 1$ uzlov. Cassandra podporuje viacero spôsobov pre umiestňovanie replík.

Jednoduchá stratégia

Táto stratégia umiestňuje replikú dát bez ohľadu na umiestnenie uzlov (serverov) v data-centre. Primárna replika je uložená na $N - 1$ susedov v smere hodinových ručičiek. Z toho vyplýva, že každý uzol je zodpovedný za dáta, ktorých kľúče spadajú do jeho rozsahu a súčasne dáta, ktoré spravuje jeho N predchodcov.

Sieťová stratégia

Pri tejto metóde a úrovni replikácie s hodnotou aspoň tri, sme schopný zabezpečiť umiestnenie dvoch replík v rozdielnych „rackoch“, tretia replika bude umiestnená do iného datacentra. Táto stratégia je výhodná v prípade ak chceme použiť časť serverov na výpočty pomocou MapReduce a zvyšné dve repliky budú slúžiť na obsluhu reálnej prevádzky.

5.4 Členstvo uzlov v systéme

Distribúovaný systém musí byť schopný odolávať chybám ako porucha uzlov alebo sieťové prerušenia. Decentralizácia a detekcia chýb využíva mechanizmy založené na *gosship protokoloch* [18]. Tieto protokoly slúžia pre vzájomnú komunikáciu uzlov vymenšujúcich si navzájom dôležité informácie o svojom stave. Periodicky v sekundových intervaloch každý uzol kontaktuje náhodne vybraný uzol, kde si overí či je tento uzol dostupný. Detekcia nedostupnosti uzla je realizovaná algoritmom s názvom Accrual Failure Detector [26].

Pridávanie nových uzlov, presun uzlov v rámci kruhu sa taktiež dejú pomocou Gosship protokolu, ktorý ďalej zabezpečuje, že každý uzol obsahuje informácie o tom, ktorý uzol je zodpovedný za daný rozsah kľúčov v kruhu. Ak sa vykonáva operácia čítania alebo zápisu dát na uzol, ktorý nie je zodpovedný za tento kľúč, dáta sú automaticky preposlané na správny uzol, ktorého výber je zabezpečený s časovou zložitou $O(1)$.

5.5 Zápis dát

Tento systém bol primárne navrhnutý tak aby spracúval vysoký tok dát pre zápis. Ak uzol obrží požiadavku pre zápis, dáta sú zapísané do štruktúry pod názvom *Commitlog*, ktorá je uložená na lokálnom súborovom systéme a zabezpečuje trvácnosť dát. Zápis do tejto štruktúry je vykonaný sekvenčne, čo umožňuje dosiahnuť vysokú priepusnosť bez nutnosti vystavovania diskových hlavičiek (disk seek operations). Dáta sú následne zoradené a nahrané do štruktúry pod názvom *Memtable*, ktorej obsah je uložený v operačnej pamäti. V prípade, že by tento zápis zlyhal alebo by došlo k neočakávanému pádu inštancie Cassandra obnovenie obsahu týchto štruktúr je možné vďaka *Commitlogu*. Po dosiahnutí predom stanovovej hodnoty, ktorá určuje maximálny počet dát uložených v *Memtable* sú tieto štruktúry zapísané do súborov pod názvom *SSTable* (Sorted String Tables), ktorých obsah je zoradený podľa hodnoty kľúčov a nie je možné ich modifikovať. Súborové SSTables sa zlievajú (merge sort) v pravidelných intervaloch na pozadí a táto operácia je neblokujúca. Počas zlievania SSTables dochádza k odstraňovaniu dát určených na vymazanie a generovaniu nových indexov. Indexy slúžia na rýchly prístup k dátam uloženým v *SSTable*. Zároveň sa generujú štruktúry pod názvom *Bloom filters*¹, ktoré sa využívajú pri čítaní dát. Požiadavok pre zápis dát je možné zaslať na ľubovoľný uzol v klastri.

5.6 Čítanie dát

V prípade požiadavky na načítanie dát, sa požadované dáta najprv hľadajú v štruktúrach *Memtable*. Ak sa dané dáta nenachádzajú v operačnej pamäti, vyhľadávanie sa uskutočňuje podľa kľúča v súboroch *SSTable*. Keďže snahou systému, je čo najefektívnejšie vyhľadávanie, využívajú sa bloom filtre. Bloom filtre sú nedeterministické algoritmy, ktoré dokážu otestovať či hľadaný prvok patrí do množiny a generujú len falošné pozitíva. Pomocou nich je možné priradiť kľúče uložené v súboroch SSTables do bitových polí, ktoré je možné uchovať v operačnej pamäti. Vďaka tomu sa výrazne redukuje prístup na disk. Požiadavok pre čítanie dát môžeme zaslať na ľubovoľný uzol.

5.7 Zmazanie dát

Počas vykonania operácie reprezentujúcej zmazanie dát sa tieto dáta nevymažú okamžite. Namiesto toho sa vykoná operácia, ktorá dané dáta označuje príznakom pod názvom *tombstone*. Po uplynutí doby, ktorá je štandardne nastavená na desať dní, sa tieto dáta odstránia pri procese zlievania súborov SSTables.

5.8 Konzistencia

Konzistencia systému je maximálne konfigurovateľná a využíva princípy techník založených na protokoloch kvóra. Klient určuje hodnotu *R*, ktorá stanovuje počet replík pre čítanie dát. Operácia je úspešná v prípade dostupnosti týchto replík. To samé platí pre čítanie dát, kde

¹http://en.wikipedia.org/wiki/Bloom_filter

počet replík je určený parametrom W . Ak je splnený vzťah $R + W > N$, kde N je počet replík tak výsledok operácie spĺňa definíciu silnej konzistencie, naopak v prípade vzťahu $R + W < N$, sa jedná o čiastočnú konzistenciu čím zaručíme vysokú dostupnosť.

5.9 Perzistentné úložisko

Cassandra využíva ako perzistentné úložisko dát lokálny súborový systém.

5.10 Bezpečnosť

Implicitne Cassandra nevyužíva žiadne prvky zabezpečujúce bezpečnostné mechanizmy. K dispozícii je modul, ktorý umožňuje nastavenie autentizácie na úrovni Keyspace-u pomocou textových hesiel alebo ich MD5 odtlačkov. Obmedzovanie prístupu k dátam je preto potrebné zabezpečiť na aplikačnej úrovni.

Kapitola 6

HBase

V úvode tejto kapitoly stručne popíšeme distribuovaný súborový systém, ktorý je súčasťou projektu Hadoop a zároveň slúži ako perzistentné úložisko pre distribuovanú databázu HBase. Následne popíšeme základné princípy fungovania tohoto databázového systému.

Hadoop

Hadoop¹ je open source projekt vytvorený v programovacom jazyku Java, ktorý tvorí distribuovaný súborový systém Hadoop Distributed Filesystem (HDFS) a framework MapReduce pre spracúvanie objemu dát v desiatkách PB [35]. Medzi hlavné vlastnosti tohoto systému patria vysoká dostupnosť, škálovateľnosť a distribuovaný výpočet. Architektúra HDFS vychádza z princípov distribuovaného súborového systému Google File System [21] (GFS) a pôvod frameworku MapReduce [13] pochádza taktiež od spoločnosti Google.

Súborový systém využíva architektúru master-slave. HDFS predpokláda prácu so súbormi rádo vo desiatkách gigabajtov, ktoré sú interne reprezentované dátovými blokmi o štandardnej veľkosti 65 MB. Informácie o priradení blokov k súborom a ich umiestnenie na uzloch typu slave sú reprezentované pomocou metadát. Tieto metadáta sú uložené v operačnej pamäti uzla typu master, ktorý sa nazýva *Namenode*. Uzly slave pod názvom *Datanode* slúžia ako fyzické úložisko dátových blokov, ktoré sú zároveň na týchto uzloch replikované. Štandardne je nastavená úroveň replikácie na hodnotu tri, každý blok je v systéme uložený trikrát.

Predtým ako klient vykoná operáciu zápisu alebo čítania dát, tak kontaktuje uzol Namenode, ktorý mu poskytne informácie, na ktorých uzloch sa nachádzajú bloky reprezentujúce súbor a dátová komunikácia následne prebehne medzi klientom a uzlami typu Datanode. HDFS je optimalizovaný pre jednorázový zápis dát a ich následné mnohonásobné čítanie.

Hlavným nedostatkom tejto infraštruktúry je fakt, že uzol Namenode tvorí kritický bod systému (SPOF), v prípade jeho nedostupnosti nie je možné pracovať s HDFS. Prípadná strata alebo poškodenie dát na tomto uzle spôsobí totálne zlyhanie súborového systému bez možnosti jeho obnovy. Súborový systém nie je vhodný pre ukladanie veľkého počtu malých súborov. Uzol Namenode alokuje pre objekt typu blok a objekt typu súbor 300 B metadát. V

¹<http://hadoop.apache.org/>

případe uloženia súboru, ktorý nepresahuje veľkosť jedného bloku je potrebné alokovať 300 B dát. Ak uložíme 10,000,000 takýchto súborov veľkosť metadát, ktoré udržiava Namenode v operačnej pamäti zaberie 3 GB. Celkový počet uložených súborov je obmedzený veľkosťou operačnej pamäti RAM, ktorou disponuje uzol Namenode.

Z týchto pozorovaní vyplýva fakt, že distribuovaný súborový systém HDFS nemá praktické využitie ako úložisko dát slúžiace k archivácii emailových správ, pre ktoré sme zadefinovali požiadavky v tretej kapitole.

HBase

Aplikácie HDFS a MapReduce slúžia na dávkové spracúvanie obrovského objemu dát. HBase je open source, distribuovaný, stĺpcovo orientovaný databázový systém, ktorý umožňuje prácu s veľkým objemom dát v reálnom čase. Ako perzistentné úložisko dát využíva distribuovaný súborový systém HDFS, je taktiež implementovaný v Jave a jeho architektonické koncepty vychádzajú z článku Bigtable od spoločnosti Google. HBase bol vytvorený spoločnosťou Powerset na konci roka 2006, pre potreby spracúvania obrovského objemu dát a začiatkom roka 2008 sa stal oficiálnym podprojektom systému Hadoop [38].

6.1 Dátový model

Dátový model je totožný s konceptom Bigtable. Dáta s ktorými pracuje HBase sa ukladajú do tabuliek. Každá tabuľka obsahuje riadky identifikované kľúčom, ktoré môžu byť tvorené ľubovoľným počtom stĺpcov. Kľúče sú reprezentované ako pole bajtov (Java byte[]). Pre zápis ich hodnoty môžeme použiť ľubovoľný typ dát, napríklad reťazec alebo serializovanú dátovú štruktúru pomocou JSON, XML. Riadky sú radené podľa názvov kľúčov, ktoré sú radené podľa hodnoty jednotlivých bajtov. Stĺpce sú organizované do skupín, ktoré sa nazývajú „rodiny stĺpcov“ (Column families). Obsahu každej bunky, ktorej pozíciu určuje riadok a stĺpec prináleži verzia, ktorá je reprezentovaná časovou značkou a jej obsah je reprezentovaný ako pole bajtov. Časovú značku určuje klient pri zápise dát alebo je automaticky generovaná systémom HBase (reprezentuje ju systémový čas). Obsah buniek tabuľky je možné prístupniť pomocou kľúča a názvu stĺpca alebo pomocou kľúča, názvu stĺpca a časovej značky. V prípade uloženia viacerých verzií v danej bunke, sú tieto dáta radené od najaktuálnejšej časovej značky po najstaršiu. Región tvorí interval riadkov, kde posledný riadok do daného intervalu nepatrí.

Dôležitým faktom je, že stĺpce, ktoré patria do rovnakej rodiny stĺpcov sú fyzicky uložené na tom istom mieste. Rodiny stĺpcov je potrebné zadefinovať počas vytvárania tabuliek. Ich názvy a počet je potrebné vhodne premyslieť už pri samotnom návrhu databázovej schémy. V prípade, že klaster HBase obsahuje viacero uzlov je na nich potrebné zabezpečiť synchronizáciu systémového času. V prípade veľkého časového rozdielu na jednotlivých uzloch kastera hrozí, že sa daná inštancia systému HBase sa nespustí.

6.2 Architektúra systému

HBase využíva architektúru master-slave. Uzol v role master sa nazýva *HBaseMaster*, uzly slave *RegionServers* (RS). Pre chod systému je ďalej potrebná služba Zookeeper ??, ktorá zabezpečuje ...

HBaseMaster

Tento uzol v systéme vykonáva priradzovanie regiónov RegionServer-om, detekujem pridanie nového RS, jeho zlyhanie a rovnomerne rozdeľuje záťaž na RS-och v prípade rozdelenia regiónu.

RegionServer

RegionServer slúži pre obsluhu klientských požiadavkov, samotný zápis alebo čítanie dát sa vykonáva medzi klientom a RS. Každý RS spravuje niekoľko regiónov, automaticky rozdeľuje regióny a informuje o tom uzol HBase Master. Tieto uzly môžeme v systéme ľubovoľne pridávať alebo odoberať, počas jeho prevádzky.

6.3 Rozdeľovanie dát

Základnou jednotkou, ktorá zabezpečuje rozdeľovanie dát a teda umožňuje horizontálne škálovanie a rovnomerné rozloženie záťaže v klastri je region. Region má náhodne vygenerovaný identifikátor. Tabuľka je tvorená regiónmi, pri jej prvotnom vytvorení ju zvyčajne reprezentuje jeden región. V prípade, že jej veľkosť dosiahne predom stanovenú hranicu (závislé na konfigurácii), dojde k rozdeleniu riadkov do dvoch nových regiónov s podobnou veľkosťou. Tieto regióny sú v klastri distribuované na uzly typu RegionServer. Tento mechanizmus zabezpečuje, že do systému je možné uložiť tabuľku o veľkosti, ktorú by nebolo možné uložiť alebo spracovať pomocou jedného fyzického počítača z dôvodu hardverových obmedzení. Región je základný element, ktorý zabezpečuje vysokú dostupnosť a rovnomerné rozloženie záťaže.

6.4 Replikácia

Primárnu replikáciu dát, ktorá spĺňa vlastnosti silnej konzistencie a zabraňuje strate dát zabezpečuje perzistentné úložisko, v tomto prípade HDFS.

HBase podporuje replikáciu v rámci viacerých geograficky oddelených datacentier. Replikácia funguje na rovnakom princípe ako v databázovom systéme MySQL², teda master-slave a je asynchrónna. Táto forma replikácie zanáša do daného distribuovaného systému čiastočnú konzistenciu, pretože negarantuje, že dáta budú zmenené na všetkých uzloch slave v rovnakom čase.

²<http://dev.mysql.com/doc/refman/5.1/en/replication-formats.html>

6.5 Perzistentné úložisko

Tento distribuovaný databázový systém je schopný pracovať v lokálnom móde, kde vyššie spomínané komponenty bežia ako samostatné služby na jednom fyzickom uzle a ako úložisko dát sa využíva lokálny súborový systém. V prípade distribuovaného módu rozlišujeme dva typy:

- pseudo-distribuovaný mód, kde všetky komponenty bežia na jednom uzle
- distribuovaný mód, jednotlivé komponenty bežia na samostatných uzloch

Obidve konfigurácie môžu využívať ako perzistentné úložisko distribuovaný súborový systém HDFS, KFS alebo Amazon S3. Štandardne sa doporučuje použitie HDFS.

6.6 Konzistencia

Systém sa vyznačuje silnou konzistenciou. Z modelu CAP splňuje CP, teda uprednostňuje konzistenciu pred dostupnosťou.

6.7 Zápis dát a čítanie dát

V prípade zápisu alebo čítania dát klient kontaktuje Zookeeper, od ktorého obrží informáciu o lokácii tabuľky *-ROOT-* a následne kontaktuje daný RS obsahujúci túto tabuľku. Z tabuľky *-ROOT-* sa určí RegionServer, ktorý obsahuje tabuľku *.META.*, tieto kroky sa lokálne kešujú na strane klienta. Následne klient kontaktuje daný RegionServer a v tabuľke *.META.* vyhľadá uzol obsahujúci cieľový región, do ktorého patria hľadané alebo cieľové dáta. V poslednom kroku prebieha celá dátová komunikácia medzi klientom a posledným vyhľadaným uzlom.

Dáta sú zapísané do štruktúry *HLog* (typu *Write-Ahead-Log*), ktorá je uložená v súborovom systéme HDFS. Po potvrdení o úspešnom zápise sú data nahrané do štruktúry *MemStore*, ktorá je uschovaná v operačnej pamäti. Štruktúry *MemStore* sú zapisované obdobne ako u systému Cassandra do zoradených súborov typu *HFile*, ktorých štruktúra je podobná ako u súborov *SSTable*. V prípade, že RegionServer obrdží požiadavok na čítanie dát, dáta sú načítane buď zo štruktúry *MemStore* alebo *HFile*. Pre zvýšenie výkonnosti je možné použiť Bloom filtre.

Zápis alebo čítanie dát na úrovni riadku identifikovaného pomocou kľúča je atomická operácia.

6.8 Zmazanie dát

Pri operácii zmazania dát je možné určiť vymazávané data konkrétnou časovou značkou, poprípade je možné vymazanie všetkých verzií dát, ktoré sú staršie ako zadaná časová značka. Zmena sa nevykoná okamžite z dôvodu, že obsah štruktúr *HFile* nie je možné modifikovať (jedným z dôvodov je aby sa zabránilo vykonávaniu nadbytočných diskových operácií).

Namiesto toho sa vykoná operácia, ktorá dané dáta označuje príznakom pod názvom *tombstone*. Tieto dáta sa odstránia obdobne ako v systéme Cassandra, pri procese zlievajúcim súbory HFile.

6.9 Bezpečnosť

Hadoop a HBase aktuálne poskytujú prvok autentizácie pomocou služby Kerberos. Možnosť pridania autorizácie na úrovni tabuliek a rodiny stĺpcov je neustále vo vývoji.

Kapitola 7

Testovanie výkonnosti

Výkonové porovnanie NoSQL systémov je zložitá úloha, neexistuje univerzálny nástroj, ktorým by bolo možné tieto systémy navzájom porovnať. Každý z týchto systémov sa vyznačuje špecifickými vlastnosťami medzi ktoré patria typ konzistencie, dostupnosť, optimalizácia pre zápis alebo čítanie a ich výber závisí na konkrétnom prípade použitia. Všeobecný nástroj pre ich porovnanie by preto nemal žiadne opodstatnenie. Aktuálne nie sú k dispozícii žiadne všeobecné techniky, ktorými by bolo možné testovať napríklad konzistenciu, spoľahlivosť a ich iné vlastnosti. Výkon týchto systémov ovplyvňuje hodnota replikácie, spôsob rozdeľovania dát a typ konzistencie. Veľmi častou a zároveň časovo náročnou metódou, ktorá slúži na porovnávanie týchto systémov je implementácia daného riešenia s využitím všetkých porovnávaných systémov. V tejto kapitole sa zameriame na popis testov, ktoré sme vykonali v reálnych podmienkach.

7.1 Testovacie prostredie

Pre výkonnostné testovanie sme mali k dispozícii 9 počítačov s rovnakou hardvérovou a softvérovou konfiguráciou, ktoré boli navzájom prepojené pomocou 10 Gbit smerovača a komunikovali po 1 Gbit linke. Konkrétnu softvarovú konfiguráciu testovaných aplikácií popíšeme jednotlivo v nasledujúcich podkapitolách.

Hardvérová konfigurácia

- 4 jádrový procesor Intel, 5506@2.13Ghz
- 4 GB RAM
- 5 pevných diskov (SATA, 7200RPM) o veľkosti 1TB zapojených v poli RAID0
- 1 Gbit sieťová karta

Softvérová konfigurácia

Každý uzol obsahoval inštaláciu operačného systému Debian GNU/Linux Lenny x64 a aplikáciu Sun Java JDK 1.6.0_₊88. Na každom uzle bol deaktivovaný odkladací priestor (angl. swap). Za účelom monitorovania bol použitý softvér Zabbix a nástroje VisualVM, htop, iostat a dstat.

Sieťová konfigurácia

Hodnota maximálnej reálnej sieťovej priepustnosti medzi dvoma uzlami bola zmeraná pomocou aplikácie nuttcp¹ s výslednou hodnotou 940 Mbit.

7.2 Popis testovacej metodológie

Nad oboma distribuovanými databázovými systémami sme vykonali testy na základe, ktorých sme sa snažili identifikovať ako tieto systémy ovplyvňuje rôzna hodnota replikácie, konzistencie, pozorovali sme ich schopnosť horizontálneho škálovania a chovanie sa pod záťažou. Testy boli zamerané na zápis dát, ktorý je kritickým prvkom pre potreby našej aplikácie.

7.2.1 Testovací klient

Testovacím klientom bola aplikácia využívajúca vlákna, založená na princípe producent - konzument, kde konzumenti predstavovali jednotlivé vlákna vykonávajúce zápis alebo čítanie dát. Optimálny počet paralelne zapisujúcich vlákien bol stanovený na hodnotu 50, pri ich navýšení sa zvyšovala hodnota latencie a nedošlo k zväčšeniu dátového toku pre zápis. Pre čítanie dát bolo použitých 250 vlákien. Dôležitým bodom bolo zabezpečenie rovnakého zaťaženia každého uzla v klastri, počas celého priebehu jednotlivých testov. Detailný popis spĺňujúci tento bod je obsiahnutý v časti popisujúcej test konkrétneho databázového systému.

7.2.2 Testovací prípad pre zápis dát

V tomto testovacom prípade sme postupne do klastra obsahujúceho jeden, tri a šesť uzlov zapisovali 8,000,000 riadkov. Každý riadok obsahoval jeden stĺpec o veľkosti 1000 B, ktorého obsah tvorili náhodne vygenerované dáta. Tento počet zapisovaných riadkov sme zvolili z dôvodu, aby počas zápisu dochádzalo k zlievaniu štruktúr Memtable a Memstore. Vďaka týmto operáciám sa chovanie klastra viac priblížilo reálnym podmienkam.

7.2.3 Testovací prípad pre čítanie dát

Dôvodom tohto testovacieho prípadu bolo určiť rýchlosť počas čítania dát z databázového systému. Táto rýchlosť bude mať výrazný vplyv na celkovú dobu trvania výpočtov štatistík, pomocou MapReduce. V tomto testovacom prípade sme náhodne načítali 1,000,000 riadkov z databázového systému o veľkosti 1000 B.

¹<http://www.wcisld.hpc.mil/nuttcp/>

7.2.4 Zaažovací test

Cieľom bolo zistiť stabilitu klastra v prípade, keď bude pod sústavným zápisom, budú v ňom prebiehať časté operácie pre zápis štruktúr SSTable a HTable na pevný disk a ich následné zlievanie. Tento test sme vykonali po dobu piatich hodín. Počas niektorých testovacích prípadov sme použili dvoch klientov z dôvodu aby sme zaručili maximálnu saturáciu prenosového pásma a vylúčili úzke hrdlo na strane klienta (1 Gbit linka umožňuje maximálny teoretický dátový tok 125 MB/s).

7.3 HDFS

Nad distribuovaným súborovým systémom HDFS sme vykonali testy určujúce maximálnu hodnotu priepustnosti pri zápise dát, z dôvodu aby sme vylúčili možné úzke hrdlo v jeho prepojení s databázovým systémom HBase. Pre účely testovania sme použili verziu Hadoop-0.20.2, veľkosť haldy pre JVM (ang. Java Virtual Machine) bola nastavená na hodnotu 1 GB.

Meranie sme vykonali v troch konfiguráciach. Každá konfigurácia obsahovala jeden uzol v role master, na ktorom boli spustené služby Namenode a JobTracker. Na zvyšných uzloch typu slave bežali služby Datanode a Tasktracker. Konfigurácia klastra bola nasledovná:

- A - tri uzly slave s faktorom replikácie jedna
- B - tri uzly slave s faktorom replikácie tri
- C - šesť uzlov slave s faktorom replikácie tri

Testy boli vykonané nástrojom TestDFSIO, ktorý je súčasťou zdrojových kódov systému Hadoop a využívali techniku MapReduce. Počas jednotlivých testov sa na súborový systém HDFS zapisovali tri rôzne veľkosti súborov, bola zachovaná štandardná veľkosť bloku 65MB. Každý test bol vykonaný trikrát a výsledná hodnota bola určená ako aritmetický priemer. Výsledky testu, ktoré zobrazuje tabuľka 7.1, reprezentujú maximálny agregovaný tok pre zápisu dát v klastri.

Veľkosť súboru [MB]	Klaster		
	A	B	C
65	287 MB/s	102 MB/s	190 MB/s
512	371 MB/s	85 MB/s	162 MB/s
2048	433 MB/s	85 MB/s	163 MB/s

Tabuľka 7.1: Priepustnosť pri zápise dát na HDFS

Na základe týchto hodnôt môžeme pozorovať, že zvýšenie faktoru replikácie má zásadný negatívny vplyv na celkový výkon distribuovaného súborového systému. Dôležitý fakt, ktorý vyplynul z výsledkov testovania je, že v prípade ak zvýšime dvojnásobne počet uzlov v klastri (prípady klastrov v konfigurácii B a C) jeho výkonnosť vzrastie lineárne, čo potvrdzuje vysokú škálovateľnosť systému.

7.4 HBase

Pre test distribuovaného databázového systému sme nainštalovali systém Hadoop 0.20.2 a HBase 0.90.1. Na jednom fyzickom uzly bežali nasledujúce služby:

- HBase Master
- Zookeeper
- Namenode

Tieto služby sú v oboch systémoch súčasťou uzla master. Na zvyšných uzloch boli spustené služby Region Server a Datanode. JVM sme v prípade systému HDFS pridelili 1 GB operačnej pamäti a v prípade systému HBase 2 GB.

Prázdna tabuľka je po vytvorení v HBase reprezentovaná jedným regiónom. Tento región je uložený na jednom uzle. K rozdeleniu tohto regiónu dochádza v prípade ak objem dát zapísaných v tabuľke prekročí štandardne nastavenú hranicu s hodnotou 256 MB. Prázdnu tabuľku sme vytvorili pomocou programového rozhrania systému HBase (API?) a to tak, že sme ju predrozdelili na počet regiónov, ktorý odpovedal počtu uzlov typu slave v klastri. Názvy kľúčov sme generovali pomocou náhodného generátora. Vďaka tejto metóde sme dosiahli rovnomerné zaťaženie všetkých uzlov po celú dobu testovania.

Tabuľka 7.2 zobrazuje výsledky meraní podľa testovacieho prípadu pre zápis dát.

Počet uzlov	Replikácia	Čas	Riadok/sek	Priepustnosť [MB/s]
1	1	551	14519	14
3	1	202	39613	39
3	3	317	25110	25
6	3	211	37864	37

Tabuľka 7.2: Hbase: zápis riadkov o veľkosti 1000 B

Škálovateľnosť

Z daných meraní vidíme, že systém je maximálne škálovateľný a dvojnásobne zvýšenie počtu uzlov zvýši priepustnosť o cca 50%.

Replikácia

Zvýšenie hodnoty replikácie (riadky 3,4) spôsobilo zníženie prenosovej rýchlosti o cca 37%.

Konzistencia

HBase vyžaduje silnú konzistenciu na úkor dostupnosti. Tento fakt sme zaznamenali počas testovania, keď v určitých intervaloch došlo k zlyhaniu operácie zápisu, ktorá bola následne zopakovaná.

Čítanie dát

Pri vytváraní tabuľky v systéme HBase chýba automatická podpora Bloom filtrov, ktoré sú neusale vo vývoji. Aktiváciu týchto filtrov je potrebné vykonať z príkazového interpreta, ktorý slúži pre manipuláciu so systémom HBase. Bloom filtre boli počas testu aktivované. Výsledky testovacieho prípadu pre čítanie dát zaznamenáva tabuľka 7.3.

Počet uzlov	Replikácia	Čas	Riadok/sek	Priepustnosť [MB/s]
1	1	246	4069	4
3	1	117	8561	8
3	3	127	7936	8
6	3	190	11904	12

Tabuľka 7.3: HBase: čítanie riadkov o veľkosti 1000 B

Zaťažovací test

V tabuľke 7.4 sú znázornené výsledky zaťažovacieho testu.

Počet uzlov				
Veľkosť riadku	3	4	5	6
1 KB	32	35	36	38
10 KB	31	37	41	49
100 KB	35	43	52	55
512 KB	25	40	51	63
1 MB	35	47	53	68

Tabuľka 7.4: Hbase: maximálna priepustnosť klastru v MB/s

7.5 Cassandra

Pri testovaní klastru bol použitý distribuovaný databázový systém Cassandra verzie 0.7.3. Adresár obsahujúci súbory typu commitlog bol na samostatnom fyzickom disku, dátový adresár bol na diskoch zapojených v poli RAID0. Veľkosť štruktúry Memtable bola 120 MB. Každý uzol zaberal na hašovom kruhu rovnaký úsek.

Tabuľka 7.5 obsahuje výsledky z viacerých meraní, podľa testovacieho prípadu pre zápis dát ???. Ako hodnoty kľúčov boli pre jednotlivé zapisované riadky použité prirodzené čísla z rozsahu nula až celkový počet riadkov. Cassandru sme nastavili tak, aby používala RandomPartitioner. Toto nastavenie zabezpečilo rovnomernú záťaž každého uzlu počas celkovej doby zápisu. Výsledok hašovacej funkcie MD5 aplikovaný na hodnotu kľúča určil uzol, do ktorého bol daný riadok zapísaný.

Škálovateľnosť

Z výsledkov meraní je vidieť, že tento distribuovaný databázový systém je maximálne škálovateľný z pohľadu rýchlosti zápisu. V prípade, keď sme zdvojnásobili počet uzlov z troch na šesť (riadok 3,5) vzrástla priepusnosť zápisu o cca 50%.

Replikácia

V prípade zvýšenia hodnoty replikácie z 1 na 3 sa automaticky znížila rýchlosť zápisu o jednu tretinu.

Konzistencia

Počas zápisu s konzistenciou kvóra, ktorá zabezpečuje silnú konzistenciu databázového systému, sa rýchlosť znížila podľa očakávaní. V tomto prípade, aby klient obdržal odpoveď o úspešnom zápise museli byť dáta zapísané na celkový počet replík $N/2 + 1$, kde N označuje počet uzlov v klastri. Počas zápisu v prípade konzistencie One, klient obdržal potvrdenie o úspešnosti po zápise na jeden uzol.

Čítanie dát

Výsledky testovacieho prípadu pre čítanie dát sú zaznamenané v tabuľke 7.6. Riadky boli čítané v náhodnom poradí a kešovanie kľúčov a riadkov, ktoré Cassandra podporuje bolo vypnuté. Počas tejto operácie boli automaticky aplikované Bloom filtre.

Zaťažovací test

Výsledky zaťažovacieho testu zobrazuje tabuľka 7.7, kde jednotlivé položky zobrazujú priemerný dátový tok počas doby testovania v MB/s. Počas testu sme klaster monitorovali a zistil viaceré závažných dôsledkov. Na všetkých uzloch prebiehali veľmi časté GC kolekcie (angl. Garbage collections) z dôvodu častého zápisu štruktúr Memtable na pevný disk. Podľa hardverovej špecifikácie pre systém Cassandra všetky uzly disponovali minimálnou veľkosťou operačnej pamäti 4 GB. Z dôvodu stability systému bola horná hodnota pri ktorej sa zapisuje Memtable z operačnej pamäti na disk 120 MB. Následkom tohto nastavenia vznikalo veľké množstvo SSTable súborov na pevnom disku, ktoré Cassandra zlievala na pozadí (angl. compactions), čo spôsobovalo záťaž vstupno výstupných operácií (angl. I/O wait). V prípade, takto zaťaženého systému a veľkého množstva SSTable súborov by bola operácia čítania veľmi pomalá, pretože dáta patriace do jednej rodiny stĺpcov by boli uložené vo veľkom množstve samostatných súborov a ich načítanie by vyžadovalo zvýšené množstvo diskových operácií (angl. seek).

Z tohoto testu ďalej vyplynulo pozorovanie, že v prípade zápisu malých súborov rádovo v kB, je hlavným úzkym hrdlom systému CPU, kdežto v prípade zápisu veľkých blokov dát zohrávajú hlavnú úlohu V/V diskové operácie.

Počet uzlov	Replikácia	Konzistencia	Čas	Riadok/sek	Priepustnosť [MB/s]
1	1	One	338	23669	23
3	1	One	207	38647	38
3	3	One	311	25723	25
3	3	Quorum	351	22792	22
6	3	One	202	39604	39
6	3	Quorum	263	30418	30

Tabuľka 7.5: Zápis riadkov o veľkosti 1000 B

Počet uzlov	Replikácia	Konzistencia	Čas	Riadok/sek	Priepustnosť [MB/s]
1	1	One	383	261	2.5
3	1	One	211	4745	4.6
3	3	One	51	19630	19
3	3	Quorum	103	9750	10
6	3	One	32	30903	30
6	3	Quorum	59	16924	17

Tabuľka 7.6: Čítanie riadkov o veľkosti 1000 B

Počet uzlov				
Veľkosť riadku	3	4	5	6
1 KB	18	28	30	33
10 KB	63	77	93	118
100 KB	71	92	111	134
512 KB	67	87	109	127
1 MB	62	92	100	126

Tabuľka 7.7: Cassandra: maximálna priepustnosť klastru v MB/s

7.6 Voľba databázového systému

Primárnym požiadavkom aplikácie je zvládať vysoký tok pre zápis, nízkonákladová administrácia (infraštruktúra HBase je komplikovaná v porovnaní s Cassandrou, ktorá je decentralizovaná), preto sme pre implementáciu zvolili systém Cassandra.

Kapitola 8

Návrh systému

V tejto kapitole popíšeme návrh systému, ktorý bude slúžiť na archiváciu emailov a splňať požiadavky, ktoré sme pre tento systém definovali. V prvej časti sa zameriame na výber vhodných open source nástrojov pre implementáciu prototypu a následne popíšeme dosiahnuté výsledky v testovacom prostredí, ktoré preukážu vhodnosť využitia NoSQL systému Cassandra pre riešenie tejto úlohy.

8.1 Zdroj dát

Základným prvkom, ktorý budeme v našom systéme archivovať je emailový objekt, ktorý definuje dokument RFC 2821 [2]. Tento objekt pozostáva z SMTP obálky a emailovej správy. Obálka obsahuje informácie, ktoré sú potrebné pre korektné doručenie správy pomocou emailového servera a patria tam napríklad odosielateľ emailového objektu a jeden alebo viacerý príjemcovia. Emailová správa predstavuje semištrukturovaný dokument [36] v textovej podobe, ktorého syntax popisuje štandard RFC 2822 [1] z roku 2001 pod názvom Formát Internetovej správy (angl. Internet Message Format). Dokument RFC 2822 nahradzuje a upravuje pôvodné RFC 822 pod názvom Štandard pre formát Internetových textových správ ARPA z roku 1982 (angl. Standard for the Format of ARPA Internet Text Messages). Obsah emailovej správy delíme na hlavičku a telo, ktoré sú od seba oddelené znakom reprezentujúcim prázdny riadok. Telo správy nie je povinné. Štruktúru tela správy a polia v hlavičke rozširujú štandardy, pod názvom MIME (ang. Multipurpose Internet Mail Extensions), RFC 2045, RFC 2046 [16, 17], RFC 2047, RFC 2048 a RFC 2049. Tieto štandardy pridávajú možnosť použitia iných znakových sád ako US-ASCII, ďalej umožňujú štruktúrovať telo správy (vnorené správy rfc822), definujú formát a typy pre zasielanie príloh atď.

Zber emailových objektov je realizovaný na unixových serveroch, ktoré budú používať emailový server Qmail¹. Tento server bude zároveň realizovať antispamovú kontrolu pomocou modulu Qmail-scanner², ktorý je naprogramovaný v jazyku Perl³. Po doručení emailového objektu na server je emailový objekt spracovaný programom Qmail, ktorý volá obslužný modul qmail-scanner a následne dokončí doručenie správy. Tento modul sme vhodne

¹<http://cr.yp.to/qmail.html>

²<http://qmail-scanner.sourceforge.net>

³<http://www.perl.org>

modifikovali pre potreby nášho systému. Modifikovaný súbor je súčasťou zdrojových kódov tejto práce. Výstupom je dvojica súborov a to obálka, ktorá obsahuje viacero štatistických údajov a samotný textový súbor reprezentujúci emailovú správu v pôvodnej podobe. Príklad obálky znázorňuje obrázok 8.1. Detailný popis tejto štruktúry sa nachádza na webovej adrese <http://qmail-scanner.sourceforge.net/>.

```
Tue, 15 Mar 2011 10:12:09 CET Clear:RC:1(88.208.65.55):SA:1 0.007811 9508
odosielatel@server prinemca@server2 predmet <1300180228103914546@aq>
1300180329.16836-0.forid1:5987 priloha1:134
```

Obr. 8.1: Obsah obálky z programu qmail-scanner

8.2 Analýza dát

Jedný z hlavných požiadavkov systému je deduplikácia príloh emailových správ z dôvodu úspory diskovej kapacity. Hlavička s názvom „Content-Type“, ktorú definuje RFC 2045 špecifikuje typ dát v tele MIME správy. Jej hodnota je tvorená z dvoch častí a to názov typu média (angl. media type) a bližšie špecifikovaný podtyp, napríklad „image/gif“. Norma definuje základných päť typov médií a to text, image, audio, video a application. V prípade, našej aplikácie má zmysel využiť deduplikáciu na všetky tieto typy s výnimkou typu „text/plain“, kde predpokladáme, že sa jedná o bežnú textovú správu napísanú užívateľom.

Program pre analýzu a deduplikáciu emailovej správy bol napísaný v programovacom jazyku Python⁴. Tento program dodržiava špecifikáciu RFC 2822 a RFC2045. Medzi povinné polia hlavičky emailu patria pole „From“ a Date. Aj napriek tomu, že tieto polia sú definované už od roku 1982 v RFC 1982 analýza nášho datasetu ukázala, že XY % emailov tento požiadavok nespĺňa. Z celkovej množiny emailov o veľkosti 98000 bolo 0.022% emailov, ktoré nespĺňali štruktúru definovanú normou RFC 2045. Metódu deduplikácie sme riešili nasledujúcim postupom:

- analýzou emailu sme určili časti, v ktorých sa nachádzajú prílohy
- H je výstup hašovacej funkcie SHA2 nad dátami reprezentujúcimi prílohu, ktorý sme použili ako unikátny identifikátor prílohy
- dáta reprezentujúce prílohu v emaille sme nahradili značkou v tvare MARK:H
- dáta reprezentujúce prílohu sme do databáze uložili pod kľúčom H

8.3 Databázová schéma

Databázovú schému sme navrhli s ohľadom na to aké operácie nad danými dátami budeme vykonávať a pri návrhu sme využili poznatky získané štúdiom architektúry databázového systému Cassandra. Schéma je tvorená pomocou štyroch rodín stĺpcov a to:

⁴<http://python.org>

- `messagesMetaData` obsahuje meta informácie identifikované v obálke emailového objektu a emailovej správy, nad ktorými budeme vykonávať štatistické výpočty pomocou metódy `MapReduce`
- `messagesContent` obsahuje obálku, hlavičku a telo správy
- `messagesAttachment` slúži na ukladanie deduplikovaných príloh emailov
- `lastInbox` v chronologickom časovom poradí, podľa hodnoty `Date` v hlavičke emailu, zaznamenáva emaily daného užívateľa

Tradičné techniky pre popis databázových schém nie je možné aplikovať na databázové systémy ako napríklad Bigtable alebo Dynamo. Jedným z dôvodom je, že na tieto schémy sa aplikuje denormalizácia, duplikácia dát a kľúče sú často komplexného charakteru. Dodnes neexistuje, žiadny štandard, ktorý by definoval popis týchto schém. Článok pod názvom techniky pre definíciu štruktúr pomocou diagramov v cloude a návrhové vzory (angl. Cloud data structure diagramming techniques and design patterns [11]) definuje stereotypy pre diagramy v jazyku UML a obsahuje vzory pre popis štruktúry týchto dát. Obrázok 8.2 znázorňujúci databázovú schému nášho modelu aplikuje tieto techniky.

Ako jedinečný identifikátor emailovej správy v databáze využívame nasledujúcu schému:

```
emailID = sha256(uid + MessageId + date)
```

V tejto schéme reprezentuje:

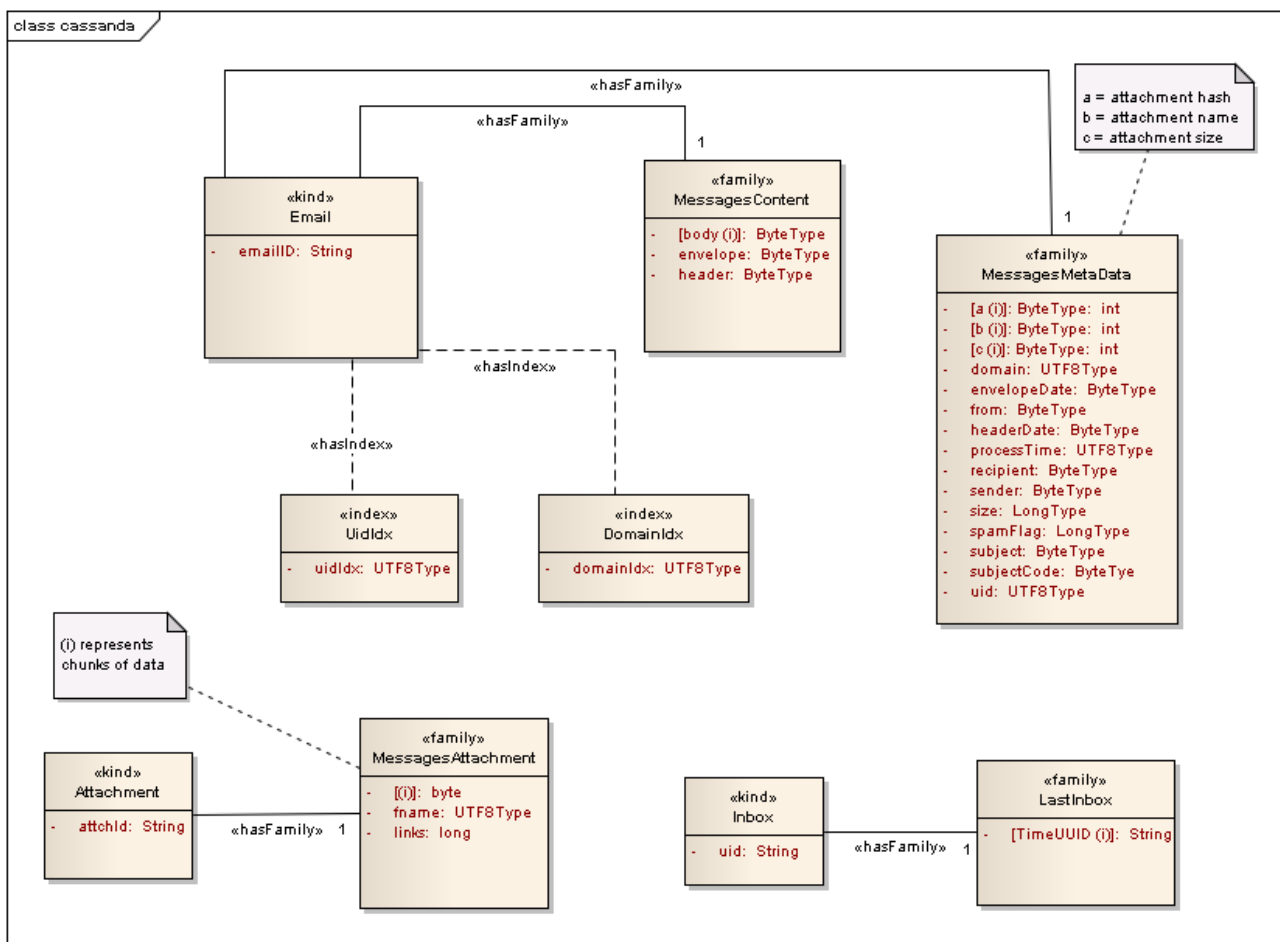
- `uid` emailovú adresu príjemcu, v tvare `jan@mak.com`
- `MessageID` je ide identifikátor z hlavičky daného emailu
- `date` je časová značka reprezentujúca čas kedy bol email prijatý emailovým serverom (formát: rok, mesiac, deň, hodina, minúta, sekunda)
- `emailID` je výstup hašovacej funkcie SHA2 v hexadecimálnom tvare

V predchádzajúcej kapitole sme zistili, že Cassandra nie je optimalizovaná pre zápis blokov dát (ang. blob), ktorých veľkosť prevyšuje 5 MB avšak optimálne výsledky pre zápis dosahuje pri veľkosti blokov 512 kB. Preto prílohy a dáta reprezentujúce email v prípade, že ich veľkosť presahuje 1 MB zapisujeme do samostatných stĺpcov o veľkosti 512 kB. Toto rozdeľovanie dát na menšie bloky vykonávame na aplikačnej úrovni na strane klienta. Názvy stĺpcov číslujeme vzostupne v rozmedzí 0-N, kde N je počet blokov. Spätnú rekonštrukciu dát vykonáva klient.

8.4 Fultextové vyhľadávanie

Fultextové vyhľadávanie realizujeme pomocou samostatného NoSQL systému Elasticsearch⁵. Hlavný index, ktorý obsahuje všetky zaindexované dáta má názov `emailArchive`, a delíme ho na dva typy s názvom `email` a `envelope`. Schéma týchto typov obsahuje polia podľa, ktorých chceme v emailovom archíve vyhľadávať a jej reprezentáciu zapísanú vo formáte JSON znázorňuje obrázok 8.3.

⁵<http://elasticsearch.org>



Obr. 8.2: Databázová schéma

8.5 Implementácia

V programovacom jazyku Python sme implementovali klienta pre zápis dát do databázy Cassandra a Elasticsearch. Jednou z najdôležitejších vlastností týchto klientov je voľba úrovne konzistencie pri zápise. Našou prioritou je integrita dát a od databázy požadujeme silnú konzistenciu. Zvolili sme mód kvóra (ang. Quorum), ktorá zabezpečí zápis dát na $N / 2 + 1$ replík a klient následne obdrží potvrdenie o úspešnosti zápisu, inak sa zápis zopakuje. Klient, ktorý slúži na čítanie dát z databázy využíva taktiež mód kvóra. Tieto vlastnosti nám zabezpečujú, v prípade použitia faktoru replikácie tri (dáta sa v databázovom systéme nachádzajú trikrát), silnú úroveň konzistencie na strane klienta a databázy. Analýza emailovej správy a jej deduplikácia spotrebuje hlavne CPU zdroje. Moderné procesory obsahujú viacero jadier, tento fakt môžeme využiť pre paralelizované spracúvanie emailov, teda každé jadro CPU bude spracúvať súčasne jednu emailovú správu.

```
mappingsEmail = {
  "inbox": {"type": "string"},
  "from": {"type": "string"},
  "subject": {"type": "string"},
  "date" : {"type": "date"},
  "messageID" : {"type": "string", "index": "not_analyzed"},
  "attachments": {"type": "string"},
  "size": {"type": "long", "index": "not_analyzed"},
  "body": {"type": "string"}
}

mappingsEnvelope = {
  "sender": {"type": "string"},
  "recipient": {"type": "string"},
  "ip": {"type": "ip"},
  "date": {"type": "date"}
}
```

Obr. 8.3: JSON schéma pre fultextové vyhľadávanie

Celery

Paralelizáciu našej aplikácie sme zabezpečili pomocou využitia asynchrónnej fronty úloh pod názvom Celery⁶, ktorá využíva architektúru distribuovaného predávania správ (ang. distributed message passing). Architektúru znázorňuje obrázok 8.4. „Pracovníci“ (angl. workers) reprezentujú samostatné procesy v našom prípade proces pre analýzu a deduplikáciu emailu, ktoré môžu bežať paralelne. Ako sprostredkovateľ (ang. broker) je použitá aplikácia RabbitMQ⁷, ktorý obdrží od klienta správu a uloží ju do fronty. Správa obsahuje identifikátor emailovej správy, v tomto prípade cestu na lokálnom súborovom systéme k súboru reprezentujúcom email. Táto správa je následne zaslaná ľubovoľnému pracovníkovi (v našom prípade klient vykonávajúci analýzu a deduplikáciu emailu), ktorý ju spracuje. Táto architektúra je plne distribuovaná, dokáže odolávať chybám (napr. v prípade výpadku elektrickej energie, správy naďalej pretrvávajú vo fronte).

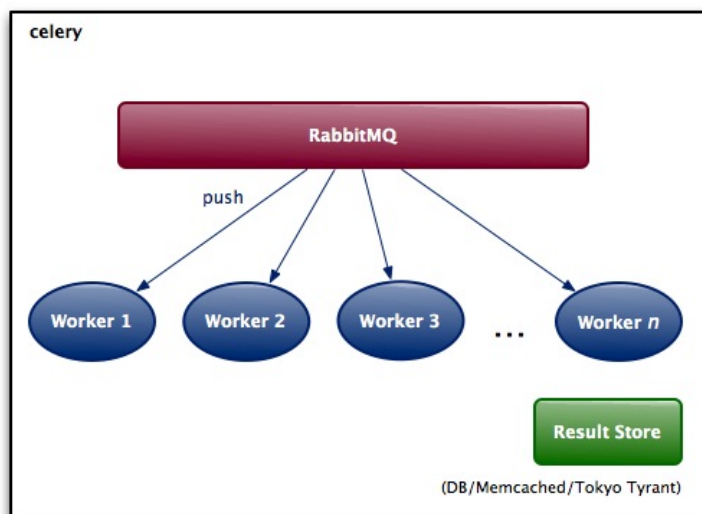
Frontu RabbitMQ sme podrobili výkonnostnému testu, zvládala XY správ pre zápis XY pre načítanie správ.

8.5.1 Klient

Proces spracovania nového emailu je znázornený pomocou sekvenčného diagramu na obrázku 8.5. Pri príchode nového emailu, ktorý je spracovaný emailovým serverom Qmail, je vytvorená nová úloha pomocou aplikácie Celery. Táto úloha uloží do sprostredkovateľa identifikátor emailu. V prípade, že je v daný okamžik k dispozícii ľubovoľný pracovník, je emailová správa

⁶<http://celeryproject.org>

⁷<http://www.rabbitmq.com>



Obr. 8.4: Architektúra Celery, Zdroj: [online], <http://ask.github.com/celery/getting-started/introduction.html>

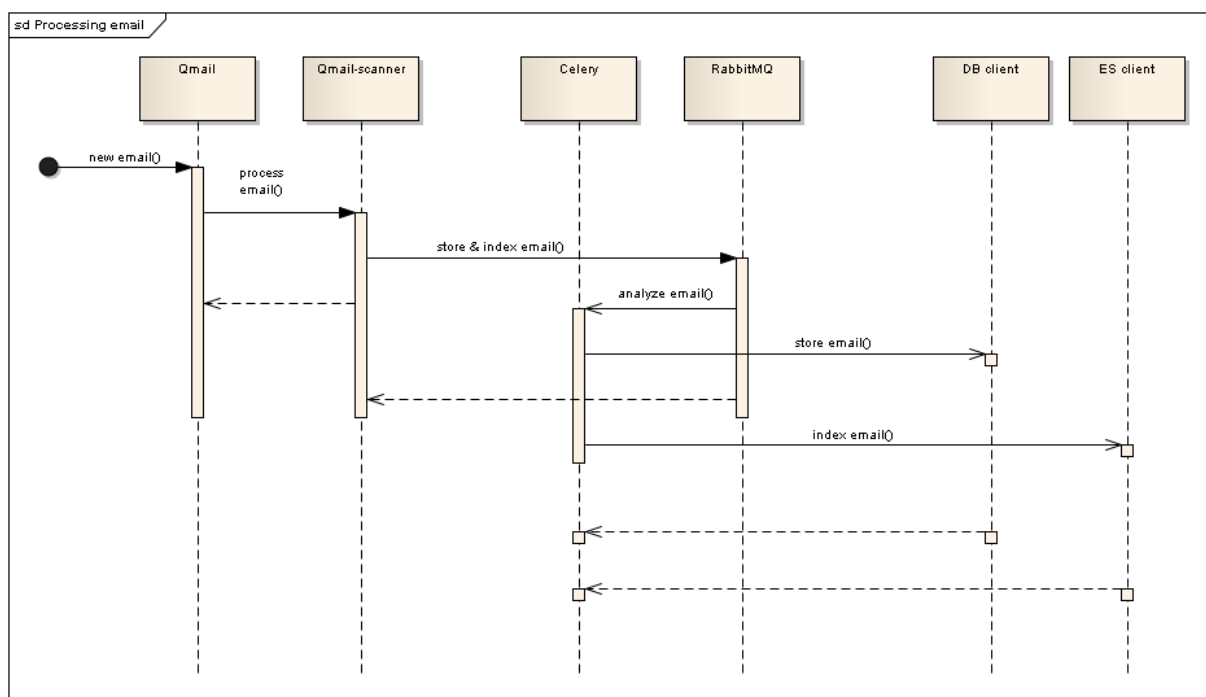
spracovaná pomocou nášho analyzátoru a následne zapísaná do databáze Cassandra a full-textového systému Elasticsearch. Na testovacie účely sme nemali k dispozícii reálny dátový tok emailov. Vrstvu reprezentujúcu Qmail sme nahradili modulom, vytvárajúcim nové úlohy prechádzaním lokálneho súborového systému, ktorý obsahoval testovaciu množinu emailových správ.

8.5.2 Výpočet štatistik

Cassandra spolupracuje so systémom Hadoop, čo nám dáva do rúk mocný nástroj na masívne paralelné spracovanie dát pomocou techniky MapReduce. Tvorba aplikácií v tomto frameworku prebieha v Jave, je náročná a okrem toho programový model MapReduce obsahuje viacero problémov. Model napríklad neobsahuje primitíva na filtrovanie, agregáciu, spájanie dát a je potrebná ich vlastná implementácia. Tieto nedostatky rieši nástroj Pig[] vďaka, ktorému sme boli schopní spracúvať meta dáta uložené v databáze. Obrázok 8.6 zobrazuje programovú ukážku, ktorá slúži na výpočet veľkosti najväčšieho emailu pre domény, ktorých emaily archivujeme. Pre jednoduchosť sme vynechali časti, ktoré slúžia na načítanie dát z databáze a obsahujú užívateľsky definovanú funkciu v programovacom jazyku Java, ktorá slúži na predprípravu vstupných dát. Podpora užívateľom definovaných funkcií je jednou z ďalších výhod nástroja Pig.

8.5.3 Webové rozhranie

TODO web IFACE – prístup koncových užívateľov k archivu a diskusia ohľadne bezpečnosti - stručny popis django aplikácie.



Obr. 8.5: Spracovanie emailu

```

notSpam = FILTER grp BY group.spam == 1;
maxSize = foreach grp {
    size = rows.size;
    generate group, MAX(size);
};
STORE maxSize into 'biggestEmailPerDomainDomain' using PigStorage(',');
  
```

Obr. 8.6: Programová ukážka v jazyku Pig

8.6 Overenie návrhu

Pomocou vyššie popísaného návrhu a implementovaných nástrojov sme overili funkčnosť nami navrhovaného modelu. Vhodná voľba daných verzií u aplikácií Celery a RabbitMQ bola určená počas písania a ladenia aplikácie. Všetky tieto aplikácie sú neustále vo vývoji, to isté platí pre databázu Cassandra a systém Hadoop. Počas písania tejto práce prebehlo viacero rozhovorov s autormi týchto aplikácií. Konkrétne databáza Cassandra na začiatku práce neobsahovala takmer žiadnu ucelenú dokumentáciu, počas začiatkov experimentov sme začínali s verziou 0.7.0. Počas ukončovania tejto práce je aktuálna verzia 0.8.⁸ a medzitým vznikla kvalitná dokumentácia od spoločnosti Datastax⁸.

⁸<https://datastax.com>

Konfigurácia

Hardverová konfigurácia bola totožná s konfiguráciou z kapitoly ???. Na šiestich serveroch bola nainštalovaná databáza Cassandra 0.7.3, Hadoop 0.20.2, dvojica serverov obsahovala klientskú aplikáciu a Celery 2.6. V úlohe sprostredkovateľa bola použitá aplikácia RabbitMQ 2.1.1, ktorá bola nainštalovaná na samostatnom serveri.

Overenie integrity dát

Databázový kluster sme naplnili testovacími dátami obsahujúcimi emaily o objeme 300 GB. Následne sme nasimulovali prípad obnovy dát z archívu, kde sme všetky emaily v náhodnom poradí z databázy načítali, zostavili ich do pôvodného tvaru klientskou aplikáciou a porovnali sme ich odtlačok pomocou hašovacej funkcie MD5 s odtlačkom pôvodných dát. Tento test prebehol bez akejkoľvek chyby.

Dosiahnuté výsledky

Dôležitým pozorovaním bol fakt, že z celkového objemu emailových správ 300 GB sa po deduplikácií príloh tento objem znížil na 99 GB, teda došlo k 30% úspore diskovej kapacity. Štruktúry do ktorých sme ukladali dáta pre potrebu štatistík zaberali 0,4% z celkového objemu dát, čo je zanedbateľná položka.

ElasticSearch rychlost pre fultextové vyhľadavanie bola do XY ms.

8.6.1 Doporučenie najvhodnejšieho systému

Použiť Brisk - nástupca Cassandri s natívnou podporou Hadoop-u, alebo nejaký key-value (to by bolo ale treba zanalyzovať a zmerať performance, je nutné to teda vobec spomínať?)

Kapitola 9

Záver

Literatúra

- [1] Internet message format, 2001.
- [2] Simple mail transfer protocol, 2001.
- [3] D. Abadi, P. Boncz, and S. Harizopoulos. Column-oriented database systems. *Proceedings of the VLDB Endowment*, 2(2):1664–1665, 2009.
- [4] D. Abadi, P. Boncz, and S. Harizopoulos. Column-oriented database systems, VLDB Tutorial. 2009.
www.cs.yale.edu/homes/dna/talks/Column_Store_Tutorial_VLDB09.pdf.
- [5] P. Bernstein and N. Goodman. Concurrency control in distributed database systems. *ACM Computing Surveys (CSUR)*, 13(2):185–221, 1981.
- [6] P. Bernstein, V. Hadzilacos, and N. Goodman. *Concurrency control and recovery in database systems*, volume 5. Addison-wesley New York, 1987.
- [7] A. Bondi. Characteristics of scalability and their impact on performance. In *Proceedings of the 2nd international workshop on Software and performance*, pages 195–203. ACM, 2000.
- [8] E. Brewer. Towards robust distributed systems. In *Proceedings of the Annual ACM Symposium on Principles of Distributed Computing*, volume 19, pages 7–10, 2000.
- [9] F. Chang, J. Dean, S. Ghemawat, W. Hsieh, D. Wallach, M. Burrows, T. Chandra, A. Fikes, and R. Gruber. Bigtable: A distributed storage system for structured data. *ACM Transactions on Computer Systems (TOCS)*, 26(2):1–26, 2008.
- [10] E. F. Codd. A relational model of data for large shared data banks. *Commun. ACM*, 13:377–387, June 1970.
- [11] David Salmen, Tatiana Malyuta, Rhonda Fettes, Normert antunes. Cloud Data Structure Diagramming Techniques and Design Patterns, 2010.
- [12] J. Dean and S. Ghemawat. MapReduce: Simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113, 2008.
- [13] J. Dean and S. Ghemawat. Mapreduce: simplified data processing on large clusters. *Commun. ACM*, 51:107–113, January 2008.

- [14] G. DeCandia, D. Hastorun, M. Jampani, G. Kakulapati, A. Lakshman, A. Pilchin, S. Sivasubramanian, P. Voshall, and W. Vogels. Dynamo: amazon's highly available key-value store. *ACM SIGOPS Operating Systems Review*, 41(6):205–220, 2007.
- [15] A. Fox, S. Gribble, Y. Chawathe, E. Brewer, and P. Gauthier. Cluster-based scalable network services. In *ACM SIGOPS Operating Systems Review*, volume 31, pages 78–91. ACM, 1997.
- [16] N. Freed and N. Borenstein. Multipurpose internet mail extensions (mime) part one: Format of internet message bodies, 1996.
- [17] N. Freed and N. Borenstein. Multipurpose internet mail extensions (mime) part two: Media types, 1996.
- [18] A. Ganesh, A. Kermarrec, and L. Massoulié. Peer-to-peer membership management for gossip-based protocols. *IEEE transactions on computers*, pages 139–149, 2003.
- [19] J. F. Gantz, C. Chute, A. Manfrediz, S. Minton, D. Reinsel, W. Schlichting, and A. Toncheva. The diverse and exploding digital universe. *IDC White Paper*, 2, 2008.
- [20] J. F. Gantz, J. Mcarthur, and S. Minton. The expanding digital universe. *Director*, 285(6), 2007.
- [21] S. Ghemawat, H. Gobioff, and S.-T. Leung. The google file system. In *Proceedings of the nineteenth ACM symposium on Operating systems principles*, SOSP '03, pages 29–43, New York, NY, USA, 2003. ACM.
- [22] D. Gifford. Weighted voting for replicated data. In *Proceedings of the seventh ACM symposium on Operating systems principles*, pages 150–162. ACM, 1979.
- [23] S. Gilbert and N. Lynch. Brewer's conjecture and the feasibility of consistent, available, partition-tolerant web services. *SIGACT News*, 33:51–59, June 2002.
- [24] J. Gray et al. The transaction concept: Virtues and limitations. In *Proceedings of the Very Large Database Conference*, pages 144–154. Citeseer, 1981.
- [25] T. Haerder and A. Reuter. Principles of transaction-oriented database recovery. *ACM Computing Surveys (CSUR)*, 15(4):287–317, 1983.
- [26] N. Hayashibara, X. Defago, R. Yared, and T. Katayama. The ϕ accrual failure detector. In *Symposium on Reliable Distributed Systems (SRDS'2004)*, pages 66–78. Citeseer, 2004.
- [27] E. Hewitt. *Cassandra: the definitive guide*. O'Reilly Media, Inc., 2010.
- [28] D. Karger, E. Lehman, T. Leighton, R. Panigrahy, M. Levine, and D. Lewin. Consistent hashing and random trees: distributed caching protocols for relieving hot spots on the world wide web. In *Proceedings of the twenty-ninth annual ACM symposium on Theory of computing*, STOC '97, pages 654–663, New York, NY, USA, 1997. ACM.

- [29] A. O. R. W. Paper. Why Cloud-Based Security and Archiving Make Sense, March 2010.
www.google.com/postini/pdf/why_cloud_based_wp.pdf, stav z 28.2.2011.
- [30] A. O. R. W. Paper. Why Cloud-Based Security and Archiving Make Sense, March 2010.
http://www.strozzi.it/cgi-bin/CSA/tw7/I/en_US/nosql, stav z 28.2.2011.
- [31] D. Pritchett. Base: An acid alternative. *Queue*, 6:48–55, May 2008.
- [32] T. Segaran and J. Hammerbacher. *Beautiful data: the stories behind elegant data solutions*. O'Reilly Media, 2009.
- [33] R. Shoup. The eBay Architecture, Striking a balance between site stability, feature velocity, performance, and cost, November 2006.
www.addsimplicity.com/downloads/eBaySDForum2006-11-29.pdf, stav z 28.2.2011.
- [34] M. Stonebraker. The case for shared nothing. *Database Engineering Bulletin*, 9(1):4–9, 1986.
- [35] A. Thusoo, Z. Shao, S. Anthony, D. Borthakur, N. Jain, J. Sen Sarma, R. Murthy, and H. Liu. Data warehousing and analytics infrastructure at facebook. In *Proceedings of the 2010 international conference on Management of data*, SIGMOD '10, pages 1013–1020, New York, NY, USA, 2010. ACM.
- [36] J. Udell. *Practical Internet GroupWare*. O'Reilly & Associates, Inc., Sebastopol, CA, USA, 1st edition, 1999.
- [37] W. Vogels. Eventually consistent. *Communications of the ACM*, 52(1):40–44, 2009.
- [38] T. White. *Hadoop: The Definitive Guide*. O'Reilly Media, Inc., 1st edition, 2009.

Dodatok A

Zoznam použitých skratiek

WPAN Wireless Personal Area Network

Dodatok B

Inštalačná a užívateľská príručka

B.0.2 Inštalácia simulátoru OMNeT++ pre platformu Linux

1. Stiahnutie archívu obsahujúceho zdrojový kód zo stránok
<http://www.omnetpp.org/omnetpp>
2. Prekopírovanie archívu do adresára /usr/local/
3. Rozbalenie archívu pomocou príkazu `tar zxvf omnetpp-4.0.src.tgz`
4. Do užívateľského profilu `.bash_profile` alebo `.profile` pridáme riadok
`export PATH=$PATH:/usr/local/omnetpp-4.0/bin`
5. Je potreba zabezpečiť prítomnosť nasledujúcich balíkov v systéme

```
sudo apt-get install build-essential gcc g++ bison flex perl tcl8.4 tcl8.4-dev  
tk8.4 tk8.4-dev blt blt-dev libxml2 libxml2-dev  
zlib1g zlib1g-dev libx11-dev
```

6. Prevedieme nasledujúce príkazy:
`cd /usr/local/omnetpp-4.0`
`./configure`
`./make`
7. Spustenie OMNeT++ s IDE pomocou príkazu `omnetpp`

B.0.3 Inštalácia mnou modifikovaného Mobility Frameworku

1. Stiahnutie súborov Mobility frameworku z svn `http://my-svn.assembla.com/svn/mframework/`,
poprípade prekopírovanie adresára `mf2o4` z priloženého CD do adresára `/usr/local/`
2. Import MF do aplikácie OMNeT++
 - (a) Po spustení aplikácie Omnet, klikneme na oblasť „Project explorer“, pravým tlačítkom
a zvolíme položku „Import...“
 - (b) Zvolíme „General->Existing project into Workspace“

- (c) V položke „Select root directory“, zvolíme cestu k adresáru mf2o4, tj. /usr/local/mf2o4
- (d) Pomocou CTRL+B, preložíme zdrojové súbory

B.0.4 Práca s modelom IEEE 802.15.4

Vo vývojom prostredí Omnetu si otvoríme v oblasti „Project explorer“ adresárovú štruktúru mf2o4, kde si následne otvoríme adresár networks a v ňom adresár ieee802.15.4. V tomto adresári sa nachádzajú aj xml súbory popisujúce antény. Otvoríme si súbor omnetpp.ini, tento súbor je hlavným konfiguračným súborom modelu simulácie. Zahŕnul som do neho ukážkové nastavenia viacerých modelov, ktoré som simuloval. Samotná simulácia sa potom spustí otvorením súboru omnetpp.ini a následným kliknutím na tlačítko „Run“ z menu aplikácie.

Dodatok C

Obsah priloženého CD

Následující obrázok ?? zobrazuje štruktúru priloženého CD.