

Sem vložte zadání Vaší práce.

ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE
FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
KATEDRA SOFTWAREVÉHO INŽENÝRSTVÍ



Diplomová práce

Umístění dat na výpočetní uzly minimalizující datové přenosy v databázi HBase

Bc. Miroslav Hrstka

Vedoucí práce: Ing. Adam Šenk

13. dubna 2015

Poděkování

Děkuji Ing. Adamovi Šenkovi za vedení diplomové práce, cenné připomínky a rady, které přispěli k úspěšnému vypracování této práce. Poděkování patří také Pavle Příbylové za nemalou pomoc s formální stránkou práce.

Prohlášení

Prohlašuji, že jsem předloženou práci vypracoval(a) samostatně a že jsem uvedl(a) veškeré použité informační zdroje v souladu s Metodickým pokynem o etické přípravě vysokoškolských závěrečných prací.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona, ve znění pozdějších předpisů. V souladu s ust. § 46 odst. 6 tohoto zákona tímto uděluji nevýhradní oprávnění (licenci) k užití této mojí práce, a to včetně všech počítačových programů, jež jsou její součástí či přílohou, a veškeré jejich dokumentace (dále souhrnně jen „Dílo“), a to všem osobám, které si přejí Dílo užít. Tyto osoby jsou oprávněny Dílo užít jakýmkoli způsobem, který nesnižuje hodnotu Díla, a za jakýmkoli účelem (včetně užití k výdělečným účelům). Toto oprávnění je časově, teritoriálně i množstevně neomezené. Každá osoba, která využije výše uvedenou licenci, se však zavazuje udělit ke každému dílu, které vznikne (byť jen zčásti) na základě Díla, úpravou Díla, spojením Díla s jiným dílem, zařazením Díla do díla souborného či zpracováním Díla (včetně překladu), licenci alespoň ve výše uvedeném rozsahu a zároveň zpřístupnit zdrojový kód takového díla alespoň srovnatelným způsobem a ve srovnatelném rozsahu, jako je zpřístupněn zdrojový kód Díla.

V Praze dne 13. dubna 2015

.....

České vysoké učení technické v Praze
Fakulta informačních technologií

© 2015 Miroslav Hrstka. Všechna práva vyhrazena.

Tato práce vznikla jako školní dílo na Českém vysokém učení technickém v Praze, Fakultě informačních technologií. Práce je chráněna právními předpisy a mezinárodními úmluvami o právu autorském a právech souvisejících s právem autorským. K jejímu užití, s výjimkou bezúplatných zákonných licencí, je nezbytný souhlas autora.

Odkaz na tuto práci

Hrstka, Miroslav. *Umístění dat na výpočetní uzly minimalizující datové přenosy v databázi HBase*. Diplomová práce. Praha: České vysoké učení technické v Praze, Fakulta informačních technologií, 2015.

Abstrakt

Tato práce si dává za cíl zmapovat možnosti optimalizace datových přenosů v databázi HBase. Je analyzováno stávající řešení navržené pro souborový systém HDFS. Na základě zjištěných skutečností je zvážena možnost použít principiálně podobné schéma optimalizace pro HBase a následně je toto řešení navrhuto a naimplementováno. Výsledek optimalizace je nutné následně otestovat a vyhodnotit.

Klíčová slova HBase, Hadoop, distribuované databáze, optimalizace

Abstract

This work is aimed to map out the options for optimizing data transfers in the database HBase. It analyzed existing solutions designed for HDFS file system. Based on the findings is considered for use principally similar pattern optimization HBase and then this solution is designed and is deployed. The result of optimization is necessary then test and evaluate.

Keywords Hbase, Hadoop, distributed database, optimization

Obsah

Úvod	1
1 Současný stav a použité technologie	3
1.1 Hadoop Ekosystém	3
1.2 HDFS	8
1.3 HBase	9
1.4 HyperGraph Partitioning	16
2 Existující řešení optimalizace distribuce dat pro MapReduce	19
2.1 Definice problému	19
2.2 MR-part	20
2.3 Testování a výsledky	23
3 Návrh řešení pro HBase	25
3.1 Klíčová specifika HBase pro návrh řešení	25
3.2 PATOH algoritmus	29
4 Implementace řešení	33
4.1 Monitoring	33
4.2 Repartitioning	34
5 Testování a vyhodnocení měření	37
Závěr	39
Literatura	41
A Generátor zkušebních dat	43
B Testovací MapReduce úlohy	45

C Instalace Hadoop prostředí	47
D Seznam použitých zkratk	49
E Obsah přiloženého CD	51

Seznam obrázků

1.1	Diagram MapReduce procesu	5
1.2	Hadoop Ekosystém	6
1.3	Diagram uložení souboru v systému HDFS	9
1.4	Datový model HBase	10
1.5	Architektura databáze HBase	11
1.6	Příklad B+ stromu.	12
1.7	Iterativní Merge Multi-Page bloků v LMS stromu	13
1.8	Schéma ukládání dat z HBase v HDFS	14
1.9	Formát souboru HFile	14
1.10	Rozložení Hfile do bloků dat v HDFS	15
1.11	Formát KeyValue struktury	15
1.12	Reprezentace hypergrafu	17
1.13	bisekce hypergrafu	17
2.1	MR-part schéma	21
2.2	Pseudokód algoritmu pro Metadata Combination	22
2.3	Pseudokód algoritmu pro Repartitioning	23
3.1	Řazení řádků za pomoci přidání prefixu	26
3.2	Vstupní formát hypergrafu	30
4.1	Rozšíření třídy TableInputFormatBase	33
4.2	Rozšíření třídy RecorReader	34

Úvod

Současný stav a použité technologie

V této úvodní kapitole je dán prostor pro seznámení s technologiemi a projekty, se kterými se bude buď přímo pracovat nebo je jejich znalost pro pochopení problematiky nezbytná. Jako první je představen projekt HadoopTM od firmy ApacheTM jako celek. Vše, co bude v této práci představeno, se bude odehrávat v rámci tohoto takzvaného ekosystému. Pro porozumění základní myšlenky projektu Hadoop je také nezbytné vysvětlit programovací model MapReduce. Po uvedení celého Hadoopu jsou detailněji uvedeny produkty, které jsou součástí tohoto ekosystému a se kterými se bude dále pracovat. Jedná se především o databázový systém HBase a také souborový systém HDFS, který je v celém modelu využíván.

1.1 Hadoop Ekosystém

Apache Hadoop je framework který sdružuje projekty vyvíjející software pro spolehlivé, škálovatelné a paralelní zpracování dat na počítačových clusterech. Je založen na dvou stěžejních technologiích pocházejících od firmy Google a to na distribuovaném souborovém systému Google File System (GFS) a na algoritmu MapReduce[1]. Všechny klíčové projekty v systému Hadoop jsou združeny pod Apache Software foundation, která poskytuje podporu pro tyto projekty. Jedná se o open.source software a všechny komponenty jsou psány v programu Java.

1.1.1 Základní principy

Podstata Hadoopu spočívá v uložení dat na velkém množství výpočetních úzlů spojených do clusterů. Většinou se jedná o běžný hardware. Na těchto uzlech jsou data uložena ve vlastním souborovém systému HDFS. K výpočtům nad clusterem se využívá princip Mapreduce, který bude osvětlen v následující

kapitole. Systém Hadoop je charakteristický především následujícími vlastnostmi, které ho odlišují od klasických databázových systémů.

Horizontální škálovatelnost a komoditní hardware

Pro objemy dat, jimiž by se měl Hadoop při svém zpracovávání primárně zabývat, je poměrně složité a především velmi drahé dosáhnout dostatečné škálovatelnosti pomocí vertikálního škálování, tedy přidávání výkonu a zdrojů ke stávajícím výpočetním uzlům. Proto Hadoop využívá horizontálního škálování. Díky horizontálnímu škálování se nabízí využití komoditního hardwaru namísto specializovaných výpočetních uzlů. Systém tedy běží na velkém množství samostatných počítačů spojených do clusteru.

Řešení selhání hardwaru

S předchozím bodem úzce souvisí řešení případných výpadků jednotlivých uzlů. Kvůli velkému počtu výpočetních uzlů a díky použití běžného hardware jsou výpadky poměrně časté. Hadoop je ale navržen tak, že se nesnaží tyto výpadky minimalizovat, ale počítá s nimi. Data jsou dostatečně replikována a pokud dojde k výpadku při zpracování dat, je přerušená úloha provedena na jiném uzlu obsahující danou replikaci a zároveň je automaticky vytvořena nová záloha dat. Defaultně je zálohovací faktor nastaven na 3, tedy každý soubor je uložen ve třech kopiích na různých výpočetních uzlech.

Přenášení kodů k datům

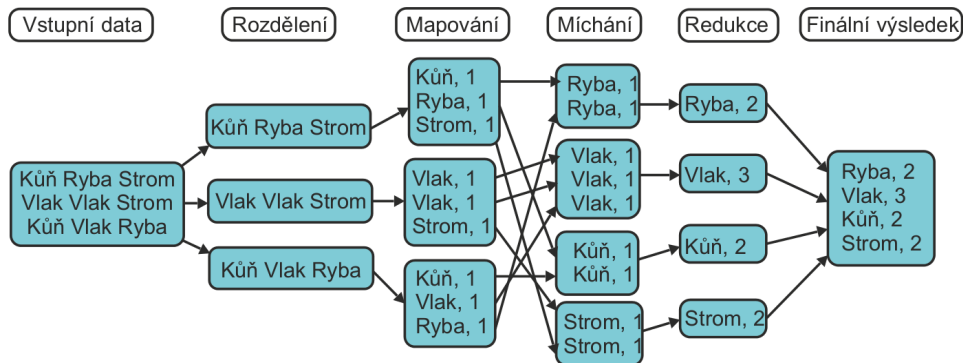
Kvůli velkým objemům dat a poměrně nízké propustnosti sítě spojující jednotlivé výpočetní uzly v clusteru je velmi výhodné namísto rozesílání dat na jednotlivé výpočetní uzly rozesílat na ně pouze kód výpočtu a minimalizovat tak přesuny dat mezi uzly. Na každém uzlu je tak vykonán výpočet pokud možno s lokálními daty.

Abstrakce od distribuovaných a paralelních aplikací

Hadoop se snaží co nejvíce odstínit vývojáře Hadoop aplikací od řešení zpracování pomocí paralelního a distribučního zpracování. Proto poskytuje poměrně jednoduché a dobře definované rozhraní pro jednotlivé komponenty. Při práci s těmito rozhraními tak není nutné řešit, jak se bude kód v clusteru distribuovat ani další záležitosti spojené s paralelním zpracováním a dovoluje se zaměřit na business logiku aplikace. Cenou za toto zjednodušení je pak právě omezené rozhraní bez možnosti detailnějšího nastavení.

1.1.2 MapReduce

MapReduce je programovací paradigma určené k provádění výpočtů, které by za normálních okolností trvaly značné množství času a to zejména z důvodů



Obrázek 1.1: Diagram MapReduce procesu

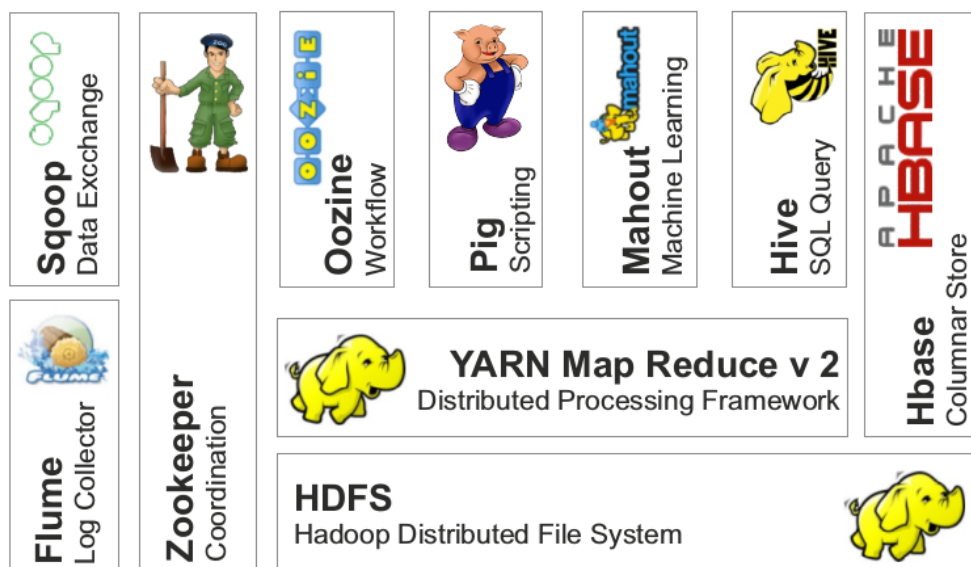
velkého množství dat. Tyto výpočty se pak snaží dokončit v přijatelném časovém horizontu. Princip byl poprvé představen v roce 2004 v práci pro firmu Google jako jedno z opatření pro zvládání obrovského množství dat, se kterým se museli potýkat.

V MapReduce jsou data modelována do párů klíč/hodnota. Tento formát je velmi jednoduchý, přesto se dají téměř všechny údaje takto prezentovat. Tato jednoduchá datová struktura tak umožňuje snadné zpracování dat efektivním způsobem. Klíč a hodnota může být cokoliv. Může se jednat o řetězce, čísla nebo komplexní struktury.

Mapreduce se skládá ze dvou hlavních fází, mapovací fáze a redukční fáze. Nejdříve je spuštěna mapovací funkce, která je dále spočtena nad jednotlivými prvky z množiny uspořádaných dvojic (klíč, hodnota), jež produkuje přechodné klíče a hodnoty.[2] Poté jsou všechny tyto přechodné hodnoty asociované se stejným klíčem seskupeny a následně poslány se do redukční fáze. Redukční fáze přijme na vstupu klíč a související množinu hodnot. Jako výstup je pak množina uspořádaných dvojic, která splňuje zadaná kritéria.

Jak vstupní data, tak mezivýsledky i finální výsledky jsou ve formátu klíč/hodnota. Jak je vidět na obrázku 1.1 probíhají mapovací a redukční fáze paralelně. Z diagramu je také zřejmé, že mezi mapovací fází a fází míchání(shuffling) dochází k přenosu průběžných výsledků mezi jednotlivými výpočetními uzly. Právě optimalizací přesunů v těchto místech se bude zabývat hlavní část této práce. MapReduce je také často popisována funkcí:

$$\begin{aligned} \text{map} &: (k_1, v_1) \rightarrow \text{list}(k_2, v_2) \\ \text{reduce} &: (k_2, \text{list}(v_2)) \rightarrow \text{list}(k_3, v_3) \end{aligned}$$



Obrázek 1.2: Hadoop Ekosystém

1.1.3 Hadoop Ekosystém

Následující kapitola je určena k seznámení s ekosystémem Hadoop, který kromě základních projektů MapReduce a HDFS obsahuje množství dalších projektů. Složení ekosystému se ale liší v závislosti na konkrétní distribuci Hadoopu. Je sice možné zvolit si tyto aplikace podle svého výběru a využít přímo zdroje od firmy Apache, ale v praxi se využívají spíše již částečně nakonfigurované distribuce, které navíc nabízejí i možnost placené podpory. Mezi největší hráče na trhu patří distribuce od firem Cloudera, MapR a Hortonworks.[3] Pro účely této práce byla vybrána distribuce Cloudera, protože se jedná o open-source projekt a také kvůli největšímu podílu na trhu.1.2

Nástroje pro vývoj

YARN

YARN je klíčovým prvkem Hadoop 2. Někdy je také zvaný MapReduce v2. Jedná se o distribuovaný operační systém, který odděluje řízení zdrojů a řízení kapacit od zpracovávající komponenty. To umožňuje podporovat větší škálu přístupů ke zpracování dat a širší pole aplikací.

Hive

Hive umožňuje dotazování nad velkými datasety uloženými v distribuovaném systému a také jejich řízení. Poskytuje mechanismus k vytvoření struktury nad těmito daty a následně nad daty provádět dotazy v SQL-like jazyku HiveQL. Kromě toho umožňuje také využití klasického map/reduce postupu v případech, kdy není výhodné použít HiveQL.

Pig

Pig poskytuje prostředí pro zpracování jednoduchého skriptovacího jazyka Pig Latin, ve kterém je přeložen na sérii MapReduce úloh. Pig Latin abstrahuje od MapReduce schématu a nabízí dotazování na vyšší úrovni, podobné jako SQL.

Mahout

Mahout je škálovatelná knihovna pro strojové učení. Jsou v ní implementovány algoritmy pro clustering, klasifikaci a kolaborativního filtrování optimalizované pro běh v prostředí Hadoopu.

Ukládání dat a správa metadat

HDFS

Jedná se o distribuovaný souborový systém navržený pro provoz na komoditním hardwaru ve velkých datových skladech, souborový systém HDFS bude detailněji představen v následující kapitole.

HBase

Hbase je sloupcově orientovaný databázový systém, který běží nad HDFS. Nepodporuje strukturovaný dotazovací jazyk a poskytuje prakticky pouze CRUD operace. HBase bude stejně jako HDFS představen detailněji v následujících kapitolách.

Nástroje pro řízení

ZooKeeper

Poskytuje provozní služby pro Hadoop cluster. Jedná se o distribuované konfigurační, synchronizační služby a o jmenné registry pro distribuovaný systém.

Oozie

Aplikace používaná pro plánování Hadoop úloh. Je složena ze dvou hlavních částí. V první části se ukládají a spouštějí různé typy hadoop úloh (Mapreduce, Pig, Hive, atd.) a z části, která koordinuje běh daných úloh na základě předdefinovaných plánů a dostupnosti dat.

Získávání a agregace dat

Sqoop

Nástroj sloužící k efektivnímu přenosu dat z relačních databází do Hadoopu k dalšímu zpracování. Zpracovat tyto data pak může buď MapReduce úloha nebo jiný nástroj (Hive, Pig). Je také možné data vložit do HBase databáze.

Flume

Služba pro efektivní získávání, agregování a přesouvání velkého množství streamovaných dat do HDFS. Typicky se používá k ukládání logů z jednoho zdroje (webové logy, bankovní logy) a agreguje je v HDFS pro pozdější zpracování.

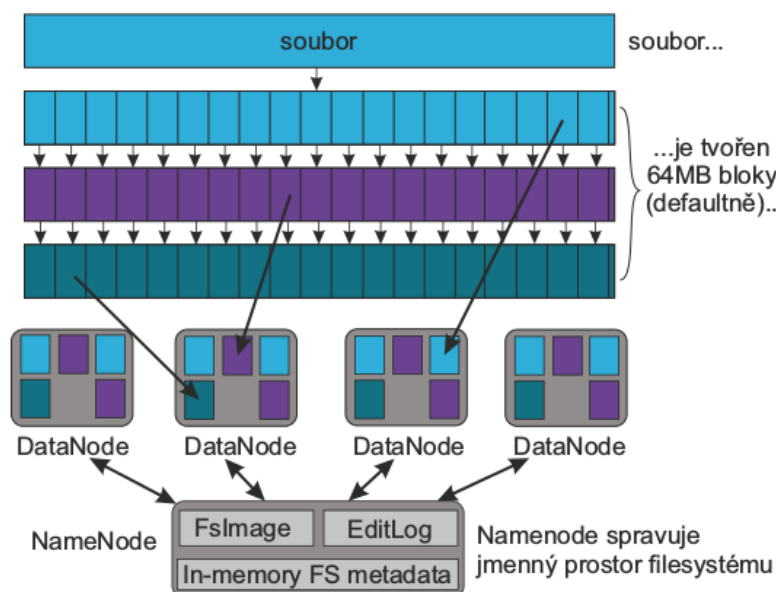
1.2 HDFS

Hadoop Distributed File Systém (HDFS) nabízí způsob skladování velkých souborů na více samostatných počítačích, který je rozdílný oproti klasickému přístupu skladování dat na jednom stroji s dostatečnou diskovou kapacitou. HDFS je navržen na základech Google File Systemu (GFS) a běží na nativním filesystému (Ext3, Ext4, XFS). HDFS je určen pro skladování především velkých souborů (100 MB a více) v menším počtu (řádově miliony) a k ukládání streamovaných dat. Systém dále není vhodný pro soubory, u nichž se očekává časté upravování, a to protože je možné připisovat data pouze na konec souboru. Systém je odolný proti chybám v replikaci a výpadkům v distribuci dat.[4]

HDFS umožňuje, stejně jako většina běžných souborových systémů, operace čtení, zápisu a mazání souborů a vytváření a mazání adresářů. Vždy, když je načten nový soubor do HDFS, je zreplikován do žádoucího počtu, který určuje replikační faktor (defaultní hodnota je 3) a rozdělen do bloků dat o fixní délce (defaultně 64MB). Tyto bloky jsou pak rozdistributedy a uloženy ve vybraných uzlech clusteru určených pro skladování, tzv. DataNodes viz.1.3 . V HDFS se informace o souborech neukládají společně s daty, ale jsou uloženy na vyhrazeném serveru nazývaném NameNode. Při přístupu k datům klient nejdříve zadá požadavek na data na NameNode, který následně vrátí adresy databloků s požadovanými daty. NameNode tedy přímo nemanipuluje s daty.

NameNode uchovává a poskytuje strom jmenného prostoru a adresy fyzického umístění bloků ve své operační paměti. Dále ukládá perzistentní záznam těchto adres (kontrolní bod) a registr modifikací (žurnál) pro zotavení z havárie v nativním systému souborů hostitelského počítače. HDFS umožňuje i vytvoření kopie kontrolního bodu a žurnálu na další výpočetní uzel nazývaný SecondaryNameNode. Ten pak slouží jako záloha dat serveru NameNode (nenahrazuje tedy funkci primárního NameNode v případě výpadku, pouze poskytuje data pro jeho obnovu). Ve verzi Hadoop 2+ je už možné mít Standby NameNode, který v případě výpadků může primární NameNode plně a okamžitě nahradit.

Přístupovat k HDFS je možné přímo a to přes nativního klienta nebo pomocí Java nebo C++ API. Dále je možný přístup přes proxy server podporující REST, Thirft a Avro server.



Obrázek 1.3: Diagram uložení souboru v systému HDFS

1.3 HBase

Jedná se o sloupcově orientovanou databázi, někdy označovanou jako Hadoop databáze. HBase podporuje náhodné real-time CRUD operace (narozdíl od HDFS). Je primárně navržena pro uchovávání velkých tabulek o miliardách řádků a milionech sloupcích a jedná se o NoSQL databázi. Nepodporuje tedy přístup založený na SQL jazycích ani relační model. Stejně jako HDFS se vyznačuje jednoduchým klientem a Java API. HBase je založena na projektu Bigtable od Googlu a stejně jako byl Bigtable postaven nad GFS je HBase postavena nad HDFS.[5]

HBase nebyla zavedena za účelem nahrazení klasických RDBMS a ani k tomuto účelu není využívána. HBase je výhodné použít, jak již bylo řečeno, v případě rozsáhlých tabulek. Výborné výsledky vykazuje při vykonávání jednotlivého náhodného výběru z databáze a při vyhledávání dle klíče. Hbase je také vhodným řešením v případě, že jednotlivé řádky tabulky jsou velmi různorodé a v případě řídkých databází, kdy je velký počet sloupců a většina z nich obsahuje nulovou hodnotu. Nevhodné využití je pak právě pro suplování úloh pro tradiční RDBMS jako jsou transakční aplikace nebo relační analýza.[6]

1.3.1 Data model HBase

Data v databázi HBase jsou uložena v tabulkách. Jednotlivé tabulky obsahují řádky. Na každý řádek odkazuje unikátní klíč. Jako hodnota klíče se bere bi-

KEY	Timestamp	JMÉNO (FAMILY)			ADRESA (FAMILY)		
		Titul	Jméno	Příjmení	Adresa	Číslo	PSČ
KEY 001	t 1	Bc.	Franta	Omáčka	Kostomlaty	33	413 01
	t 2	Mgr.					
	t 5						
KEY 002	t 10		Laura	Tlustá	Praha	133	15000
	t 15				Olomouc	480	68300
	t 17			Hubená	Brno	1153	77100

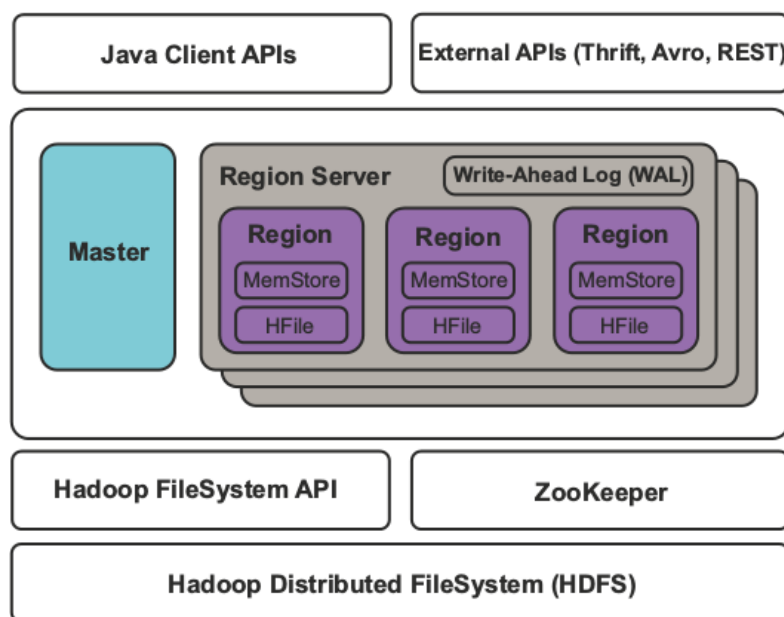
Obrázek 1.4: Datový model HBase

tové pole. Klíč u HBase tedy může být cokoli, string, long nebo vlastní datová struktura. Každý řádek je složen ze sloupců, které jsou sdruženy do rodin (column families). Tyto rodiny sloupců jsou definovány staticky při vytváření databáze narozdíl od samotných sloupců, které se mohou přidávat libovolně. Data jako taková jsou pak uložena v buňkách. Tyto buňky jsou identifikovány pomocí řádku, rodiny, sloupce a časovou značkou(timestamp). Obsah každé buňky je pak uchováván také jako pole bitů. Data v buňkách jsou navíc verzovány. Každá buňka defaultně uchovává poslední tři zadané hodnoty s tím, že pokud není v dotazu specifikována konkrétní verze, vrací vždy tu nejmladší. Řádky jsou v každé tabulce seřazeny lexikograficky podle svého klíče. Příklad takové tabulky je uveden v obrázku 1.4.

1.3.2 HBase Architektura

HBase je distribuovaná databáze. Proto je i architektura složitější než u databází běžících na jednom výpočetním uzlu. HBase musí řešit všechny problémy typické pro distribuované aplikace jako je koordinace a řízení vzdálených procesů, blokování, distribuce dat a příliš velká síťová komunikace. HBase však k tomuto z velké části využívá služeb v rámci Hadoop a Zookeeper. Následující obrázek 1.5 popisuje hlavní architektonické komponenty HBase.

Jednotlivé tabulky jsou složeny z regionů. Region je vždy určitý rozsah řádků uložený pohromadě. Protože jsou řádky v databázi ukládány v lexikografickém pořadí, je nutné počítat s tím, že se velikost těchto rozsahů, tedy regionů, bude v čase měnit. Proto se v případě, kdy velikost regionu překročí stanovenou hranici, rozdělí region na dva přesně v půli podle prostředního klíče. Naopak v případě, kdy se regiony příliš zmenší, dojde k jejich sloučení. Regiony jsou uloženy v region serverech. Každý region server může obsahovat jeden a více regionů. Region je však vždy jen na jednom serveru. Master server je zodpovědný za správu region serverů. Pro koordinaci se využívá Zookeeper. Na každém regionu je uložen určitý rozsah klíčů. Rozdělení dat do region serverů umožňuje rychlou obnovu v případě pádu region serveru a také ulehčuje load balancing pokud dochází k přetěžování některých serverů. Všechny tyto



Obrázek 1.5: Architektura databáze HBase

činnosti včetně rozdělování velkých regionů jsou prováděny automaticky bez zásahu uživatele.

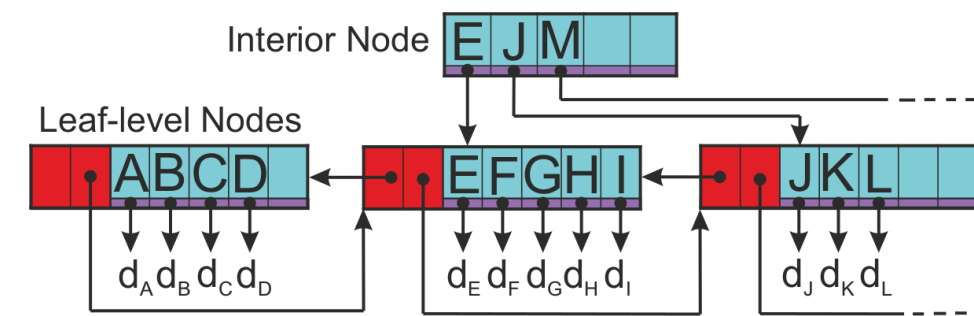
1.3.3 Architektura uložště RDBMS vs HBase

Dříve než bude uveden detailní pohled na uložení dat v databázi HBase, je na místě popsat základní rozdíl mezi typickou architekturou uložení dat v RDBMS a v HBase. Typické RDBMS ukládá data do struktury B+ Trees oproti tomu HBase a ostatní Big Table architektury využívají Log-Structured Merge Trees[7].

1.3.3.1 B+ Trees

Jedná se o stromovou datovou strukturu, která vychází z B-stromu. Umožňuje rychlé vkládání, vyhledávání a mazání dat. Záznamy v tabulkách jsou identifikovány za pomoci klíčů. Všechna data jsou uložena jen na listech stromu, oproti klíčům, které jsou uloženy i ve vnitřních uzlech. V implementaci těchto stromů se přidává do všech listů kromě vlastních klíčů i odkaz na následujícího sourozence. To umožňuje velmi efektivní sekvenční prohledávání aniž by výrazněji stoupla paměťová náročnost na uložení stromu, odkaz na sourozence je v obrázku znázorněn červenými políčky 1.6.

V B+stromech je data-locality dosažena na úrovni stránek, kde stránky odpovídají blokům v jiných systémech. Stránka tak může vypadat například takto:

Obrázek 1.6: Příklad B+ stromu. Obsahuje data d_1 až d_7

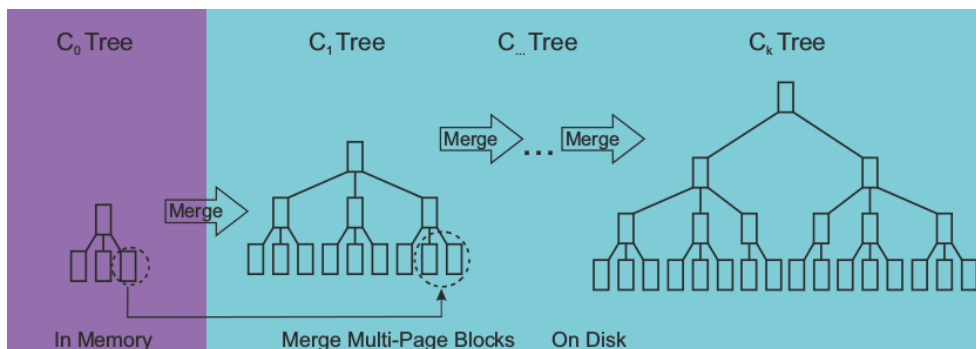
- Odkaz na předchozí stránku
- Odkaz na následující stránku
- key: $A \rightarrow d_A$
- key: $B \rightarrow d_B$
- key: $C \rightarrow d_C$
- key: $D \rightarrow d_D$

Vždy, když se vkládá nový záznam, doplní se daná stránka požadovaným záznamem. Pokud dojde k případu, že stránka už je plná, rozdělí se na 2 poloprázdné stránky a upraví se patřičně rodič těchto stránek. Tímto dochází k fragmentaci dat na disku, když jednotlivé logicky sousedící stránky neleží vedle sebe fyzicky.

1.3.3.2 Log-Structured Merge Trees

Log-Structured Merge Trees využívají způsob odlišný od B+ stromů. Všechny příchozí data jsou nejprve ukládány v logovacích souborech a to kompletně sekvenčně. Jakmile jsou informace uloženy v logu, uloží se data v in-memory uložišti, kde jsou uloženy naposledy upravené záznamy. Vždy, když je k dispozici dostatečný počet záznamů, dojde k zapsání už seřazených dat do uložišť souborů. Po zapsání dat na disk je log soubor smazán.

Datové soubory jsou pak strukturovány podobně jako B stromy, s tím že jsou optimalizovány pro sekvenční přístup na disku. Všechny uzly stromu jsou zaplněny a uloženy v sigle-page nebo multi-page bloku. Při přidávání dat se uložená data na disku v multi-page blocích spojí s příchozími in-memory daty. Tento proces je znázorněn na obrázku 1.7. Vždy, když data využijí celou kapacitu bloků, dojde k vytvoření nového. Postupně se tak zvyšuje počet vytvořených souborů. Tyto soubory jsou pak spojovány do větších celků. Všechny tyto soubory jsou uloženy sekvenčně za sebou a tak je umožněn optimální



Obrázek 1.7: Iterativní Merge Multi-Page bloků v LMS stromu

sekvenční přístup k datům. Strom může být také rozdělen v případě, kdy je potřeba vložit nová data s klíči mezi ostatními. Mazání záznamů se provádí dávkově. Každý záznam, který je určený ke smazání je označen jako smazaný, a při nahlížení do dat se ignoruje. Při přepisování stránek pak dojde k odstranění těchto záznamů ze stromu.

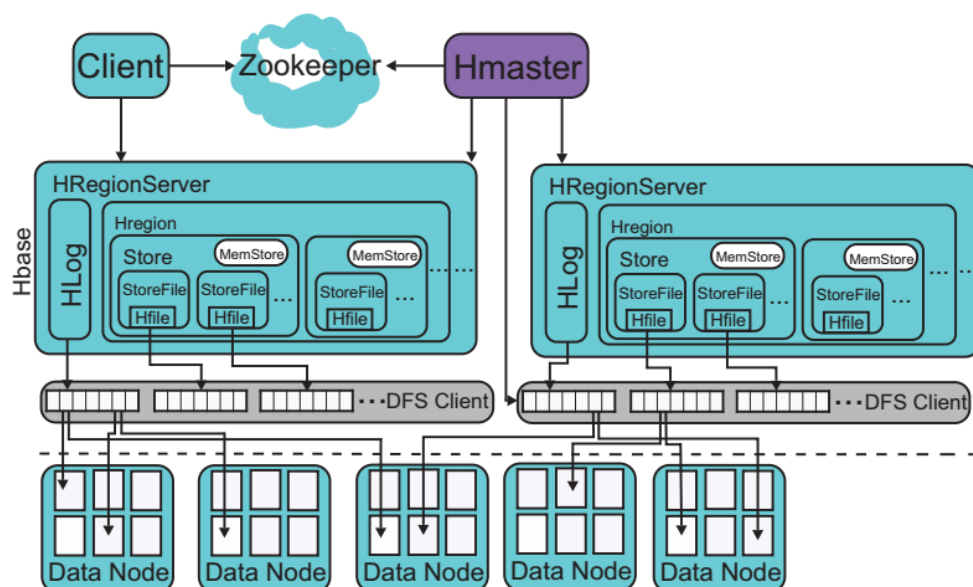
Z těchto rozdílných přístupů k uložení dat vyplývá, že B+ stromy jsou určeny pro úlohy, kde se očekává častá modifikace již vložených dat, zatímco LSM stromy jsou určeny k ukládání velkého množství dat a jeho následnému čtení.

1.3.4 Fyzické uložení dat

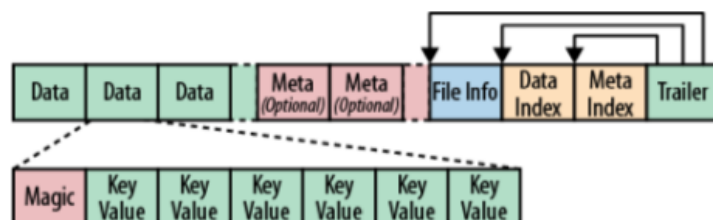
Pro většinu uživatelů je forma fyzického uložení dat v HBase zcela zkrývá a k plnohodnotnému využívání databáze prakticky nevýznamná. Avšak pro potřeby této práce, je nezbytné tuto část osvětlit více, než je zapotřebí v běžném využití. Na obrázku 1.8 je znázorněno schéma uložení dat, které zobrazuje uložení dat v souborovém systému HDFS. Hbase pracuje především s dvěma hlavními typy souborů. se soubory HLog reprezentující write-ahead log (WAL) a HFile pro uložení dat.[8]

1.3.4.1 formát HFile

Formát souboru HFile je navržen tak, aby ukládal data co nejefektivněji. Je založen na formátu souboru TFile z Hadoopu. Soubor obsahuje několik bloků, jak je vidět na obrázku 1.9, z nichž jsou fixní pouze info a trailer bloky. Tyto bloky jsou uloženy na konci souboru a ukončují tak daný soubor, který se stává neměnným. V index blocích jsou uloženy offsety data a metadata bloků. Jak metadata bloky tak i data bloky jsou nepovinné, nicméně už z podstaty věci jsou data bloky téměř vždy součástí Hfile souboru. Velikost jednotlivých bloků je defaultně definována na 64 KB. Je možné tuto velikost změnit v závislosti na očekávané struktuře dotazů na databázi. Čím větší bloky, tím efektivnější



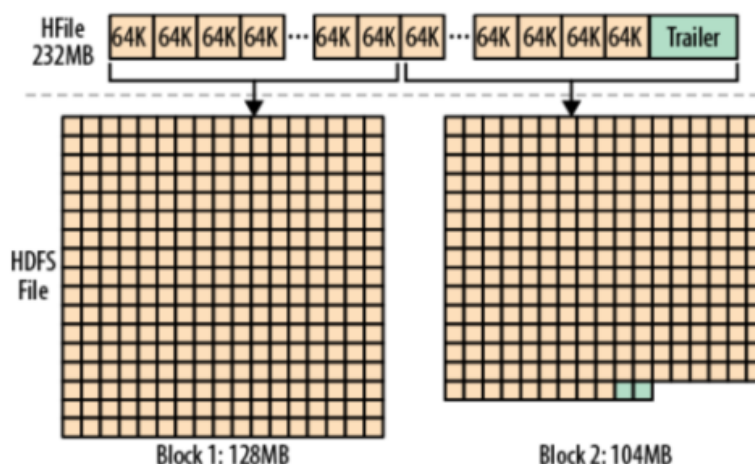
Obrázek 1.8: Schéma ukládání dat z HBase v HDFS



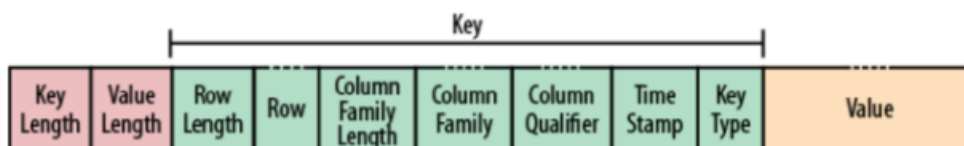
Obrázek 1.9: Formát souboru HFile

bude sekvenční prohledávání databáze, naopak náhodný přístup bude efektivní méně. V jednotlivých blocích se sekvenčně ukládají Key Value instance v Key Value formátu. Velikost bloků není fixní, odvíjí se od velikosti poslední vložené key value instance, která se do bloku uloží celá a pak teprve dojde k uzavření bloku. V praxi tedy boky bývají o něco větší než je definovaná hodnota. Je možné také použít libovolný komprimační algoritmus na ukládané data, což ale neovlivní celý proces, protože checksum je dopočítáván až po přidání poslední key value instance.

Ačkoli se nabízí myšlenka, že velikost bloků v HFile nastavená na 64KB souvisí z velikostí souborů v HDFS, který je nastaven na 64 MB, není mezi těmito hodnotami žádná spojitost. HFile jsou v HBase ukládány jako běžné soubory a HDFS je vnímá pouze jako binární data. Na obrázku 1.10 je zobrazeno schéma ukládání HFile souborů do HDFS.



Obrázek 1.10: Rozložení Hfile do bloků dat v HDFS



Obrázek 1.11: Formát KeyValue struktury

Formát KeyValue KeyValue je struktura, která reprezentuje fyzické uložení jedné buňky z tabulky. Z obrázku 1.11 vyplývá, že jako část klíče se používá mimo jiné název tabulky, family a sloupce. Tento klíč se vyskytuje u každé buňky a proto se doporučuje, mít vše pojmenováno co nejkratšími identifikátory.

1.3.4.2 Write-Ahead Log

Region servery ukládají příchozí data do paměti, do doby než se jich nahromadí dostatečné množství a poté je najednou zapíše na disk. Ve fázi, kdy se data hromadí v paměti, jsou ale zranitelná a v případě výpadku dojde k jejich ztrátě. Právě pro zamezení tohoto problému se používá WAL. Jeho funkce tedy je omezit dopady výpadku. Do logu se proto zapisují všechny změny dat, a až v případě, že jsou úspěšně uloženy na disk, je klient informován o úspěšném uložení dat.

1.4 HyperGraph Partitioning

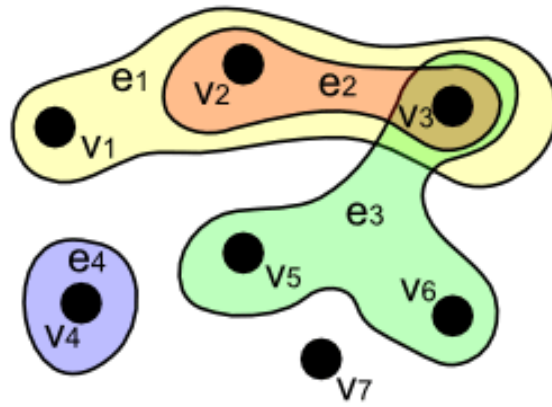
Hypergraph partitioning je známý problém, řešený hlavně při návrhu rozložení VLSI (Very-large-scale integration). Obecně lze říci že se jedná o operaci, která má za cíl rozdělit hypergraf do dvou nebo zhruba stejně velkých částí tak, aby při oddělování bylo přerušeno co nejméně hran spojující jednotlivé uzly v grafu.

Pro řešení problému dělení grafů a hypergrafů se používají především heuristiky založené na algoritmu Kernighan-Lin (KL). Především díky rychlosti běhu a poměrně dobrým výsledkům. KL algoritmus je navržen průvodně pro bipartitioning grafů. Na jeho základech však vznikla většina pokročilejších algoritmu pro dělení grafů. KL algoritmus začíná počátečním rozdělením na dvě části, následně provede určitý počet kroků, dokud se nenajde lokální minimum bipartitivity. Každý krok obsahuje sekvenci prohození vrcholů. Stejnou strategii prohazování vrcholů při řešení problému dělení hypergrafů použili Schweikert-Kernighan[9]. Fiduccia-Mattheyses[10] uvedli rychlejší implementaci KL algoritmu pro dělení hypergrafů. Byla založená na nahrazení vyměňování uzlů za jejich přesouvání. V kombinaci s použitím správných datových struktur výrazně klesla časová náročnost algoritmu. Na těchto základech a mnoha dalších poznatcích je pak vyvinut nástroj PaToH (PaToH:Partitioning Tools for Hypergraphs), který bude využit právě pro řešení dělení hypergrafu v implementační části této práce.

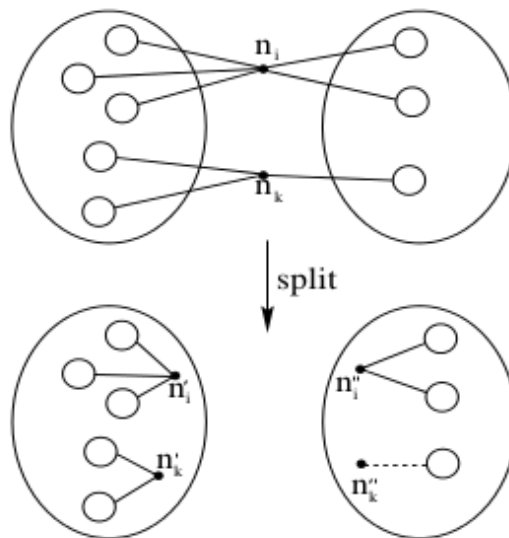
1.4.1 Definice problému

Hypergraf $H = (V, N)$ je definován jako množina vrcholů V a množina sítí (multihran) N propojujících tyto vrcholy. Každá síť $n \in N$ je podmnožina $n \subseteq V$. Vrcholy v síti n jsou nazývány *piny*. Velikost sítě se rovná počtu pinů, které obsahuje. Stupeň vrcholu je počet sítí, kterými je spojen. Hypergraph může být i vážený a v tom případě se mohou ohodnotit jak hrany, tak uzly.

....



Obrázek 1.12: Reprezentace hypergrafu



Obrázek 1.13: bisekce hypergrafu

Existující řešení optimalizace distribuce dat pro MapReduce

V této kapitole bude představeno řešení, které poskytl M. Liroz-Gistau et al. ve svém článku Data Partitioning for Minimizing Transferred Data in MapReduce [11]. V tomto článku se zaměřují na redukování datových přenosů mezi mapovací a redukční fází. V tomto místě přichází na řadu fáze, kde probíhá míchání přechodných klíčů (shuffle phase). Od tohoto řešení se bude poté odvíjet návrh řešení pro databázi HBase. Protože v předchozích kapitolách už byl vysvětlen princip zpracování dat pomocí Map Reduce, může tato kapitola plynule navázat na tyto poznatky.

2.1 Definice problému

Nejdříve je zapotřebí formálně definovat, jaký problém chceme řešit. Mějme tedy sadu MapReduce úloh, které reprezentují typické zatížení systému a sadu vstupních dat. Předpokládejme, že budoucí MapReduce úlohy budou vykonávat podobné úlohy na podobných datech a budou generovat podobné intermediate key (předpokládá se, že v praxi se vykonávají pořád stejné úlohy, jen dochází například ke zvětšování datasetu o nově zapsané data).

Cílem navrhovaného systému je automatické rozdělení vstupních dat tak, aby u budoucího vykonávání MapReduce úloh byl minimalizován přenos dat mezi jednotlivými uzly v shuffle fázi. Při tomto rozdělování se nebere v úvahu plánování mapovacích a redukčních fází, ale pouze inteligentní rozdělení intermediate klíčů mezi jednotlivé redukční uzly.

Definujme daný problém formálně. Mějme vstupní data pro MapReduce úlohu job_α složená z jednotlivých souborů $D = \{d_1, \dots, d_n\}$, které jsou rozděleny do množiny bloků (chunks) $C = \{c_1, \dots, c_p\}$. Funkce $loc : D \rightarrow C$ přiřazuje data do bloků. Nechť job_α je složen z $M_\alpha = \{m_1, \dots, m_p\}$ mapovacích úloh a $R_\alpha = \{r_1, \dots, r_q\}$ jsou redukční úlohy. Předpokládejme, že každá mapovací

úloha m_i zpracuje blok c_i . Nechť $N_\alpha = \{n_1, \dots, n_s\}$ je množina výpočetních uzlů použitých pro provedení úlohy. $node(t)$ reprezentuje výpočetní uzel, kde se vykonává úloha t .

Nechť $I_\alpha = \{i_1, \dots, i_m\}$ je množina intermediate párů klíč-hodnota produkovaných mapovací fází jako je $map(d_j) = \{i_{j_1}, \dots, i_{j_t}\}$. $k(i_j)$ reprezentuje klíč z intermediate páru i_j a $size(i_j)$ reprezentuje celkovou velikost v bytech. Definujeme $output(m_i) \subseteq I_\alpha$ jako množinu intermediate párů produkovaných mapovací úlohou m_i , tedy $output(m_i) = \bigcup_{d_j \in c_i} map(d_j)$. Dále definujeme $input(r_i) \subseteq I_\alpha$ jako množinu intermediate párů přiřazených k redukční úloze r_i . Funkce $part : k(I_\alpha) \rightarrow R$ přiřazuje intermediate klíč k redukční úloze.

Nechť i_j je intermediate klíč-hodnota pár, pak $i_j \in output(m)$ a $i_j \in input(r)$. Nechť $P_{i_j} \in [0, 1]$ je proměnná, která se rovná 0 pokud intermediate pár i_j je vyprodukován na stejném výpočetním uzlu jako je následně zpracováván v redukční části a 1 v opačném případě.

Nechť $W = job_1, \dots, job_w$ je množina všech úloh. Cílem je pak najít optimální loc a $part$ funkce tak aby $\sum_{job_\alpha \in W} \sum_{i_j \in I_\alpha} size(i_j) P(i_j)$ bylo minimální.

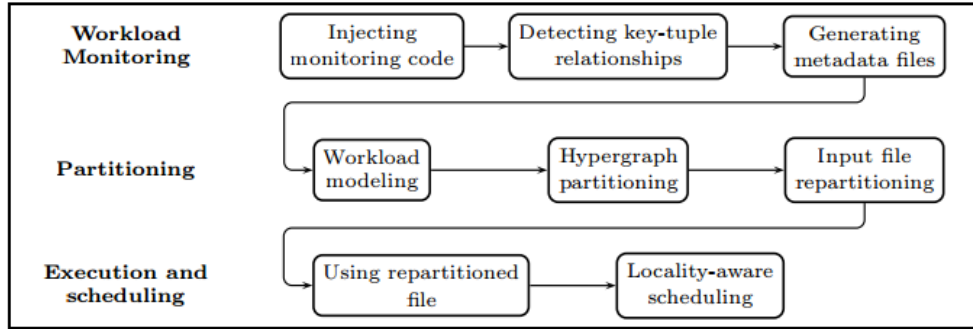
2.2 MR-part

Pro vyřešení zadaného problému byla navržena technika pojmenovaná MR-Part. Tato technika, za pomoci automatického dělení vstupních souborů, dovoluje využít maximální výhody data-locality při plánování redukčních úloh a výrazně snižuje množství dat, které je potřeba přesunout v shuffle fázi. MR-part se skládá ze tří hlavních fází. Jedná se o fáze Workload Monitoring, Partitioning a Execution and scheduling, tak jak je vidět z obrázku 2.1. V první fázi se shromáždí informace o vykonávání jednotlivých MapReduce úloh, které jsou následně zkombinovány. Z těchto informací je vytvořen model zatížení pomocí hypergrafu. V druhé fázi se na vytvořený hypergraf aplikuje dělicí algoritmus, který rozdělí data na požadovaný počet bloků a následně jsou vstupní soubory upraveny na základě tohoto rozdělení. V poslední fázi se využije upravených vstupních souborů a za pomoci optimalizace přiřazování redukčních úloh se dosáhne minimalizace přenosu dat v shuffle fázi.

2.2.1 První fáze - Workload Characterization

Pro zajištění minimalizace přenosů mezi výpočetními uzly při přechodu z mapovací do redukční fáze je nejdříve zapotřebí zjistit, jaké páry hodnot se generují pro vstupní data a následně je vhodně seskupit. K tomu dochází v monitorovací a kombinační části první fáze.

Monitoring Nejprve je zapotřebí získat potřebná data z typických MapReduce úloh, u kterých se očekává jejich častější vykonávání. K zachycení těchto



Obrázek 2.1: MR-part schéma

dat se využívá třída `RecordReader`¹, která je rozšířená o monitorovací funkci. Monitorovací funkce unikátně identifikuje vstupní páry klíč-hodnota a jejich pozici ve vstupních datech. Pro každou mapovací úlohu se tak vytvoří soubor s metadaty. Vždy, když je načten nový blok s daty, je zároveň vytvořen i nový soubor, obsahující informace o bloku. Následně je iniciován record counter(rc). Pokaždé kdy je načten vstupní pár, inkrementuje se counter o 1. Poté pokud dojde k vytvoření vstupního páru, je vygenerován pár (k, rc). Po dokončení zpracování bloku dat jsou takto vygenerované páry uloženy do již vytvořeného souboru ve formátu $\langle k, \{rc_1, \dots, rc_n\} \rangle$.

Combination Následující fáze již neběží zároveň s jinými úlohami, ale pustí je uživatel ideálně v čase, kdy systém není vytížen jinými výpočty. V kombinční fázi se shromáždí a zkombinují metadata z monitorování a na jejich základě se vygeneruje pro každý vstupní soubor hypergraf. Hypergraf $H = (H_V, H_E)$ je graf, kde každá hyperhrana $e \subseteq H_E$ může propojovat více jak dva vrcholy $v \subseteq H_V$. Po zpracování metadat se pak do tohoto hypergrafu uloží každý zpracovávaný prvek (vygenerovaný unikátní identifikátor reprezentující typicky řádek ve vstupním souboru). Poté se přidá hyperhrana, reprezentující klíč a propojí vrcholy, které tento klíč vygenerovaly. Detailní popis algoritmu v pseudokódu je zobrazen na obrázku. 2.2

2.2.2 Druhá fáze - Repartitioning

Nyní, když je vygenerován hypergraf modelující rozložení dat v jednotlivých souborech, je na každý hypergraf aplikován min-cut k-way dělicí algoritmus. Tento algoritmus má jako vstup hodnotu k a hypergraf, ze kterého následně vygeneruje k disjunktních podmnožin vrcholů tak, aby byla minimalizována suma hran mezi uzly rozdílných podmnožin. Parametr k je nastaven podle

¹`RecordReader` je třída, která parsuje vstupní soubor a generuje vstupní páry. Každý datový formát má jiný `RecordReader`. Soubory tedy obvykle používají stále stejný.

Data: F : Input file; W : Set of jobs composing the workload
Result: $H = (H_V, H_E)$: Hypergraph modeling the workload
begin
 $H_E \leftarrow \emptyset$; $H_V \leftarrow \emptyset$
 foreach $job \in |W|$ **do**
 $T \leftarrow \emptyset$; $K \leftarrow \emptyset$
 foreach $m_i \in M_{job}$ **do**
 $md_i \leftarrow getMetadata(m_i)$
 if $F = getFile(md_i)$ **then**
 foreach $\langle k, \{rc_1, \dots, rc_n\} \rangle \in md_i$ **do**
 $\{t_1.id, \dots, t_n.id\} \leftarrow generateTupleID(c_i, \{rc_1, \dots, rc_n\})$
 $T[k] \leftarrow T[k] \cup \{t_1.id, \dots, t_n.id\}$; $K \leftarrow K \cup \{k\}$
 foreach *intermediate key* $k \in K$ **do**
 $H_V \leftarrow H_V \cup T[k]$; $H_E \leftarrow H_E \cup \{T[k]\}$
end

Obrázek 2.2: Pseudokód algoritmu pro Metadata Combination

počtu bloků ve vstupním souboru. Po provedení tohoto algoritmu by měli být v jednotlivých vygenerovaných podmnožinách seskupeny uzly generující stejný klíč. Následně se použijí tyto podmnožiny k vygenerování nových vstupních souborů, kde už jsou data seřazena tak, aby řádky generující stejný klíč byly maximálně seskupeny. Tímto nově vzniklým souborem je následně nahrazen starý vstupní soubor, který je smazán. Pseudokód algoritmu je uveden na obrázku. 2.3 V algoritmu je uvedená funkce RR , která reprezentuje funkci třídy *RecordReader* použitou pro parsování vstupních souborů. Dále se v kódu objevuje funkce RW znamenající *RecordWriter*. Její funkce je inverzní k funkci *RecordReader*. V této části je výpočetně nejsložitější vykonání min-cut algoritmu. Min-cut algoritmus spadá do skupiny NP-Complete problémů. Existuje však několik aproximačních algoritmů, které byly navrženy k řešení tohoto problému. V tomto případě byl použit algoritmus PATOH ²

2.2.3 Třetí fáze - Execution and scheduling

K tomu, aby bylo možné plně využít výhody získané přeskupením záznamů v předchozích fázích, je zapotřebí maximalizovat data locality při plánování redukčních úloh. K tomuto účelu byl upraven algoritmus fairness-locality poskytnutý v [12], který pro každý pár key-value vypočítá skóre reprezentující poměr mezi vyvážeností vstupů do redukční fáze a lokací dat. Každý klíč je zpracován nezávisle pomocí greedy algoritmu. Pro každý klíč jsou pak možné uzly seřazeny podle jejich frekvence výskytu v sestupném pořadí (uzly s vyššími frekvencemi mají lepší data locality). Avšak namísto vybrání uzlu s ma-

²<http://bmi.osu.edu/~umit/software.html>

Data: F : Input file; $H = (H_V, H_E)$: Hypergraph modeling the workload; k : Number of partitions

Result: F' : The repartitioned file

begin

```

 $H_V \leftarrow H_V \cup t_{\text{virtual}}$ 
 $\{P_1, \dots, P_k\} \leftarrow \text{mincut}(H, k)$ 
for  $i \in (1, \dots, k)$  do
   $\mid$  create  $\text{tempf}_i$ 
foreach  $c_i \in F$  do
   $\mid$  initialize( $RR, c_i$ );  $rc \leftarrow 0$ 
  while  $t.\text{data} \leftarrow RR.\text{next}()$  do
     $\mid$   $t.\text{id} \leftarrow \text{generateTupleID}(c_i, rc)$ 
     $\mid$   $p \leftarrow \text{getPartition}(t.\text{id}, \{P_1, \dots, P_k\})$ 
     $\mid$   $RW.\text{write}(\text{tempf}_p, t.\text{data})$ 
     $\mid$   $rc \leftarrow rc + 1$ 
 $(j_1, \dots, j_k) \leftarrow \text{reorder}(\text{tempf}_1, \dots, \text{tempf}_k)$ 
for  $j \in (j_1, \dots, j_k)$  do
   $\mid$  write  $\text{tempf}_i$  in  $F'$ 

```

Obrázek 2.3: Pseudokód algoritmu pro Repartitioning

ximální frekvenci se upřednostní uzel s lepším fairness-locality skóre. To má za následek maximální vyvážení vstupů do redukčních fází.

V MapReduce frameworku je tedy nutné provést následující modifikace:

- Upravení partitioning funkce tak, aby každému intermediate klíči přiřadila unikátní partition.
- Po dokončení mapovací části odeslat do master uzlu seznam s vygenerovanými intermediate klíči a jejich frekvencemi. Tato informace je přiložena k Heartbeat zprávě, která je odesílána po dokončení mapovací úlohy.
- Master přiřadí intermediate klíč k redukční úloze na základě dodaných informací a zajistí tak maximální vyvážení a data-locality v redukční části výpočtu.

2.3 Testování a výsledky

MR-Part byl implementován v Hadoop-1.0.4 a otestován na Grid5000³. Jedná se o rozsáhlou infrastrukturu složenou s rozdílných sítí s několika clustery výpočetních uzlů. Nutno podotknout, že tato infrastruktura není typická pro nasazení Hadoopu a vykonávání MapReduce úloh, především právě kvůli nesooudnosti clusterů v síti, jejich fyzické vzdálenosti a tím i kapacitně omezeném

³<https://www.grid5000.fr/mediawiki/index.php/Grid5000:Home>

propojení. Částečně i proto bylo zrychlení zpracování tak znatelné. K testování byly využity datasety z TPC-H ⁴.

2.3.1 Výsledky

Množství přenesených dat pro různé typy dotazů

Při spuštění několika různých úloh TPC-H ⁵, na původních datasetech bylo v shuffle vázi přeneseno cca. 80% dat. Po provedení dotazů na již upravených datech pak docházelo k přenosu dat nižším než 10% pro všechny provedené dotazy.

Množství přenesených dat v závislosti na velikosti clusteru

Při provedení vybraného dotazu nad konfiguracemi systému s různým počtem clusterů nedocházelo k žádné výrazné změně v podílu přenesených dat. Jednalo se o konfigurace o 5 až 25 uzly v clusteru.

Časová náročnost dotazů

Jak již bylo řečeno, časová náročnost je velmi vázaná na propustnost sítě. Bylo provedeno vykonání dotazů na různých šířkách pásma a výsledek byl očekávaný. Tedy čím pomalejší síť mezi výpočetními uzly, tím výraznější zrychlení při vykonávání dotazů.

⁴<http://www.tpc.org/tpch/>

⁵Implementace použitých dotazů: <http://www.cs.duke.edu/starfish/mr-apps.html>

Návrh řešení pro HBase

Při návrhu řešení pro optimalizaci umístění dat na výpočetní uzly minimalizující datové přenosy V databázi HBase jsem vycházel z řešení pro souborový systém HDFS, které jsem popsal v předchozí kapitole. Nebudu proto uvádět znovu celý návrh řešení, ale pojmu tuto kapitolu jako výčet změn, oproti návrhu pro implementaci v HDFS.

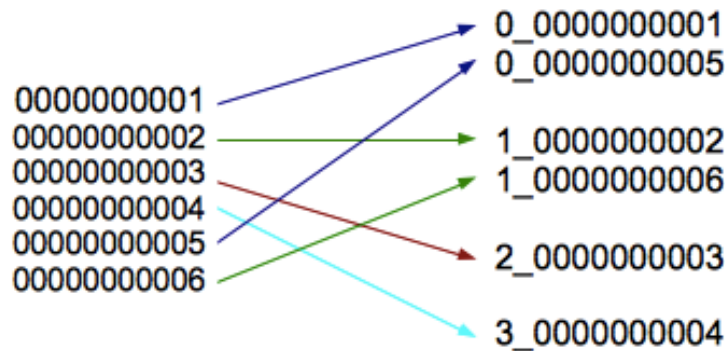
3.1 Klíčová specifika HBase pro návrh řešení

3.1.1 Řazení záznamů v databázi

Jeden z největších změn při návrhu optimalizace pro databázi HBase bylo fyzické uložení jednotlivých řádků databáze. Nativní funkcí HBase je ukládání záznamů v lexikografickém pořadí. Tato funkce je pro HBase klíčová a je charakteristickým prvkem, kterým se odlišuje od jiných databází. Právě tato vlastnost dává HBase velkou sílu při sekvenčním prohledávání ať už celého datasetu nebo určité "výseče" z dat. Tento fakt se ale přímo rozchází se základní myšlenkou optimalizace pro HDFS, kdy je nezbytné změnit fyzické pořadí a tím i uložení záznamů. Bylo tedy nutné navrhnout možná řešení, jak docílit toho, aby bylo možné změnit pořadí záznamů v závislosti na požadavcích optimalizačního algoritmu. V úvahu přicházely dva možné varianty jak se s tímto omezením vyrovnat. V úvahu přicázela změna klíče dle potřeb optimalizačního algoritmu nebo přidání prefixu k již existujícímu klíči 3.1.

Změna klíče řádku podle potřeb optimalizačního algoritmu

Toto řešení se nabízí jako nejjednodušší. Naráží však na výrazný problém. Při návrhu logického modelu pro databáze v HBase se typicky jako klíč pro jednotlivé záznamy používá klíč nesoucí určitou informační hodnotu (například část klíče odpovídá času pořízení daného záznamu). Změnu klíčů by jsme si tedy v tomto případě nemohli dovolit, pro aplikování tohoto přístupu by se nabízely tabulky, kde je klíč automaticky gene-



Obrázek 3.1: Řazení řádků za pomoci přidání prefixu

rován a nemá žádnou spojitost s daty které se nachází v řádku, který reprezentuje.

Přidání prefixu k již existujícímu klíči

Druhým možným řešením je přidání prefixu k již existujícímu klíči. Tento návrh již řeší problém s automatickým řazením, zároveň přidání krátkého prefixu, neznehodnotí informaci uloženou v klíči, jen je nutné při dalších dotazech toto zohlednit a před zpracováním dat prefix s klíče odstranit. Nabízí se více možností, na jakém základě prefixy přiřazovat. Nejefektivnější je přidání určité hodnoty v bitech, na základě požadovaného počtu částí výsledného rozdělení grafu.

3.1.2 Splitting HBase regionů

V návrhu pro implementaci v HBase je zapotřebí zohlednit, jak pracovat s distribucí HBase regionů do souborového systému. Jak již bylo řečeno, v HBase se data v řádcích ukládají do HBase regionů, kde je vždy uložen určitý rozsah lexikograficky seřazených řádků. Řádek je vždy uložen jako celek jen na jednom HRegionu. Není tedy potřeba zohledňovat případné dělení řádku.

Tabulky se typicky skládají z mnoha regionů, které jsou umístěny na mnoha region serverech. Regiony tedy slouží jako mechanismus k rozložení zátěže napříč clusterem. Vždy když se vytváří nová tabulka, vytvoří se pouze jeden region, který se postupně při nabírání objemu štěpí na více regionů, které se distribuují v clusteru. Proto ve fázi nahrávání dat do DB není využita plná kapacita clusteru. Toto chování je zapříčiněno především tím, že dopředu není zřejmé, kde se budou ve výsledné tabulce nalézat split points, tedy místa, kde dojde k rozdělení tabulky na více regionů. Tyto body jsou totiž špatně predikovatelné, protože rozložení klíčů v tabulce nemusí být vždy rovnoměrné. V našem případě však již víme, s jakým rozložením dat budeme pracovat při vytváření tabulky a zároveň budeme vědět, na kolik částí budeme chtít

data rozdělit. Ideální stav by tedy byl, že výsledná data určená dělicím algoritmem jako homogenní část by byla uložena na určitém počtu regionů aniž by v těchto regionech byla uložena data z ostatních částí tabulky. K tomuto rozdělení, můžeme využít pre-splitting [13].

3.1.2.1 Pre-Splitting

Pre-splitting je nástroj, který nám umožňuje vytvořit novou tabulku s již vytvořenými split pointy. Použití tohoto nástroje nám umožní vytvořit si tabulku podle našich potřeb. První z možných řešení je předpočítání splitting pointů podle zadaných kritérií. K tomu slouží `regionSplitter` utility. `RegionSplitter` vytvoří splitting points pomocí `SplitAlgorithm`. Předdefinované algoritmy jsou `HexStringSplit` a `UniformSplit`. První algoritmus je možné použít, pokud klíče obsahují prefixy ve formě hexadecimálního řetězce. Druhý algoritmus určuje splitting point pomocí uniformního rozdělení klíčů. Pro potřeby naší aplikace se nabízí využití prvního algoritmu, kdy dojde k rozdělení podle prefixů, které byli ke klíčům přidány na základě výsledku dělicího algoritmu. Je možné implementovat i vlastní `SplitAlgorithm`, nicméně pro naše potřeby je plně dostačující již implementovaný algoritmus.

Následující příkaz vytvoří tabulku `testTable`, kde `-c 10` definuje počet regionů a `-f f1` definuje požadované families v tabulce.

```
$ hbase org.apache.hadoop.hbase.util.RegionSplitter \
> testTable HexStringSplit -c 10 -f f1
```

Následně dojde k vygenerování tabulky o deseti regionech.

```
15/04/08 18:49:32 DEBUG hbase.HRegionInfo:
> Current INFO from scan results =
> {NAME => 'test_table',
> 135856377.acclad1b7962564fc3a43e5907e8db33.',
> STARTKEY => '', ENDKEY => '19999999',
> ENCODED => acclad1b7962564fc3a43e5907e8db33,}
15/04/08 18:49:32 DEBUG hbase.HRegionInfo:
> Current INFO from scan results =
> {NAME => 'test_table',
> 19999999,135856377.37ec12df6bd0078f5573565af415c91b.',
> STARTKEY => '19999999', ENDKEY => '33333332',
> ENCODED => 37ec12df6bd0078f5573565af415c91b,}
...
```

3.1.2.2 Auto Splitting

Jak již bylo řečeno, pre-splitting je nástroj, který má na starosti rozdělení tabulky při jejím vytvoření v případě, že nechceme začít vkládat do defaultní tabulky s jedním regionem. Nehledě na to, zda pre-splitting použijeme nebo ne, je potřeba nadále spravovat regiony, v závislosti na nově přidávaných datech. K tomu slouží Auto splitting, který automaticky rozděluje regiony, které dosáhly své maximální velikosti. Nastavení automatického dělení regionů je možné provést pomocí `RegionSplitPolicy` API. K dispozici jsou tři předdefinované varianty:

- `ConstantSizeRegionSplitPolicy`
- `IncreasingToUpperBoundRegionSplitPolicy`
- `KeyPrefixRegionSplitPolicy`

ConstantSizeRegionSplitPolicy Jedná se o jednoduchou politiku (ve starších verzích před verzí 0.94 to byla defaultní a jediná předdefinovaná) k určení, kdy má dojít k rozdělení regionu. Rozdělení nastane v případě, kdy dojde k překročení maximální velikosti alespoň jednoho file store v regionu (uložiště jedné family v rámci regionu). Tato hodnota je konfigurována v `hbase.hregion.max.filesize` a její defaultní hodnota je 10 GB.

IncreasingToUpperBoundRegionSplitPolicy Pro novější verze je defaultní `IncreasingToUpperBoundRegionSplitPolicy`. Tato politika provádí více agresivní splitting založený na počtu regionů uložených v region serverech. K určení velikosti při které dojde k rozdělení regionu se používá vzorec $MIN(R^2 * hbase.hregion.memstore.flush.size, hbase.hregion.max.filesize)$, kde R je počet regionů stejné tabulky na stejném regionu. Například uvažujme situaci, kdy máme nastavenou default memstore flush size na defaultní hodnotu 128MB a hodnotu max store size na defaultních 10GB. K prvnímu splitu dojde po provedení první operace flush, při velikosti 128MB. Postupně, jak bude docházet ke zvyšování počtu regionů v region serveru, bude zároveň docházet i ke zvětšování velikosti jednotlivých regionů a to konkrétně na 512MB, 1152MB, 2GB, 3.2GB, 4.6GB, 6.2GB, etc. Po dosažení 9 regionů se zvětšování zastaví díky `hbase.hregion.max.filesize` nastavené na 10 GB. U obou předchozích metod dochází při dělení vždy k rozdělení v místě, kde je prostřední bod největšího store file.

KeyPrefixRegionSplitPolicy Poslední z předdefinovaných politik pro automatický splitting je `KeyPrefixRegionSplitPolicy`. V tomto případě se rozdělení provádí za pomoci prefixu. Na základě určení velikosti prefixu, dojde k jejich rozdělení do skupin. Tato funkce pak zajistí, že při dělení regionů nedojde k rozdělení uprostřed takto definované skupiny. Dojde tedy k situaci,

kdy jsou klíče se stejným prefixem uloženy v rámci jednoho regionu (nebo více regionů v případě, že velikost těchto dat překročí maximální velikost region serveru).

Při porovnání těchto předem definovaných politik jsem došel k závěru, že politika `KeyPrefixRegionSplitPolicy` přesně splňuje požadavky plynoucí z potřeby ukládat předem definované rozsahy řádku pomocí prefixu v rámci jednoho regionu. Je zde také možnost naimplementovat vlastní politiky, nicméně v tomto případě plně vyhovuje již definovaná politika.

3.1.3 optional kapitola, jak jsou splity implementovány

<http://hortonworks.com/blog/apache-hbase-region-splitting-and-merging/>

3.1.4 Fyzické uložení family column

Ačkoli jsou řádky ukládány lexikograficky, v každém stores jsou takto uloženy pouze hodnoty v rámci jedné family. Každá family sloupců se tedy ukládá do oddělených stores. Platí však omezení, že ačkoli jsou jednotlivé části řádku uloženy rozděleně, vždy jsou tyto části uloženy v rámci jednoho regionu. Dojde-li tedy k přejmenování identifikátoru řádku, repsektive vytvoření nové tabulky s prefixem, budou řádky o určitém rozsahu klíčů vždy uloženy logicky v jednom HRegionu.

3.2 PATOH algoritmus

Při kombinační části je zapotřebí rozdělit co nejoptimálněji vzniklý hypergraf. Tato fáze je klíčová a právě správné rozdělení grafu je klíčem k následnému optimálnímu rozložení dat. K vyřešení této úlohy použijí již naimplementovaný nástroj pro dělení hypergrafů PaToH⁶. Program nabízí 2 možnosti použití. Buď jako stand-alone program spustitelný z příkazové řádky nebo jako knihovna v jazyku C. Vzhledem k tomu, že pro naše potřeby plně dostačuje rozhraní spustitelného programu, využijí této možnosti a v implementaci provedu výpočet spuštěním příkazu z kódu.

3.2.1 Použití programu

3.2.1.1 Vstupní formát dat

Vstupní hypergraf je reprezentován textovým souborem. První řádek souboru obsahuje informace o celém grafu. Následující řádky pak obsahují informace o každé hyperhraně v grafu. První řádek obsahuje 4 až 6 hodnot. Hodnoty postupně udávají:

⁶<http://bmi.osu.edu/~umit/software.html>

```

1 8 9 28
8 6 3 5 2
4 5 1 7
4 2 5 7
4 7
3 5
8 2 4
6 5 2
5 7 2
8 4

```

Obrázek 3.2: Vstupní formát hypergrafu

- index base (0 nebo 1) značící od jakého čísla indexujeme elementy v grafu
- počet uzlů v grafu
- počet hyperhran v grafu
- počet pinů (propojení hrany a vrcholu)
- schéma vážení grafu (volitelné)
- počet vah pro uzel (volitelné)

Následující řádky reprezentují vždy po jedné hyperhraně a udávají indexy uzlů, které hyperhrana propojuje. Možné podoby vstupních souborů jsou ukázány na obrázku 4.1

3.2.1.2 Výstupní formát dat

Po vykonání výpočtu na hypergrafu reprezentovaném souborem input.hygr pomocí PaToH programu, je výsledek uložen v souboru input.hygr.part.K. Kde K značí počet vygenerovaných podgrafů. Tento soubor obsahuje $|V|$ čísel v rozsahu $[0, K - 1]$. Každé i -té číslo ukazuje do jakého hypergrafu patří i -tý vrchol z původního grafu.

```

> patoh ken-11.u 4 RA=6 UM=U

+++++
+++ PaToH v3.2 (c) Nov 1999-, by Umit V. Catalyurek
+++ Build Date: Sun Mar 13 17:41:19 2011 -0400
+++++
*****

```

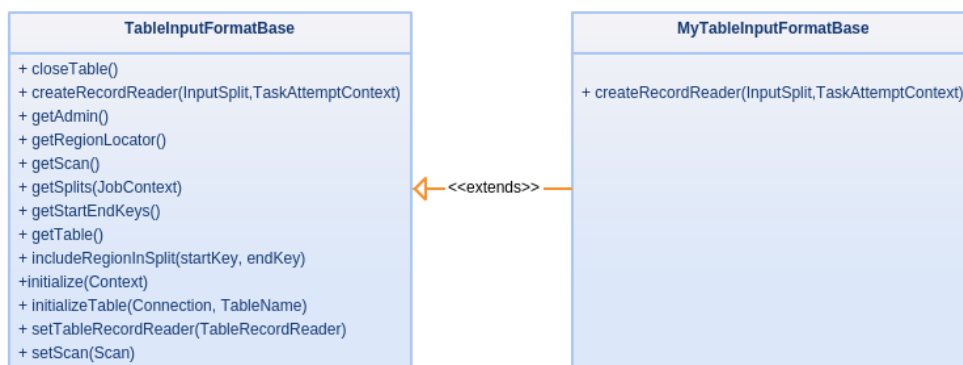
```
Hypergraph:ken-11.u #Cells:14694 #Nets:14694 #Pins:82454
*****
4-way partitioning results of PaToH:
Cut Cost: 4775
Part Weights : Min= 20408 (0.010) Max= 20947 (0.016)
-----
I/O : 0.007 sec
I.Perm/Cons.H: 0.002 sec ( 0.7%)
Coarsening : 0.087 sec (37.9%)
Partitioning : 0.011 sec ( 4.8%)
Uncoarsening : 0.128 sec (55.8%)
Total : 0.229 sec
Total (w I/O): 0.236 sec
-----
```


Implementace řešení

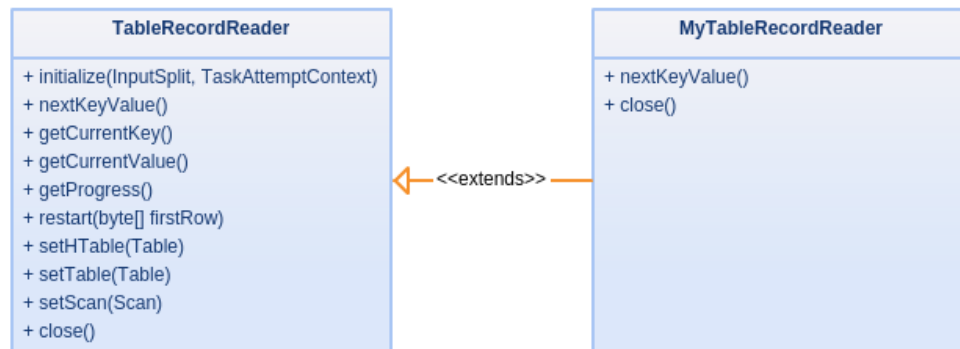
V této kapitole bude představena implementace navrženého řešení. Při popisu implementace budu postupně probírat všechny klíčové partie. Implementace byla provedená v jazyku Java. V javě je napsán celý balík Hadoop a tak se toto řešení nabízí jako nejvhodnější. Pouze pro provedení dělicího algoritmu byl využit již existující program naimplementovaný v jazyku C.

4.1 Monitoring

V prvním kroce je nutné získat potřebná data, podle kterých se následná optimalizace bude provádět. Je zapotřebí shromáždit informace o tom, který řádek tabulky generuje výstupy do reduceru a i informaci do kterého. Sběr těchto dat se realizuje v průběhu vykonávání sady MapReduce úloh, pro které se bude následně celý dataset optimalizovat. K tomu abychom mohli nasbírat potřebná data je nutné rozšířit HBase třídy `RecordReader` třídou `MyRecordReader` a `TableInputFormatBase` třídou `MyTableInputFormatBase`. Funkcionality těchto tříd je následující:



Obrázek 4.1: Rozšíření třídy `TableInputFormatBase`



Obrázek 4.2: Rozšíření třídy RecordReader

TableInputFormatBase

Nejdříve ověří vstupní konfiguraci pro MapReduce úlohu. Následně vytvoří instanci třídy **RecordReader**, která bude použita k vytvoření key/value párů z načtených dat. Tyto páry jsou pak po jednom zaslány k mapování.

TableRecordReader

TableRecordReader generuje z dat získaných od **TableInputFormatBase** key/value páry. Podle zadaného rozsahu rozsahu projde požadovanou část datasetu, nebo celou tabulku.

4.1.1 MyTableInputClassBase Class

V této třídě je pouze nutné změnit metodu **CreateRecordReader** tak, aby namísto defaultní TableRecordReader instance nastavila pro prováděnou úlohu instanci record readeru ze třídy **MyTableRecordReader**.

4.1.2 MyTableRecordReader Class

V této třídě je nutné vytvořit soubor, který uchová vygenerované páry key/value, během zpracovávání MapReduce. Při zpracovávání každého záznamu se vyhodnotí, zda byl vygenerován klíč. v případě že ano, uloží se do dočasné struktury daný klíč a identifikátor řádku, který ho vygeneroval. Při ukončování úlohy se poté zapíšou získané hodnoty do nově vytvořeného souboru ve formátu: *Key; počet_řádků_generujících_klíč; id_řadku₁...id_řadku_n*. výsledkem je tedy seznam vygenerovaných klíčů s řádky které tento klíč generují.

4.2 Repartitioning

Poté co jsme shromáždili potřebná metadata shromážděná při vykonávání MapReduce úloh, je čas přejít k fázi optimalizace rozložení dat. Nyní máme k

dispozici soubor nebo několik souborů, kde jsou definovány jednotlivé řádky databáze a k nim seznam klíčů, které generují. Za pomoci dat z těchto souborů je zapotřebí vytvořit strukturu hypergrafu, na který pak následně použijeme partitioning algoritmus.

4.2.1 Vytvoření a reprezentace hypergrafu

Pro reprezentaci hypergrafu jsem použil knihovnu *Jung*⁷ poskytující implementace různých typů grafů mezi nimiž je i námi požadovaný hypargraf. Při parsování vstupních souborů se postupně vytváří hypergraf. Každý řádek reprezentuje hyperhranu a vrcholy které jsou touto hranou propojeny. Vždy se tedy vytvoří nová hyperhrana, nesoucí jako svůj identifikátor jméno generovaného klíče, který reprezentuje a postupně ověřuje existenci uzlů, které by měla propojovat. V případě že uzel neexistuje, vytvoří ho a přidá jako jeden ze svých vrcholů. V opačném případě si pouze přidá již existující uzel. Po dokončení parsování souborů s metadatami tedy máme k dispozici strukturu *setHypergraph* reprezentující kompletní rozložení zátěže datasetu při vykonávání zadaných dotazů.

4.2.2 příprava dat pro Min-cut algoritmus

Samotné rozdělení grafu pomocí min-cut algoritmus je prováděno za pomoci externího programu PaToH, nicméně, tento program požaduje specifický formát vstupního souboru. Tento soubor je tedy zapotřebí vygenerovat ze stávající reprezentace grafu. Jak již bylo popsáno v dřívější kapitole. PaToH očekává jako vstupní soubor graf v podobě od 0 nebo 1 indexovaný seznam hran a vrcholů jež propojuje. Vytvořil jsem tedy dvě pole *nodesArray* a *edgesArray*, do kterých se postupně uloží všechny vrcholy a hrany. Indexy těchto polí pak slouží k vygenerování požadovaného formátu pro PaToH program. Do vytvářeného souboru nejprve uložím hlavičku souboru ve tvaru 0 znamenající výchozí hodnotu indexování a následně počet vrcholů, hran a pinů v grafu. Tyto hodnoty lze jednoduše získat z reprezentace hypergrafu pomocí metod *getVertices()*, *getEdges()* a počet pinů pomocí metody pro každý uzel *degree(node)* a sečtené přes všechny uzly v grafu.

4.2.3 Spuštění PaToH a zpracování výsledků

4.2.4 Vytvoření a konfigurace nové tabulky

4.2.4.1 Nastavení prozdělení tabulky na regiony

4.2.5 přesun upravených dat

⁷<http://jung.sourceforge.net/>

Testování a vyhodnocení měření

5.0.6 Vyhodnocení

5.0.7 Testování

5.0.7.1 JUnit testy

5.0.7.2 MRUnit testy

Závěr

Literatura

- [1] deRoos, D.: *Hadoop For Dummies*. O'Reilly Media, Inc., 2014, ISBN 978-1-118-60755-8.
- [2] White, T.: *Hadoop: The Definitive Guide*. O'Reilly Media, Inc., třetí vydání, 2012, ISBN 0596521979, 9780596521974.
- [3] Cloudera trainings. 2015. Dostupné z: <http://cloudera.com/content/cloudera/en/training/library.html>
- [4] HDFS User Guide. 2014. Dostupné z: <http://hadoop.apache.org/docs/r2.3.0/hadoop-project-dist/hadoop-hdfs/HdfsUserGuide.html>
- [5] George, L.: *Hadoop: The Definitive Guide*. O'Reilly Media, Inc., první vydání, 2011, ISBN 978-1-4493-9610-7.
- [6] Team, A. H.: *Apache HBase Reference Guide*. 2015. Dostupné z: <http://hbase.apache.org/book.html>
- [7] George, L.: *HBase: The Definitive Guide*. O'Reilly Media, první vydání, 2011, ISBN 1449396100.
- [8] HBaseCon. 2015. Dostupné z: <http://hbasecon.com/archive.html>
- [9] Schweikert, D. G.; Kernighan, B. W.: A Proper Model for the Partitioning of Electrical Circuits. In *Proceedings of the 9th Design Automation Workshop*, DAC '72, New York, NY, USA: ACM, 1972, s. 57–62, doi:10.1145/800153.804930. Dostupné z: <http://doi.acm.org/10.1145/800153.804930>
- [10] Fiduccia, C. M.; Mattheyses, R. M.: A Linear-time Heuristic for Improving Network Partitions. In *Proceedings of the 19th Design Automation Conference*, DAC '82, Piscataway, NJ, USA: IEEE Press, 1982, ISBN 0-89791-020-6, s. 175–181. Dostupné z: <http://dl.acm.org/citation.cfm?id=800263.809204>

- [11] et al, M. L.-G.: *Data Partitioning for Minimizing Transferred Data in MapReduce*. INRIA & LIRMM, Montpellier, France.
- [12] Ibrahim, S.; Hai, J.; Lu, L.; aj.: Locality/fairness-aware key partitioning for mapreduce in the cloud. *CloudCom 2010*, 2010: s. 17–24.
- [13] Enis, S.: Apache HBase Region Splitting and Merging. 2013. Dostupné z: <http://hortonworks.com/blog/apache-hbase-region-splitting-and-merging/>

Generátor zkušebních dat

Testovací MapReduce úlohy

Instalace Hadoop prostředí

Seznam použitých zkratek

HDFS Hadoop Distributed File System

REST Representational State Transfer

CRUD Create Read Update Delete

API Application Programming Interface

SQL Structured Query Language

NoSQL Not only SQL

RDBMS Relational DataBase Management System

WAL Write-Ahead Log

Obsah přiloženého CD

	readme.txt.....	stručný popis obsahu CD
	exe	adresář se spustitelnou formou implementace
	src	
	impl.....	zdrojové kódy implementace
	thesis	zdrojová forma práce ve formátu \LaTeX
	text	text práce
	thesis.pdf	text práce ve formátu PDF
	thesis.ps	text práce ve formátu PS