VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
FACULTY OF INFORMATION TECHNOLOGY

ÚSTAV INFORMAČNÍCH SYSTÉMŮ
DEPARTMENT OF INFORMATION SYSTEMS

DISTRIBUOVANÝ REPOSITÁŘ DIGITÁLNÍCH FORENZNÍCH DAT

DISTRIBUTED FORENSIC DIGITAL DATA REPOSITORY

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE — **Bc. MARTIN JOSEFÍK**
AUTHOR

VEDOUCÍ PRÁCE — **RNDr. MAREK RYCHLÝ, Ph.D.**
SUPERVISOR

BRNO 2018

Abstrakt

Do tohoto odstavce bude zapsán výtah (abstrakt) práce v českém (slovenském) jazyce.

Abstract

Do tohoto odstavce bude zapsán výtah (abstrakt) práce v anglickém jazyce.

Klíčová slova

Sem budou zapsána jednotlivá klíčová slova v českém (slovenském) jazyce, oddělená čárkami.

Keywords

Sem budou zapsána jednotlivá klíčová slova v anglickém jazyce, oddělená čárkami.

Citace

JOSEFÍK, Martin. *Distribuovaný repositář digitálních forenzních dat*. Brno, 2018. Diplomová práce. Vysoké učení technické v Brně, Fakulta informačních technologií. Vedoucí práce RNDr. Marek Rychlý, Ph.D.

Distribuovaný repositář digitálních forenzních dat

Prohlášení

Prohlašuji, že jsem tuto bakalářskou práci vypracoval samostatně pod vedením pana X... Další informace mi poskytli... Uvedl jsem všechny literární prameny a publikace, ze kterých jsem čerpal.

.....
Martin Josefík
29. října 2017

Poděkování

V této sekci je možno uvést poděkování vedoucímu práce a těm, kteří poskytli odbornou pomoc (externí zadavatel, konzultant, apod.).

Obsah

Kapitola 1

Úvod

Kapitola 2

Forenzní analýza digitálních dat

Forenzní analýza digitálních dat je věda identifikující, zachovávající, obnovující, analyzující a předávající fakta ohledně digitálních důkazů nalezených v počítačích nebo digitálních úložištích mediálních zařízení. Nezabývá se tedy pouze počítači, ale také ostatními digitálními technologiemi včetně mobilních telefonů a tabletů, mobilních sítí, internetového bankovnictví, datových médií apod. Pod výše uvedenými aktivitami se skrývá [8]:

- Identifikace - Jedná se o první část celého procesu. Předtím, než je cokoliv zkoumáno a analyzováno, je důležité identifikovat, kde jsou data uložena. Typicky jsou uložena na diskových jednotkách, serverech, flash klíčenkách, síťových zařízeních.
- Zachování - Důležitá je ochrana důkazů, tzn. pro sběr a analýzu informací je potřeba zachovat původní data, musí se zabránit jejich změně a ztrátě. Bez integrity je důkazní materiál nepoužitelný.
- Obnovení - Součástí procesu je i obnova dat, která může zahrnovat obnovu smazaných dat procesy operačního systému, úmyslně smazané soubory, soubory chráněné heslem a také poškozené soubory.
- Analýza - Jedná se o hlavní část vyšetřování. Cílem je shromáždit co nejvíce relevantních artefaktů.
- Předání - Po analýze jsou artefakty důkladně zdokumentovány a odevzdány například ve formě protokolu.

Vyšetřování digitálních forenzních dat obvykle zahrnuje vytvoření a prozkoumání obrazu disku. Obraz disku je kopie celého disku nebo jeho části bit po bitu. Obraz je statický snímek, který může být analyzován za účelem odhalení nebo stanovení událostí ohledně incidentů a být tak použitý jako důkaz v soudní síni. Analýza je prováděna na kopii pro zachování integrity originálu.

Vyšetřovatel zanalyzuje obraz pomocí snímacích technik, aby získal relevantní data z disku. Forenzní obraz obsahuje soubory z disku, ale také nealokovaný prostor a tzv. `slack space`. Slack space je pozůstatek diskového prostoru, který byl alokován pro nějaký počítačový soubor a ten všechen prostor nepotřebuje. Právě v těchto prostorech mohou být nalezeny relevantní informace jako například smazané soubory, či jejich fragmenty. [10]

2.1 Formáty digitálních forenzních dat

Typů digitálních forenzních dat existuje spousta. Každý typ takových dat může být reprezentován jiným formátem. Tato sekce čerpá informace převážně z [3].

Mnoho forenzních počítačových programů používají své vlastní formáty pro uložení informace. Můžeme je rozdělit na nezávislé (anglicky `Independent File Formats`) a programově-specifické formáty (anglicky `Program-Specific File Formats`).

- Independent File Formats - Tyto formáty byly vyvinuty nezávisle na konkrétním forenzním programu. Patří mezi ně `AFF`, `AFF4`, `gfzip`, `Raw Image Format`.
- Program-Specific File Formats - Byly vyvinuty pro použití specifickými forenzními programy. Většinou každý takový formát je unikátní a proto je pro přečtení potřeba unikátního nástroje. Zástupci jsou například `Encase image file format`, `ProDiscover image file format`, `IXimager file formats`.

Vyznamným zástupcem je systém AFF4, který bude popsán detailněji.

2.2 AFF4

Celým názvem `The Advanced Forensics File Format 4` se jedná o open source formát pro ukládání digitálních důkazů a dat. Jeho výhodami jsou správa metadat a možnost komprese. Tato sekce čerpá převážně z[1]. Je založen na objektově orientované architektuře. Veškerá množina známých objektů je označována jako `AFF4 universe`. Takový prostor je definovaný jako nekonečný, protože AFF4 je navržen pro škálování obrovského množství důkazního materiálu. Všechny objekty jsou adresovatelné jejich jménem, které je v rámci AFF4 universe unikátní.

Příkladem jména nějakého AFF4 objektu může být:
`urn:aff4:f3eba626-505a-4730-8216-1987853bc4d2`
Jedná se o standardní URN notaci, URN je unikátní.

AFF4 universe používá RDF notaci pro specifikaci atributů objektů. V nejjednodušší podobě je RDF množina tvrzení o objektu ve formátu:
`Subject Attribute Value`

Příklad:

```
******** Object urn:aff4:f3eba626-505a-4730-8216-1987853bc4d2 ***********
aff4:stored = urn:aff4:4bdbf8bc-d8a5-40cb-9af0-fd7e4d0e2c9e
aff4:type = image
aff4:interface = stream
aff4:timestamp = 0x49E9DEC3
aff4:chunk_size = 32k
aff4:compression = 8
aff4:chunks_in_segment = 2048
aff4:size = 10485760
```

Příklad ukazuje, že objekt má tyto atributy a hodnoty. Nazýváme je relace nebo fakta. Celý AFF4 universe je sestavený z takových faktů.

AFF4 objekty existují, protože dělají něco užitečného, což závisí na rozhraní, které představují. Aktuálně existuje několik rozhraní, nejvýznamnější jsou `Volume` a `Stream`. Rozhraní objektu je fakt o objektu, který nalezneme v atributu `aff4:interface`.

2.2.1 Rozhraní Volume

Rozhraní Volume definujeme jako mechanismus ukládání, který dokáže uložit segment (bit binárních dat) pod nějaké jméno a získat jej podle tohoto jména. Aktuálně existují dvě implementace: `Directory` a `ZipFile`.

- `Directory Volume` - Tato implementace ukládá segmenty jako soubory uvnitř běžného adresáře v souborovém systému. Hodí se zejména, pokud potřebujeme uložit obraz na souborový systém FAT, přičemž velikost segmentu je malá a nenarazíme tak na omezení velikosti souboru. Je také možné založit adresář na nějaké http adrese, což nám umožní používat obraz přímo z webu.
- `ZipFile Volume` - Jak napovídá název, tato implementace ukládá segmenty uvnitř zip archivu. Malé soubory lze tak bez problémů otevřít obyčejným průzkumníkem (windows explorer) a data extrahovat. Zase je možné zapsat zip archiv přímo na HTTP server a používat obraz přímo ze serveru.

Je možnost převádět mezi oběma formáty z jednoho na druhý, extrahovat zip archiv do adresáře a vytvoření Directory volume.

2.2.2 Rozhraní Stream

Streamy jsou základním rozhraním pro ukládání dat obrazu. Stream obsahuje metody typu `read`, `seek`, `tell` a `close`. Podporuje ještě `write`, ale ne k modifikaci obrazu, nýbrž k jeho vytvoření. Pokud nějaký AFF4 objektu podporuje rozhraní stream, lze provést náhodné čtení jeho dat. Existuje několik specifických implementací rozhraní stream, některými z nich jsou:

- `FileBackedObjects` - Stream, který ukládá data v souboru v souborovém systému, jehož pozice je určena URN souboru.
- `HTTPObject` - Pozice souboru je udána pomocí URL. Objekt lze ukládat a číst z HTTP serveru. Implementace umožňuje přečíst určité rozmezí bajtů, režie síťového provozu mezi klientem a serverem je minimální. Je možné vyšetřit vzdálený obraz přes http bez potřeby celé kopie obrazu. Server musí být samozřejmě zabezpečen, ale to už AFF neřeší.
- `Segments` - Segmenty jsou komponenty uloženy přímo ve `Volume`. Volume je zjednodušeně řečeno objekt uchovávající segmenty. Segmenty by měly být použity pro malé streamy, protože prohledávat v komprimovaných segmentech může být drahá operace. Segmenty jsou užitečné, pokud potřebujeme vytvořit logický obraz nějaké podmnožiny souborového systému (pouze některé soubory) a ne forenzní obraz celkového systému.
- `Image streams` - Tyto streamy jsou opakem segmentů. Pro velké obrazy nemůžeme použít segmenty, protože by nebyly zkomprimovány efektivně. Image stream ukládá obraz v tzv. `chunks`.

Kapitola 3

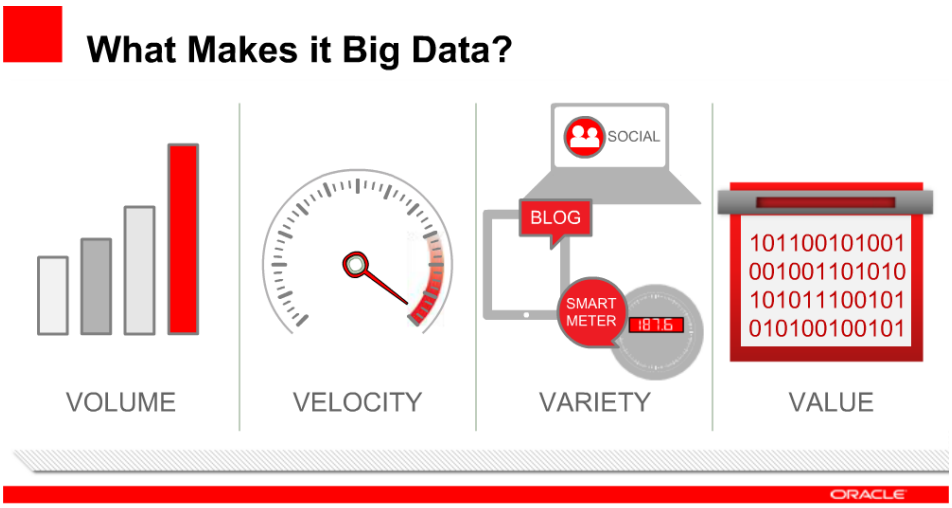
Úložiště pro rozsáhlá strukturovaná i nestrukturovaná data

V této kapitole budou vysvětleny termíny Big data, distribuované databáze a NoSQL databáze, včetně jejich vlastností, výhod a nevýhod.

3.1 Big data

Definicí pro frázi Big data existuje několik. Jedná se o termín použitý na soubory dat, které jsou příliš komplexní z hlediska velikosti a různorodosti, a které je nemožné zpracovávat běžně používanými přístupy a softwarovými nástroji v rozumném čase.

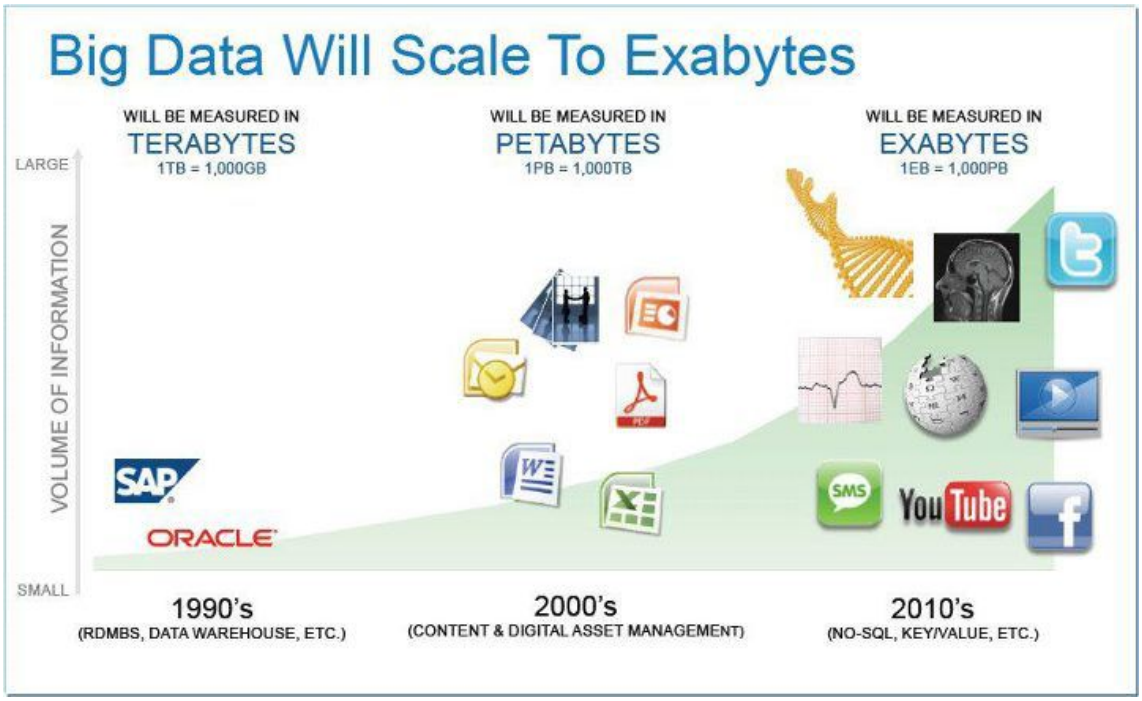
Objem takových dat rychle roste. Vyskytují se v mnoha odvětvích, například sběr informací o počasí, sociální sítě, energetické a telekomunikační společnosti, ekonomie a finančnictví, či data z kamer, měření z různých senzorů apod. Z toho plyne, že se jedná o data různorodých typů, mohou být strukturovaná i nestrukturovaná. Proto je potřeba existence různých technologií pro jejich uložení, zpracování i zobrazení.



Obrázek 3.1: Definice Big data podle Oracle. [5]

Big data je často definováno jako 4V z anglických slov Volume, Velocity, Variety a Value. [4]

- Volume – značí množství nebo velikost dat. Big data vyžaduje zpracování vysokých objemů dat neznámých hodnot, například síťový provoz, data sesbírána ze senzorů apod.
- Velocity – vyjadřuje rychlost z hlediska vzniku dat a potřeby jejich analýzy, některá vyžadují zpracování v reálném čase. Nejdůležitější data se zapisují přímo do paměti, a ne na disk, z důvodu co nejrychlejšího zpracování.
- Variety – znamená různorodost typů. Jedná se především o nestrukturovaná data, například text, audio, video, data o geografické poloze a další. Jsou na ně kladeny velmi podobné požadavky jako na data strukturovaná – sumarizace, monitorování, důvěrnost. [4]
- Value – data mají vlastní hodnotu, která musí být analyzována a zjištěna. Nejedná se o jednoduchý proces, je stále potřeba nových metod a technik zpracování.



Obrázek 3.2: S novými technologiemi se masivně zvyšuje růst dat a přibývají nové typy. [6]

Tato práce se zabývá Big daty hlavně typu – PCAP soubory, logy ze síťových zařízení a komunikací. Možnosti uložení Big data budou popsány v následujících podkapitolách.

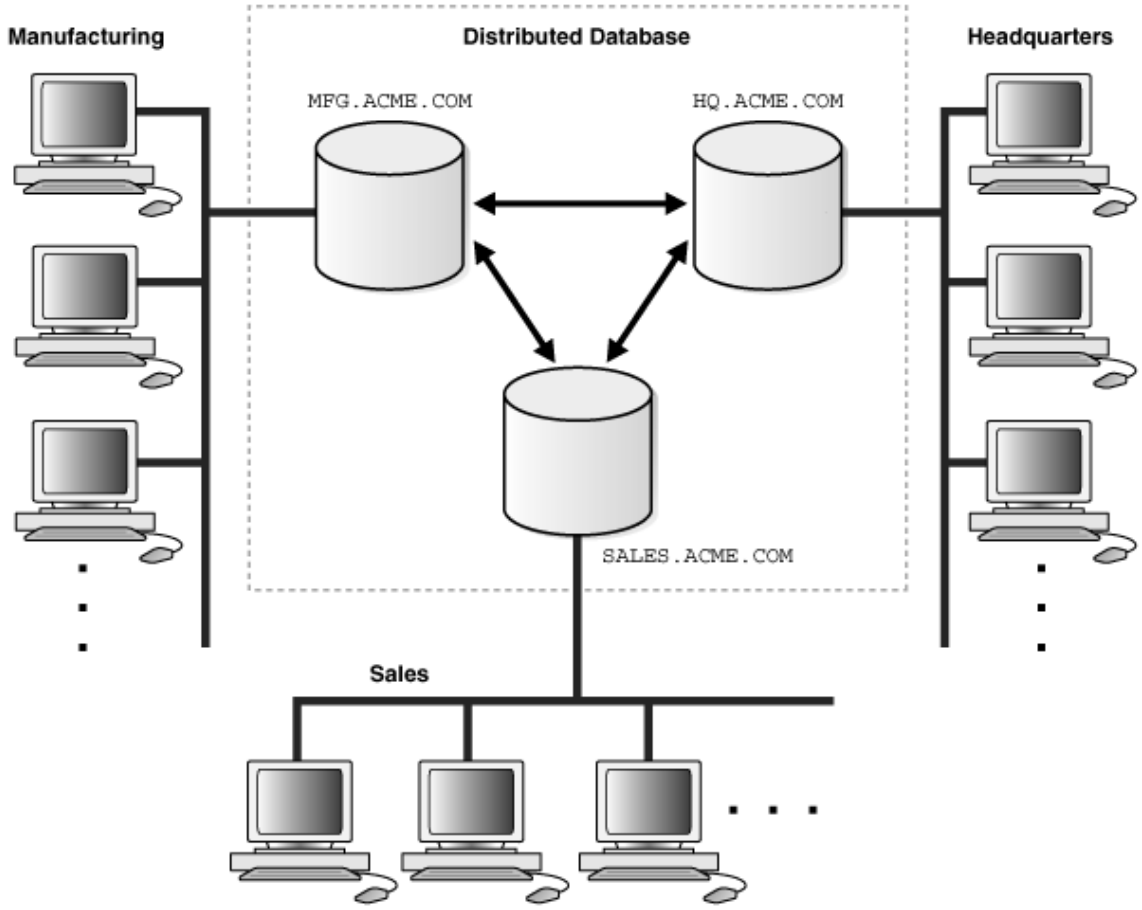
3.2 Distribuované databáze

Distribuovaná databáze se skládá z většího počtu samostatných databází, které mohou být geograficky rozmístěny na jiných pozicích. Jednotlivé uzly spolu komunikují přes počítačovou síť. Každý uzel je sám o sobě databázový systém. DSŘBD neboli systém řízení

distribuované báze dat (anglicky Distributed Database Database Management System) zajišťuje, že se distribuovaná databáze uživatelům jeví jako jedna jediná databáze. Data jsou fyzicky uložena na různých pozicích. Mohou být spravována rozdílnými SŘBD nezávisle na ostatních pozicích. [9]

Systém řízení distribuované báze dat je centralizovaný systém s těmito vlastnostmi [9]:

- Umí vytvářet, získávat, upravovat a mazat distribuované databáze. Zajišťuje důvěrnost a integritu databází.
- Periodicky synchronizuje databázi a poskytuje mechanismy přístupu tak, aby se databáze uživatelům jevila transparentní.
- Zajišťuje, že změna dat v kterémkoliv uzlu se promítne i v ostatních uzlech.
- Je využíván v aplikacích, kde se předpokládá zpracování velkých objemů dat, ke kterým přistupuje současně mnoho uživatelů.



Obrázek 3.3: Schéma distribuované databáze a současný přístup více zařízení k ní. [7]

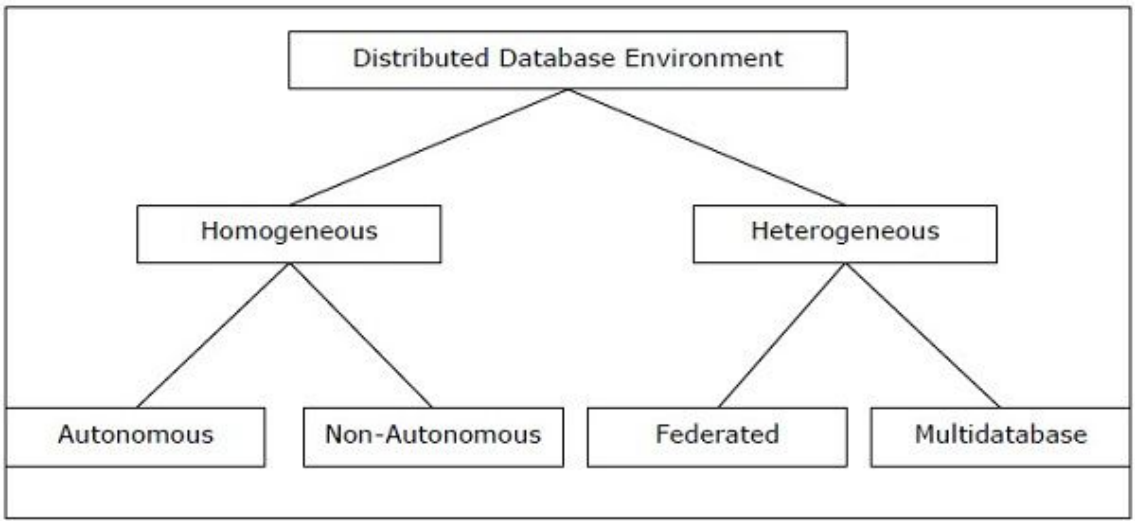
Výhody

- Rozšiřitelnost – pokud je potřeba databázový systém rozšířit do nových míst nebo přidat další uzly, stačí přidat nový(é) počítač(e) a lokální data v nové pozici, a nakonec je připojit k distribuovanému systému, bez jakéhokoliv přerušení funkcionality. Podobný postup je při odebrání uzlu.

- Spolehlivost – když nějaký z připojených uzlů selže, nepřestane distribuovaná databáze fungovat, sníží se maximálně výkon.
- Ochrana (záloha) dat – při zničení jednoho uzlu a smazání dat z něj, mohou být stejná data zálohována i na jiných uzlech.
- Výkonnost – pokud jsou data efektivně distribuována, může být uživatelův požadavek uspokojen rychleji. Transakce mohou být také distribuované a provedeny rychleji.

Nevýhody

- Integrita dat – data musí být průběžně synchronizována na více uzlech, aby na stejné dotazy nebyly z různých uzlů vraceny rozdílné odpovědi.
- Komunikační režie – i zdánlivě jednoduchá operace může vyžadovat spoustu zbytečné komunikace.
- Cena – DSŘDB vyžaduje drahý a složitý software ke koordinaci uzlu a zajištění transparentnosti. [9]
- Mezi další patří – složitost, zabezpečení, řízení souběžného přístupu k datům.



Obrázek 3.4: Distribuované databáze můžeme rozdělit na homogenní a heterogenní, a tyto ještě dále dělit. [9]

Homogenní – všechny uzly používají identické SŘBD a operační systémy. Uzly mají informace o ostatních uzlech a spolupracují při zpracování uživatelských požadavků. Homogenní distribuovaná databáze se navenek jeví uživateli jako jeden systém. Je jednodušší jej navrhnout a spravovat.

Heterogenní – uzly mohou mít rozdílné operační systémy a SŘBD, které nejsou kompatibilní. Mohou také využívat rozdílná schémata (relační, objektově orientované, hierarchické, ...). Rozdílnost schématu je hlavním problémem při zpracování dotazu a transakcí. Kvůli tomu je také složité dotazování. [2]

Architekturami distribuovaných databází jsou centrální architektura, klient-server, peer-to-peer, multi-databázová architektura.

3.3 NoSQL, disky, úložiště

Kapitola 4

Návrh distribuovaného úložiště

4.1 Přístup k datům

4.1.1 Sekvenční, náhodný

4.1.2 Dotazování

4.1.3 Big data přístupy

4.2 Architektura

4.3 Aplikační rozhraní

4.4 Technologie

4.4.1 Docker

4.4.2 HDFS, Hadoop, Spark

4.4.3 Cassandra / MongoDB

4.4.4 Zookeeper

4.4.5 MQ broker

Kapitola 5

Implementace

5.1 Rozšiřitelnost, znuvupoužitelnost

Kapitola 6

Testování

6.1 Výkon

Kapitola 7

Závěr

Literatura

[1] *Advanced Forensic Framework 4 (AFF4)*. [Online; navštíveno 28.10.2017]. URL http://forensicswiki.org/wiki/AFF4

[2] *Distributed database*. [Online; navštíveno 02.10.2017]. URL https://en.wikipedia.org/wiki/Distributed_database

[3] *Forensics File Formats*. [Online; navštíveno 28.10.2017]. URL http://www.forensicswiki.org/wiki/Category:Forensics_File_Formats

[4] Heller, P.; Piziak, D.; Stackowiak, R.; aj.: *An Enterprise Architect's Guide to Big Data*. [Online; navštíveno 26.09.2017]. URL http://www.oracle.com/technetwork/topics/entarch/articles/oea-big-data-guide-1522052.pdf

[5] Louwers, J.: *Big Data is sometimes Fast Data*. [Online; navštíveno 27.09.2017]. URL http://johanlouwers.blogspot.cz/2013/01/big-data-is-sometimes-fast-data.html

[6] Nambiar, R.: *What is Big Data ?* . [Online; navštíveno 27.09.2017]. URL http://rrnamb.blogspot.cz/2012/09/what-is-big-data.html

[7] Oracle Help Center: *Distributed Database Architecture*. [Online; navštíveno 29.09.2017]. URL https://docs.oracle.com/cd/B28359_01/server.111/b28310/ds_concepts001.htm

[8] Stephens, B.: *What Is Digital Forensics?* [Online; navštíveno 28.10.2017]. URL https://www.interworks.com/blog/bstephens/2016/02/05/what-digital-forensics

[9] Tutorials Point (I) Pvt. Ltd.: *Distributed DBMS Tutorial*. [Online; navštíveno 29.09.2017]. URL https://www.tutorialspoint.com/distributed_dbms/

[10] Vandeven, S.: *Forensic Images: For Your Viewing Pleasure*. [Online; navštíveno 28.10.2017]. URL https://www.sans.org/reading-room/whitepapers/forensics/forensic-images-viewing-pleasure-35447