Úvod

Digitální forenzní data

Formáty digitálních forenzních dat

Způsob uložení

Existující systémy

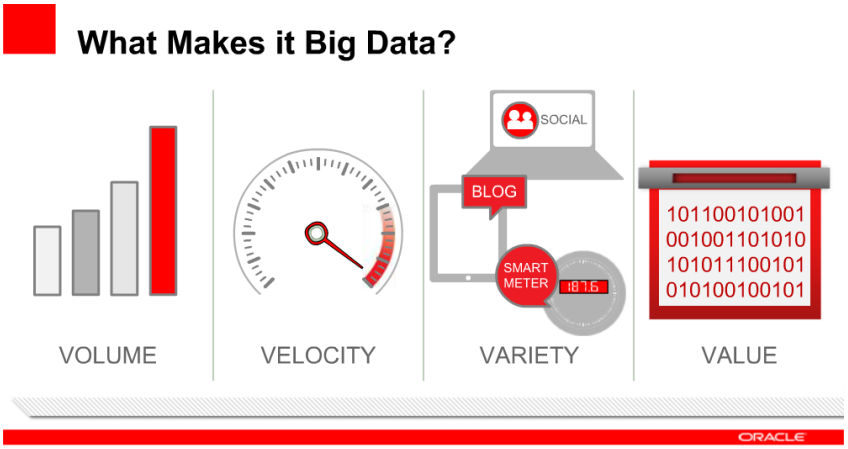
Úložiště pro rozsáhlá strukturovaná i nestrukturovaná data

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

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.



[3]

Big data je často definováno jako 4V z anglických slov Volume, Velocity, Variety a Value. [2]

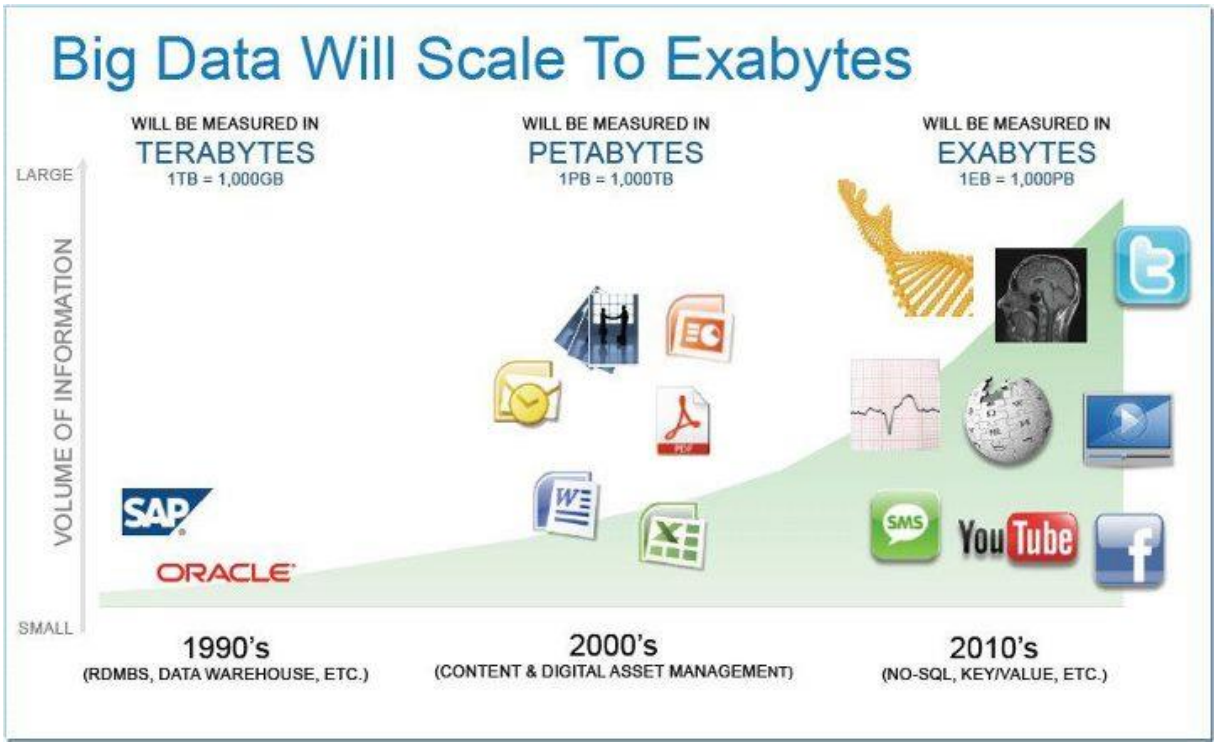
Volume – značí množství nebo velikost dat. Big data vyžaduje zpracování vysokých objemů dat neznámých hodnot, například síťový provoz, data sesbírána ze senzorů apod.

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. [2]

Value – data mají vlastní hodnotu, která musí být analyzována a zjištěna. Nejedná se o jednoduchý proces, je stále potřeba nových metod a technik zpracování.



S novými technologiemi se masivně zvyšuje růst dat a přibývají nové typy. [4]

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.

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. [6]

Systém řízení distribuované báze dat je centralizovaný systém s těmito vlastnostmi [6]:

- 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.

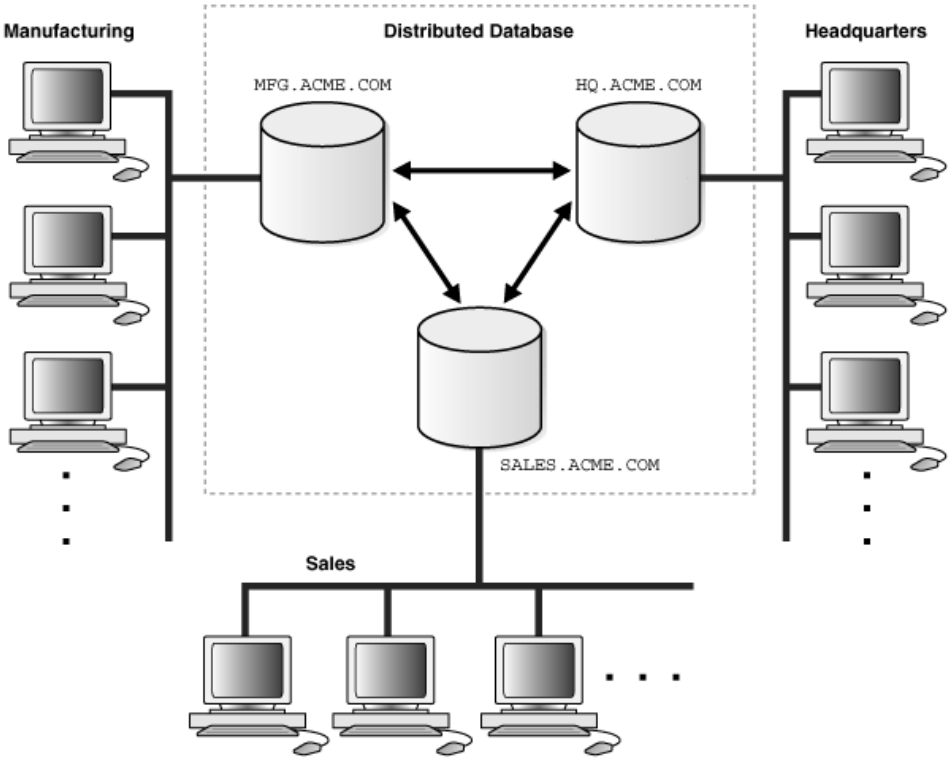


Schéma distribuované databáze a současný přístup více zařízení k ní. [7]

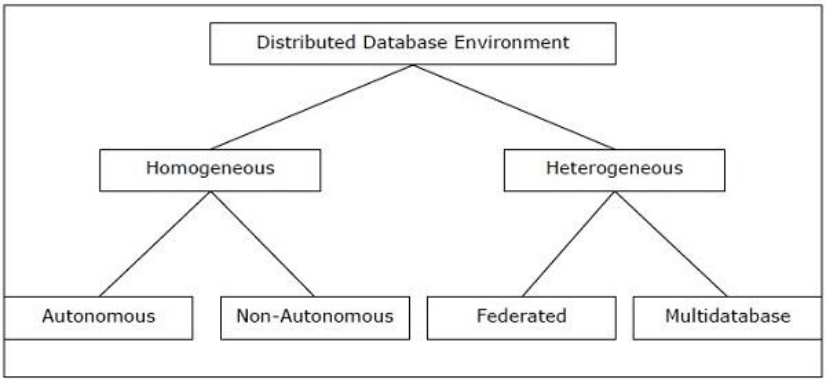
Výhody

- 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. [6]
- Mezi další patří – složitost, zabezpečení, řízení souběžného přístupu k datům.

Distribuované databáze můžeme rozdělit na homogenní a heterogenní, a tyto ještě dále dělit, jak ukazuje obrázek:



Převzato z [6].

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. [8]

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

NoSQL, disky, úložiště

Návrh distribuovaného úložiště

Přístup k datům

Sekvenční, náhodný

Dotazování

Big data přístupy

Architektura

Aplikační rozhraní

Technologie

Implementace

Rozšiřitelnost, znuvupoužitelnost

Testování

Výkon

Závěr

Reference

[1] https://www.systemonline.cz/clanky/big-data.htm

[2] http://www.oracle.com/technetwork/topics/entarch/articles/oea-big-data-guide-1522052.pdf

[3] http://mattiasdrefs.com/wp-content/uploads/2015/12/what_makes_data_big_data.png

[4] http://rrnamb.blogspot.cz/2012/09/what-is-big-data.html

[5] https://www.slideshare.net/OracleMKTPR20/oracle-big-data-y-database-analytics-andres-araujo

[6] https://www.tutorialspoint.com/distributed_dbms/

[7] https://docs.oracle.com/cd/B28359_01/server.111/b28310/ds_concepts001.htm

[8] https://en.wikipedia.org/wiki/Distributed_database