cesnet

Od

mongoDB®

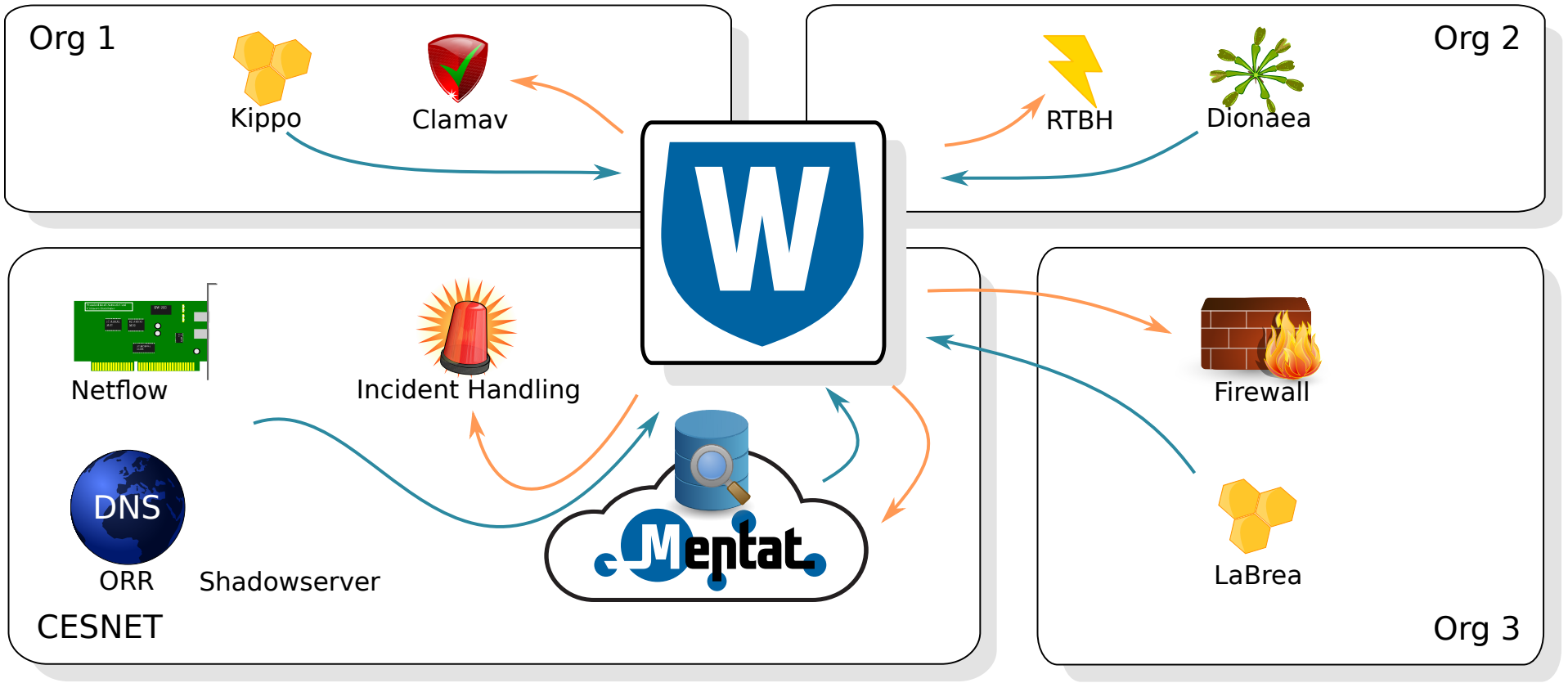
k

PostgreSQL

Pavel Kácha
ph@cesnet.cz

Jan Mach, Radko Krkoš, Radomír Orkáč, Tomáš Plesník

SABU – Sdílení a analýza bezpečnostních událostí



- Warden – obousměrný komunikační kanál https://warden.cesnet.cz
  - Co pošleš, dostanou všichni ostatní (a naopak)

- Mentat – jeden z klientů Wardenu https://mentat.cesnet.cz
  - Databáze, historie, reportování, analýza, „SIEM“
  - Lokální nástroj, který posouvá dosah událostí do konvenční reality (email, web)

Formát – IDEA

Botnet C&C


```
{
  "Format": "IDEA0",
  "ID": "cca3325c-a989-4f8c-998f-5b0e971f6ef0",
  "DetectTime": "2014-03-05T15:52:22Z",
  "Category": ["Intrusion.Botnet"],
  "Description": "Botnet Command and Control",
  "Source": [
    {
      "Type": ["Botnet", "CC"],
      "IP4": ["93.184.216.119"],
      "Proto": ["tcp", "ircu"],
      "Port": [6667]
    }
  ]
}
```


Honeypot


```
{
  "Format": "IDEA0",
  "ID": "2E4A3926-B1B9-41E3-89AE-B6B474EB0A54",
  "DetectTime": "2014-03-22T10:12:31Z",
  "Category": ["Recon.Scanning"],
  "ConnCount": 633,
  "Description": "EPMAPPER exploitation attempt",
  "Ref": ["cve:CVE-2003-0605"],
  "Source": [
    {
      "IP4": ["93.184.216.119"],
      "Proto": ["tcp", "epmap"],
      "Port": [24508]
    }
  ],
  "Target": [
    {
     "Port": [135]
    }
  ]
}
```


- Jednoduchý, rozšiřitelný JSON
- Více zdrojů (DDoS), více cílů (sken)
- https://idea.cesnet.cz

Mentat – hlavní use-case

- Umožnit uživatelům vyhledávat podle důležitých polí.
  (zdroje, cíle, porty, služby, kategorie, tagy, detektory, časy, …)
- Málo složitých analytických dotazů.
  (nelze čekat paralelizaci na úrovni uživatelů)
- Rozměry
  - Průměrná velikost dokumentu 2 kB
  - Cca 1 milion zpráv denně
  - Cca měsíc dat
    (Tj. cca 30 mil záznamů, 50G raw dat)
  - Cca 12 zápisů za sekundu
  - Prakticky žádné updaty.

Tohle nejsou Big Data!

MongoDB

- Umí přece libovolně strukturovaná JSON data.
- Má bezva agregační framework.
- Je to nativní C kód (jsme staromilci)
- A používá je velké G, velké T i velké FB.

- Použili jsme
  - v začátcích 2.6.x s MMAPv1 backendem,
  - v závěru 3.2.x s WiredTiger backendem.

- Disclaimer:

  Takže za cca poslední rok a půl může být v Mongu leccos jinak.

Konverze, indexy

- IP - nejsou nativní IP typy. Neva, IP je jen omalovaný integer.
  - IPv4 - 32 b, ale IPv6 - 128 b (tj. nevejde se do longu)
  - Ale - binární typy jsou v Mongu uspořádatelné, takže binData.
- Rozsahy IP
  - CIDR? (prefix) Není obecný, neumožňuje jednoduché indexování.
  - Řešení: IP.min, IP.max

- Časté dotazy jsou na průnik zdrojů a cílů. Ale:
  - Nelze dvě pole ve složeném indexu (dává smysl, byl by to kartézský součin).
- Navíc ale potřebujeme porovnávat související min a max v poli
  - V Mongu $elemMatch, v PQ "&& ANY(...)", tj. neindexovatelné.

Algoritmická odbočka: Range search

- Časté dotazy jsou na rozsahy. Dotaz na více atributů lze urychlit složeným indexem – pokud se ale jedná vždy o konkrétní jednu hodnotu atributu.
- Jedná-li se o dva rozsahy přes dva sloupce (např. Zdroj a Cíl)
  - Máme-li složený index:

    A 1
    A 2
    A 3
    B 2
    B 3
    C 3

- Hledání „Zdroj: <A, C> & Cíl: 2“ znamená projít v indexu všechny hodnoty od A do C a v každé z nich teprve konzultovat druhý sloupec indexu.
- Algoritmický problém – hledání v ortogonálních rozsazích
  - Řešitelný pomocí K-D tree, K-D-B tree, Z-order curve (Mortonova křivka).
  - V běžných databázích ale pouze kombinacemi indexů a chytrým plánovačem.

Výkon

- Máme 200GB RAM, vejde se celý working set včetně indexů.

  Přesto:
  - nedeterministická doba odpovědi
  - "nevykonatelné" dotazy (desítky minut, po 45 to Mongo vzdá)
  - vybydlená jádra

- Po víceméně neúspěšném zkoumání jsme alespoň
  - Omezili možnosti dotazování a zkrátili defaultní délku dotazů na týden.

    (Mongo pak občas použilo časový index, ostatní podmínky muselo dohledat bruteforce, ale zase nemuselo řadit.)
  - Omezili a optimalizovali na stroji běhy čehokoliv jiného, co by mohlo interferovat (na procesoru, paměti nebo disku).
  - Víceméně okometricky poladili konfiguraci.

    (Dokumentace tristní, občas se dozvíte, kam se máte koukat, ale ne co s tím udělat.)
- Mírné zlepšení, ale minutové dotazy nebyly výjimkou, patologie stále desetiminutové.

Výkon – různé

- Mongo by mělo od 2.6 umět průnik indexů.
  - (což by dávalo smysl pro filtrování podle různých polí za rozumného počtu indexů)
  - Potíž: v životě jsem ho v explainu neviděl.
- Skip+Limit
  - Vypadá to, že algoritmy Monga příliš neberou v úvahu top-K.
  - Skip+limit na poslední stránku někdy delší, než vyčtení kompletně všech dat.
  - Opačné řazení na indexu má cca 10% dopad na výkon.
- Občas vyloženě špatné plány, nutnost ručního hintování.
  - Planner se neřídí daty, jen "query shape" - dotaz, řazení, projekce.
- Konkurence
  - MMAPv1: "the number of cores can improve performance but does not provide significant return"
  - WiredTiger: "multithreaded and can take advantage of additional CPU cores"
    - Po migraci ale žádný znatelný rozdíl.
- Všelék (dle vývojářů) replikace, sharding.
  - Zkusili jsme různé kombinace, spíše degradace výkonu.
  - Ale - kvůli 50G dat postavíme cluster?

Údržba

- Příprava číselníků – distinct
  - Rychlý…
  - … pokud nepoužijete podmínku.

  Error in getting detector selector data from db. recv timed out (800000 ms)

- Pomohl inkrementální výpočet v aplikaci,
- a později ve 3.2 partial index (předindexujeme jen to, co je v podmínce).

- Vacuum?
  - Naplánujte výpadek a připravte si jednou tolik volného místa.
  - Anebo všelék - postavte repliku, přemigrujte, zrušte…

Umělé limity

- BSON size limit 16MB.

- Takže co je větší, musí do GridFS…
  - … ale co když nad tím stále potřebujete vyhledávat?
  - Kreativně lámat a skládat výjimky?

- Totéž platí i pro výsledky z aggregation frameworku.

Exception: aggregation result exceeds maximum document size (16MB)

PostgreSQL

- Zkusili jsme různá rozlámání (tj. alespoň 1NF), ale JOINy byly drahé.
- Zkusili jsme podporu JSONu – některé dotazy ale nedokážeme vyjádřit.

- Nejlépe se osvědčilo vytáhnout potřebná data do vlastních sloupců ploché tabulky a využít složené typy PostgreSQL, tj. např. pole rozsahů v jednom atributu.

- Trvá tím ale dřívější problém $elemMatch, tady "&& ANY(...)", tj. nutnost procházet pole, ale:
  - PostgreSQL má rychlé sekvenční skeny (rozhodně proti Mongu).
  - A pokud jsou v dotazu další podmínky, má velmi dobrý plánovač.

- Není bez problémů:
  - NULL terminated strings: JSON s "\0000" nelze uložit.
  - Ale s vytaženými daty JSON nepotřebujeme, takže BYTEA.

Paralelizace

- 9.6 - sekvenční skeny paralelizované
- 10 - index skeny paralelizované
  (Na našich datech škáluje prakticky lineárně.)

- effective_io_concurrency = 8 (default 1)
  - Počet současných IO requestů - máme NVME.
- parallel_setup_cost = 10.0 (default 1000)
  - Použij paralelizaci prakticky kde můžeš.
- default_statistics_target = 1000 (default 100)
  - Dělej si přesnější statistiku pro plánovač.

```
CREATE TABLE IF NOT EXISTS events(
    id text PRIMARY KEY,
    detecttime timestamp NOT NULL,
    category text[] NOT NULL,
    description text,
    source_ip iprange[],
    target_ip iprange[],
    source_port integer[],
    target_port integer[],
    source_type text[],
    target_type text[],
    protocol text[],
    node_name text[] NOT NULL,
    node_type text[],
    cesnet_storagetime timestamp NOT NULL,
    cesnet_resolvedabuses text[],
    cesnet_eventclass text,
    cesnet_eventseverity text,
    cesnet_inspectionerrors text[],
    event bytea
);
```

- Těšíme se na 11
  - Předávání LIMIT paralelním workerům (mohlo by zlepšit dotazy s nízkou selektivitou).
  - Zvýšení výkonu sekvenčního paralelního skenu.
  - Optimalizace monotónně doplňovaných BTREE indexů (časy, id).
  - Zlepšení statistiky pro neuniformní rozložení hodnot.

Závěr

- 119 milionů událostí (tj. už 2 měsíce): 161GB DB + 24GB indexů
  (Průměrná velikost dokumentu se nyní zvedla cca na 4 kB.)
- Patologie 20 s, žádné "nevykonatelné" dotazy, většina dotazů v jednotkách sekund.
- Kromě hlavního usecase máme k dispozici celou mašinerii PostgreSQL pro různé ad-hoc složité analytické dotazy.

commit 016883a78de0dd20777117fcd0804017f5035fb1
Author: Jan Mach
Date: Thu Aug 2 11:35:38 2018 +0200

Removed all mentions of MongoDB from documentation, added deprecation warnings to code libraries.

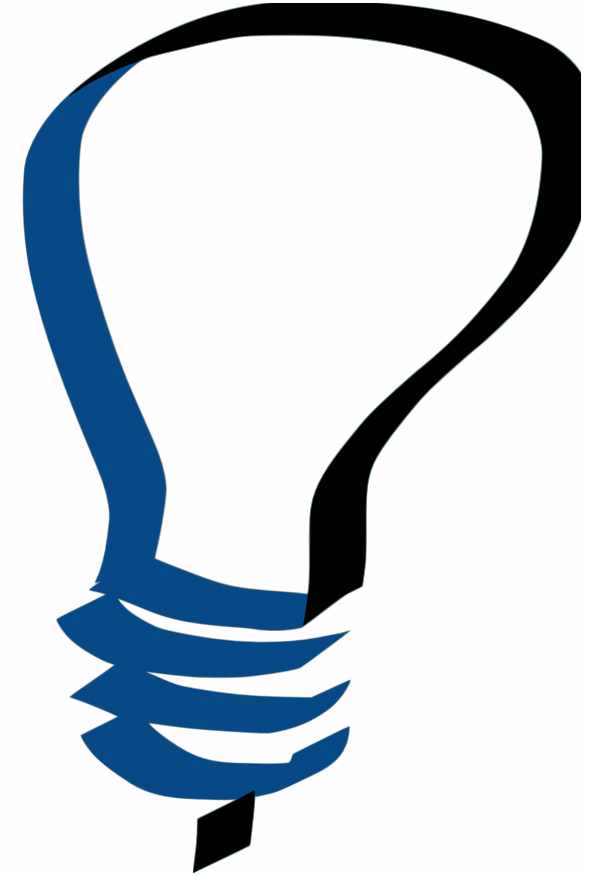
The only places in documentation where MongoDB mentions were kept are for obvious reasons installation and migration pages. (Redmine issue: #4225)

Děkuji za pozornost.

Hledáme nové pracovníky na pozice:

- správce a vývojář internetových služeb
- analytik digitální forenzní laboratoře

Více informací na stánku sdružení CESNET.



The Catch by TREND MICRO & cesnet