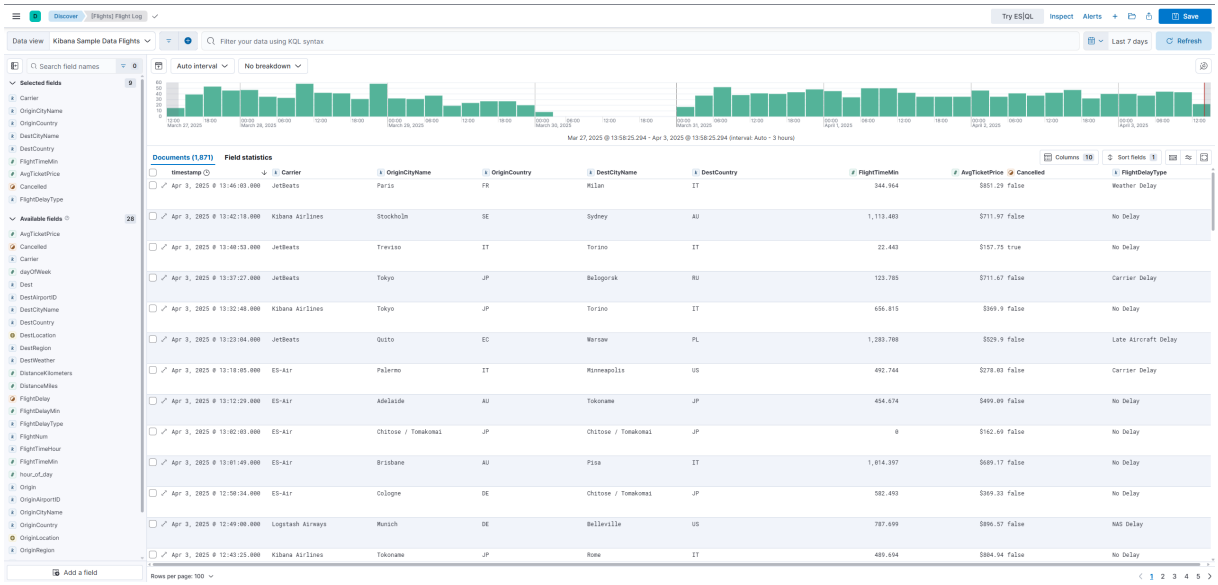
Ukážka používania Elasticsearch

Na demonštráciu práce s Elasticsearch a Kibana použijeme vstavaný dataset Sample flight data, ktorý simuluje reálne údaje o letoch. Dataset obsahuje polia ako dátum letu, krajinu pôvodu a cieľovú destináciu, odletové a príletové časy, meškanie a ďalšie podrobnosti. Na tomto príklade si ukážeme, ako možno tieto dáta analyzovať pomocou nástrojov Elasticsearch a Kibana – vrátane filtrovania, vizualizácie a tvorby dashboardov.

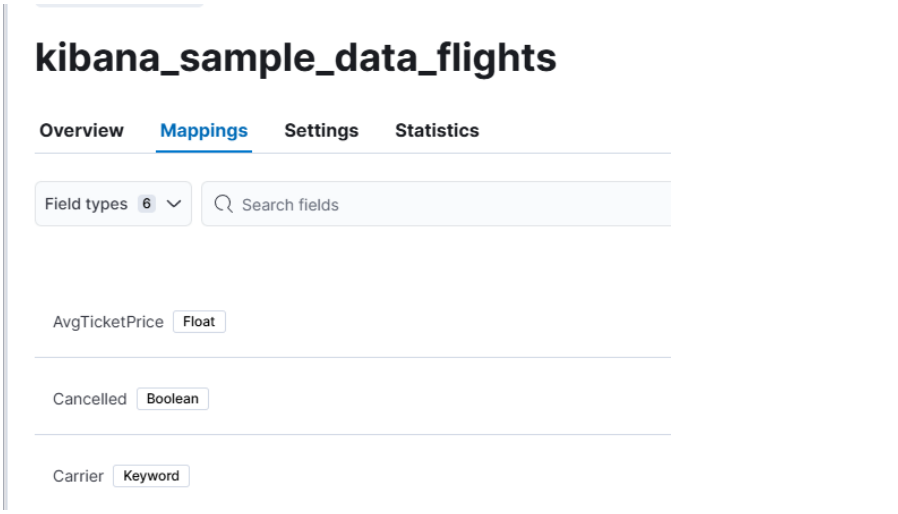
V sekcii Discover v Kibane môžeme interaktívne prehliadať všetky dokumenty uložené v Elasticsearch. Na pravej strane sa zobrazujú samotné dokumenty (záznamy o letoch), na ľavej strane vidíme zoznam dostupných polí. Každé pole môžeme použiť na filtrovanie alebo tvorbu vizualizácií.



Na tomto obrázku je vidieť použitie filtrovania – v hornej časti zadávame podmienky na výber záznamov. Napríklad môžeme zobraziť všetky lety, ktoré odlietajú z Abu Dhabi, zadaním filtra OriginCityName: "Abu Dhabi". Tento spôsob filtrovania je veľmi praktický pri analýze väčšieho množstva údajov.

ndexy a typy polí

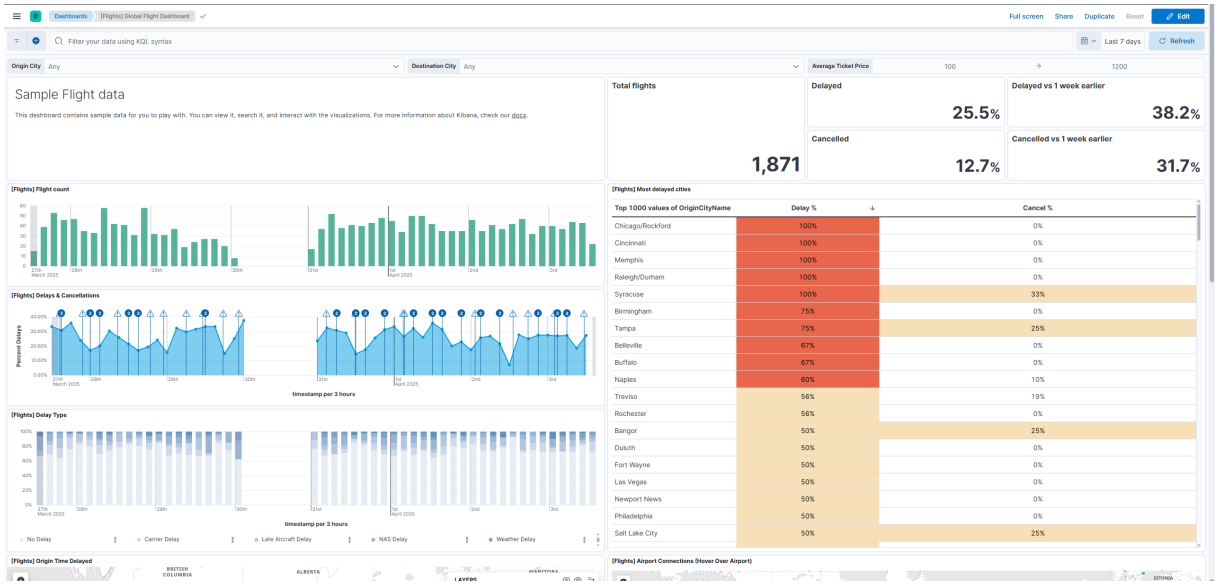
Elasticsearch ukladá dáta do indexov, ktoré predstavujú základnú štruktúru pre ukladanie a vyhľadávanie údajov. Každý index obsahuje dokumenty vo formáte JSON, ktoré sú podobné riadkom v databázovej tabuľke. Každý dokument má polia s rôznymi typmi, napríklad text, keyword, integer, date, geo_point alebo boolean.



Pre efektívne a rýchle vyhľadávanie je dôležité správne navrhnúť tzv. mapping, ktorý definuje dátové typy a spôsob spracovania jednotlivých polí. Napríklad pole OriginAirportID, ktoré obsahuje kódy letísk, by malo byť typu keyword, aby sa dosiahlo presné porovnávanie. Naopak, pole s popisom alebo menom mesta je vhodné typu text pre potreby fulltextového vyhľadávania.

Pri textových poliach sa používa tzv. analyzátor, ktorý rozkladá text na jednotlivé tokeny. Napríklad veta „Flight delayed by weather“ sa rozloží na slová: flight, delayed, by, weather. To umožňuje efektívne fulltextové vyhľadávanie. Avšak na presné hodnoty (napr. letiskové kódy) nie je vhodné analyzovanie, pretože by mohlo viesť k chybným výsledkom.

V časti Dashboards môžeme kombinovať rôzne vizualizácie na jednom mieste. Vytváranie grafov nevyžaduje programovanie – stačí si vybrať typ vizualizácie (napr. bar chart, pie chart, map, data table), určiť zdroj dát a agregácie (napr. počet letov podľa krajiny, priemerné meškanie podľa dopravcu). Tieto vizualizácie môžeme interaktívne filtrovať – napríklad kliknutím na konkrétnu krajinu na mape sa automaticky aktualizujú všetky ostatné vizualizácie.



Elasticsearch umožňuje pracovať s aliasmi – alternatívnymi názvami pre indexy. Alias môže smerovať na jeden alebo viac indexov a pomáha pri prechode na nový index bez potreby zmeny dotazov v aplikáciách. To je užitočné napríklad pri rotácii indexov podľa dátumu (napr. mesačné indexy s aliasom flights_current).

Templates zas definujú vopred pripravené štruktúry indexov. Ak vytvárame indexy automaticky (napr. podľa dátumu), môžeme šablónou určiť mapping, nastavenia shardov, repliky a aliasy. Tým sa zabezpečí jednotnosť a optimalizácia bez nutnosti manuálneho zásahu pri každom novom indexe.

V časti Index Management môžeme sledovať všetky vytvorené indexy. Vidíme ich názvy, počet dokumentov, veľkosť na disku, stav shardov a tzv. health status:

- Green znamená, že všetky primárne aj replikačné shardy sú dostupné,
- Yellow znamená, že chýba niektorá replika (napr. keď máme len jeden node),
- Red značí problém – niektoré primárne shardy sú nedostupné.



Okrem stavu vieme sledovať aj:

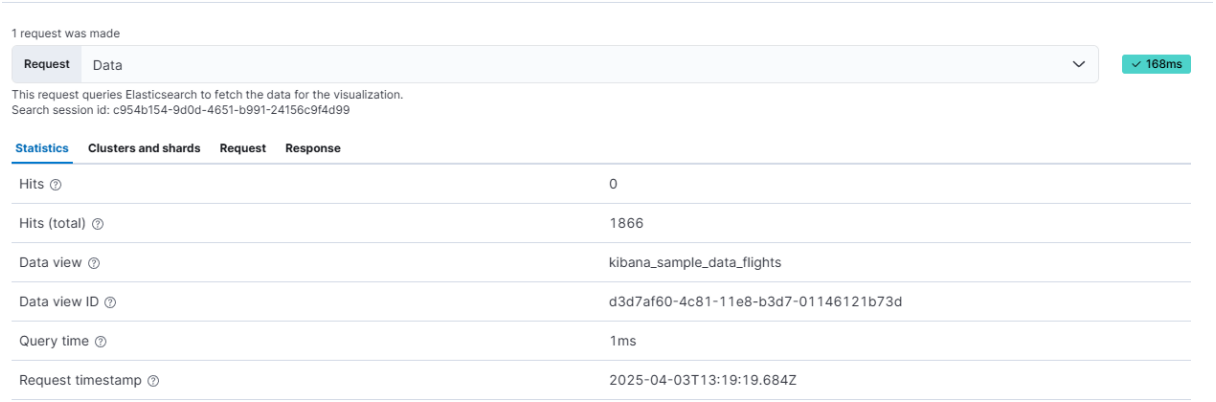
- veľkosť každého indexu,
- počet dokumentov,
- počet shardov (delení) a replík (kópií pre dostupnosť).

Tieto informácie sú dôležité pri plánovaní výkonu, odhade zaťaženia a škálovaní. Každý index sa pri vytvorení rozdelí na primárne shardy, ktoré obsahujú samotné dáta. K nim môžeme nastaviť repliky – záložné kópie shardov, ktoré slúžia na zvýšenie dostupnosti a zároveň umožňujú paralelné spracovanie dotazov. Napríklad index s 3 primárnymi shardmi a 1 replikou bude mať spolu 6 shardov (3 originály a 3 kópie).

Správna konfigurácia závisí od:

- očakávaného objemu dát (veľmi malé indexy môžu mať len 1 shard),
- počtu node-ov v clustri (viac replík má zmysel len pri viacerých serveroch),
- potreby rýchlosti vs. redundancie.

Elasticsearch umožňuje pomocou Profile API zistiť, ako presne bol dotaz vykonaný – ktoré fázy trvali najdlhšie a aké operácie boli vykonané v pozadí (napr. filtering, scoring, aggregácie). V Kibane to vieme zobraziť cez Inspect panel (dostupný pri vizualizáciách alebo v sekcii Discover).



Získame tu:

- čas vykonania dotazu (v milisekundách),
- podrobnosti o jednotlivých fázach spracovania,
- použité filtre a ich náročnosť,
- výstupné JSON dotazy, ktoré môžeme ďalej upravovať a testovať.

Tento nástroj je výborný pre:

- ladenie zložitých dotazov s viacerými podmienkami,
- hľadanie tzv. bottleneckov – napríklad drahých aggregácií,
- validáciu, či správne využívame indexy a typy polí.