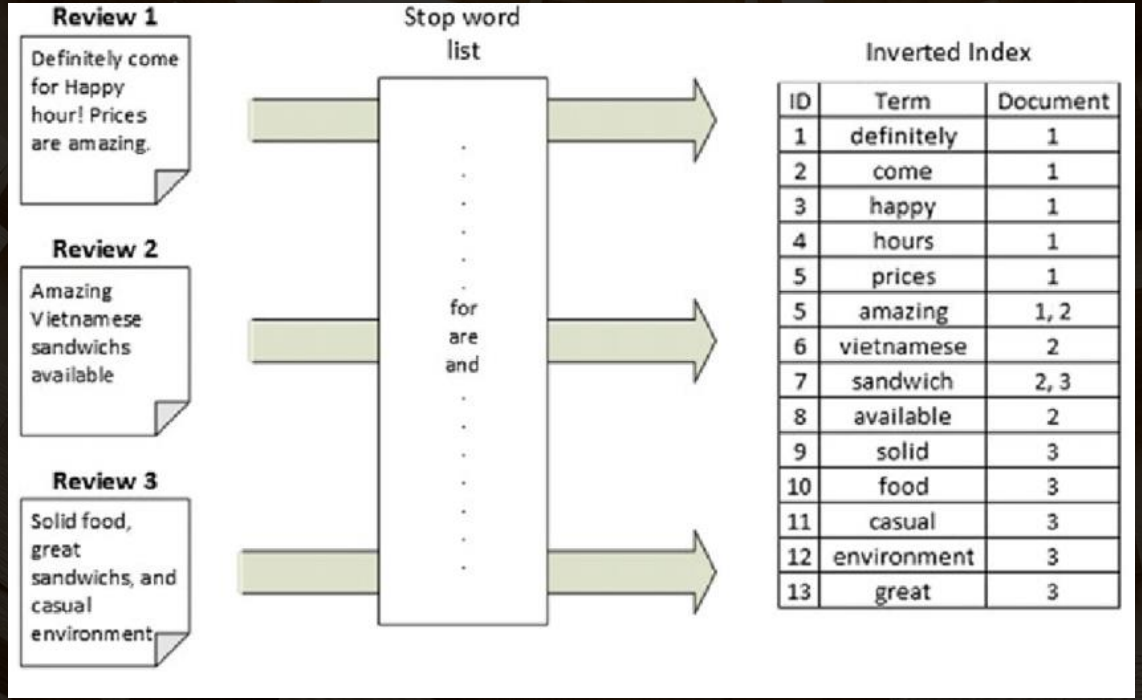
Full-Text Search e applicazioni: Confronto tra Microsoft SQL Server ed Elasticsearch

Tesi di Corradi Davide
Relatore Prof. Sonia Bergamaschi – Correlatore Luca Gagliardelli
Università degli Studi di Modena e Reggio Emilia
Dipartimento di Ingegneria «Enzo Ferrari»
Corso di Laurea Triennale in Ingegneria Informatica
Anno Accademico 2019/2020

Full-Text Search

- *Scopo:* *1. Trovare documenti pertinenti alle domande*
  2. Valutazione del matching
- *Full-Text database*
- *Indicizzazione e Ricerca*
- *Query ricche, flessibili e sofisticate unite ad algoritmi di ricerca specifici*
- *Inverted index*



ID	Term	Document
1	definitely	1
2	come	1
3	happy	1
4	hours	1
5	prices	1
5	amazing	1, 2
6	vietnamese	2
7	sandwich	2, 3
8	available	2
9	solid	3
10	food	3
11	casual	3
12	environment	3
13	great	3

Microsoft SQL Server

- *Permette la ricerca Full-Text, con installazione della componente opzionale specifica.*

- *Indicizzazione:* - *Una o più colonne*
  - *Catalogo full-text*
  - *Popolazione*

- *Per scrivere query full-text sono a disposizione:*

  - *Predicati: CONTAINS e FREETEXT*
  - *Funzioni: CONTAINSTABLE e FREETEXTTABLE*

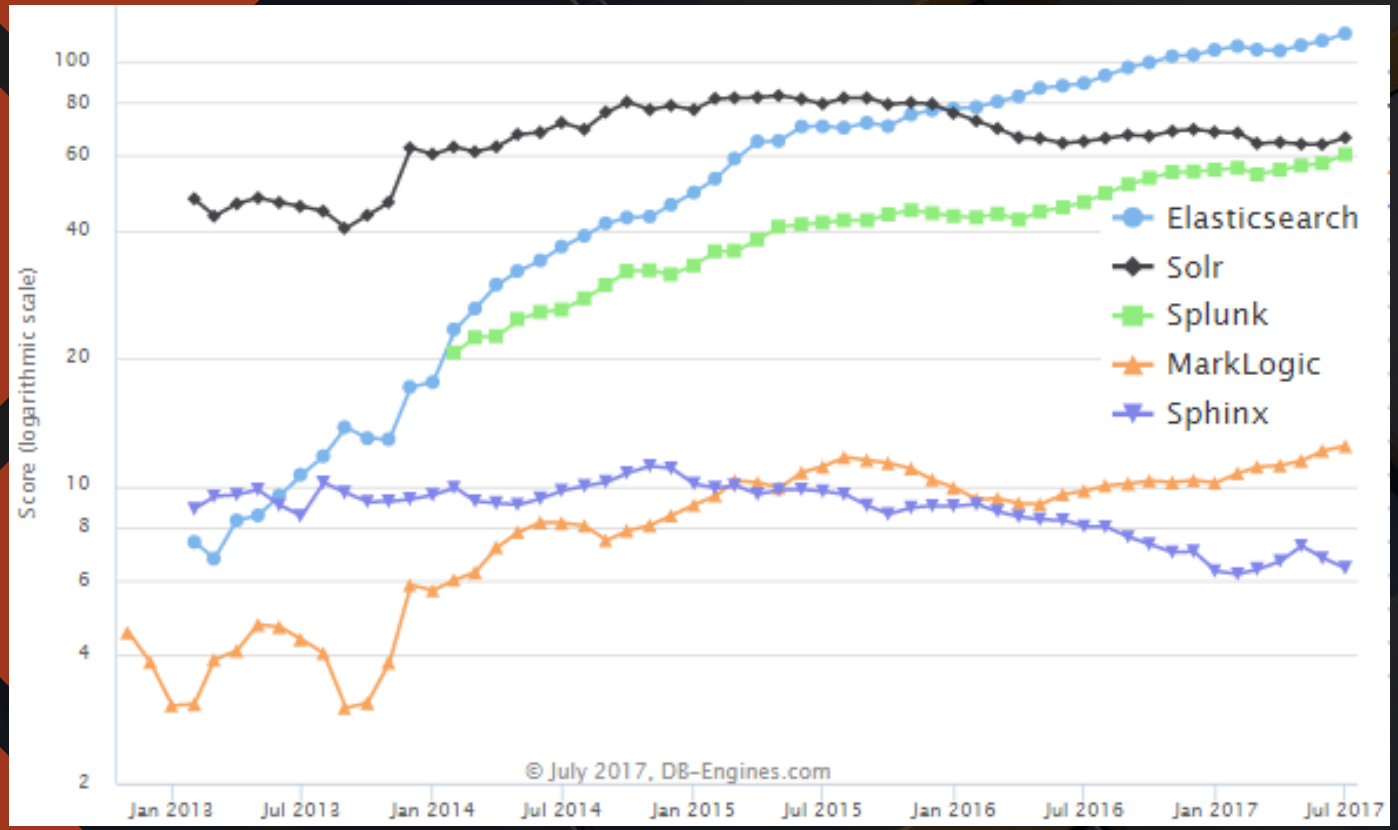
Problematiche

- *Sono nate nuove esigenze col tempo*
- *Ricerca di grandi moli di dati in tempi vicini al real-time.*
- *MS SQL Server offre una buona ricerca ma limitata dalle join, che lo rendono inefficiente*
- *Molte tabelle o molto grandi inficiano sulle operazioni di join*
- *Nascono strumenti differenti per il supporto della ricerca full-text*

Apache Lucene

- *Prodotto ideato da Apache Software Foundation*
- *Leader del mercato*
- *Open source*
- *Java*
- *Libreria con funzionalità di ricerca e indicizzazione*
- *Disposizione di API*

Elasticsearch



- *Creato nel 2004 da Shay Banon e proprietario di Elastic.*
- *Motore di ricerca NoSQL, basato su JSON.*
- *Elastic Stack: ELK (Elasticsearch, Logstash, Kibana)*
- *Altamente scalabile e resiliente*
- *Basato sulla tecnologia di Apache Lucene per l'archiviazione di dati e ricerca.*

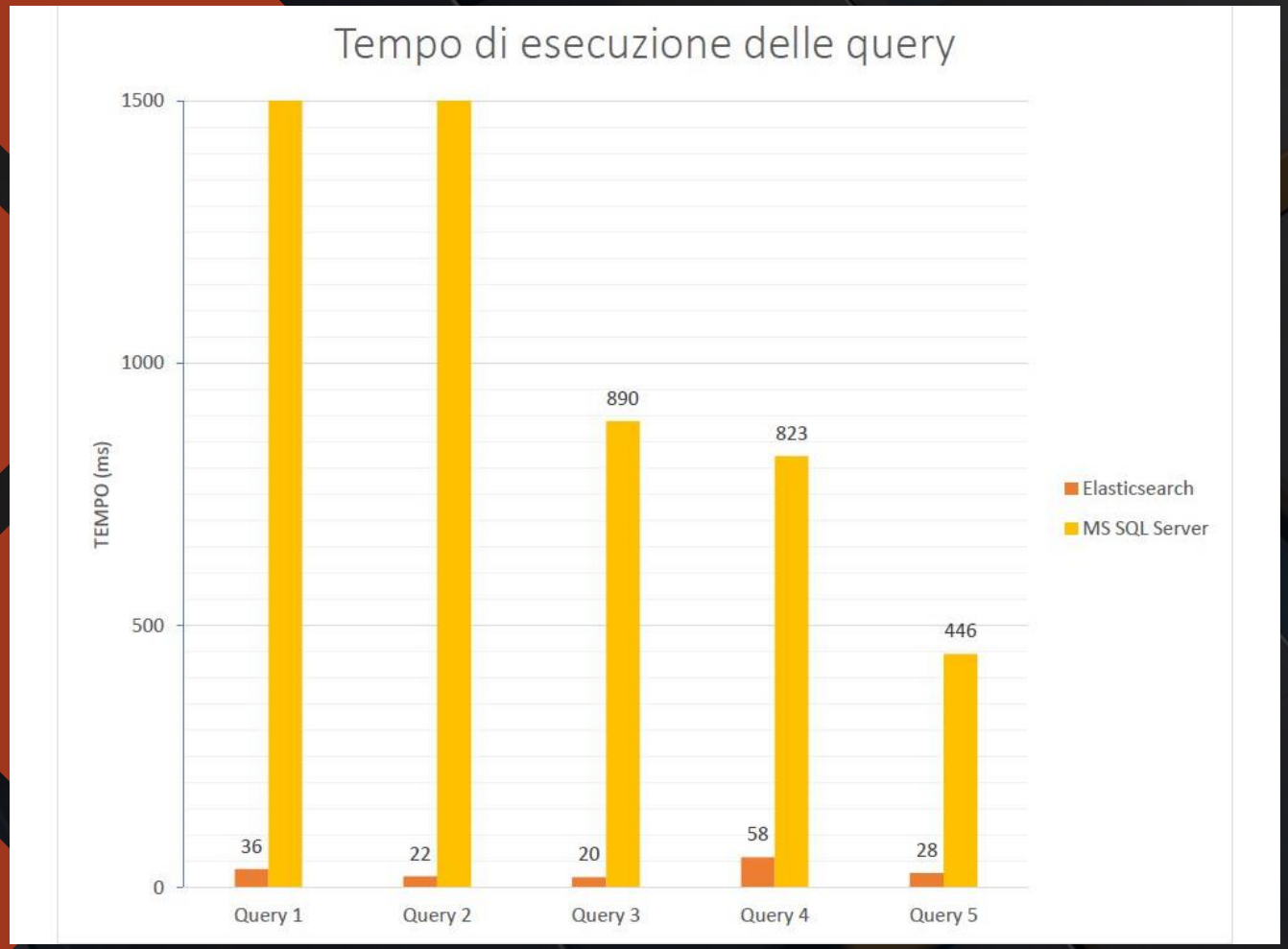
Caso di studio: Us Flights

	Elasticsearch	MS SQL Server
Query 1: tutti i voli presenti nel database	36	10344
Query 2: voli effettuati dal 20 al 28 Feb (*CONTAINS*)	22	4155
Query 3: voli che hanno "AS" nel campo/colonna "OP_CARRIER" (*FREETEXT*)	20	890
Query 4: voli effettuati in diversi giorni che hanno peso diverso (*CONTAINSTABLE*)	58	823
Query 5: voli effettuati in determinati giorni del mese (*FREETEXTTABLE*)	28	446

*Tempi di esecuzione calcolati in **ms***

- *Database di 71.4 MB nel quale sono inseriti più di 500.000 voli effettuati negli Stati Uniti nel mese di Febbraio 2020.*
- *Una sola tabella di grandi dimensioni, condizioni a sfavore di SQL Server che avrà l'operazione di join molto lenta.*
- *Eseguite le medesime ricerche su SQL Server ed Elasticsearch e sono riportati in tabella i tempi di calcolo.*
- *Computer con processore Intel® Core™ i5-8250, CPU 1.60 GHz – 1.80 GHz, RAM 8 GB.*
- *Le prime 2 query restituiscono centinaia di migliaia di righe, negli altri 3 casi decine di migliaia.*

Caso di studio: US Flights



- *MS SQL Server ha prestazioni altalenanti. Nel caso si debbano restituire molti dati, il tempo di esecuzione non è accettabile e per niente competitivo.*

- *Elasticsearch ha performance costanti e stabili. La velocità di calcolo è veloce indipendentemente dalla complessità delle query e dalla quantità di dati restituita.*

Considerazioni

- *Lentezza di join su set di grandi dimensioni.*
- *Struttura dati omogenea*
- *Non concepito per la scalabilità orizzontale*
- *Ancora molto utilizzato*
- *Stabile, ben funzionante, tecnologia consolidata*

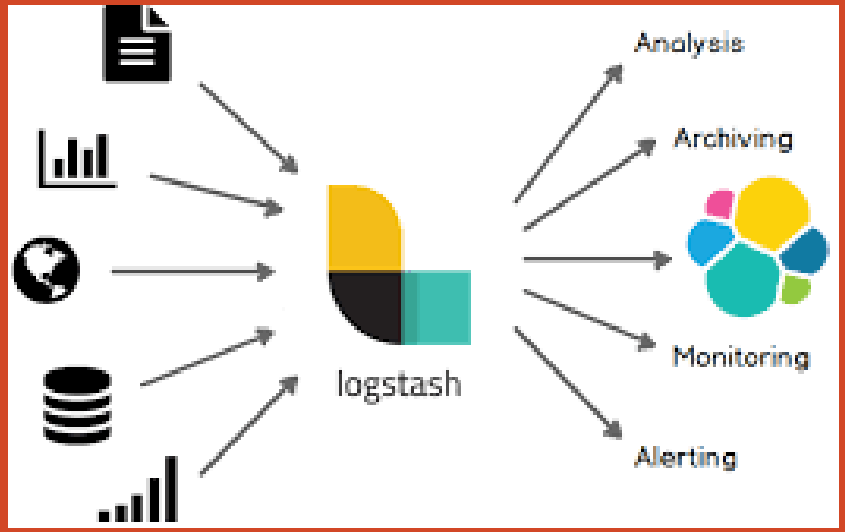
- *Funzionalità di ricerca più complete*
- *Complessità strutturale che causa possibile perdita di integrità*
- *Indicizzazioni lente e dispendiose*
- *Elastic Stack*
- *Scalabile e resiliente*
- *Non perfetto come database, ottimo come motore di ricerca*

Condivisione dati con Elasticsearch

- *Archiviare dati attraverso MS SQL Server e utilizzare Elasticsearch per le sue capacità di ricerca Full-Text.*

- *«Logstash ingerisce, trasforma e spedisce dinamicamente i dati indipendentemente dal formato o dalla complessità» (sito ufficiale Elastic).*

- *Logstash si interpone tra due software, nel nostro caso permette la condivisione di dati tra SQL Server ed Elasticsearch.*



Conclusioni

- *L'aumentare dei dati da dover ricercare in un ambiente vario e ampio come quello odierno ha reso la Full-Text Search dei sistemi SQL non compatibile con le velocità richieste dal mercato.*

- *Anche i sistemi NoSQL con grandi moli di dati risultano inefficienti rispetto ai motori di ricerca sviluppati negli ultimi anni.*

- *Affiancare tecnologie tradizionali a sistemi in grado di cercare enormi quantità di dati con rapidissimi tempi di risposta può essere ad oggi la soluzione migliore.*

- *Il futuro è strettamente legato ai database, ai dati e alla loro gestione e ricerca. Scegliere la giusta tecnologia per scopi specifici è importante e bisogna prenderne atto in tutti i contesti progettuali.*

Grazie per l'attenzione!

Tesi di Corradi Davide