**Spark integrácia s technologiami Elasticsearch a Kubernetes**

Juraj Lapčák, Martin Kubečka, Juraj Budai, Ľubomír Ševčík, Viliam Alakša

Architektúra softvérových systémov

Ing. Stanislav Marochok

8.12.2022

# Elasticsearch

Elasticsearch je distribuovaný vyhľadávací a analytický nástroj pre všetky typy údajov. Je postavený na Apache Lucene. Po prvýkrát bol vydaný v roku 2010 spoločnosťou ElasticSearch N.V.

## Základné pojmy

### Index

Index je zbierka dokumentov s podobnou dátovou štruktúrou a používa sa na ukladanie a čítanie dát. Index je definovaný jedinečńym názvom indexu respektive kľúča. Používa sa na vyhľadávanie, skenovanie alebo mazanie dát z dokumentov. Invertovaný index môžeme prirovnať k hľadaniu stránky knihy, ktorá obsahuje určité slovo. Jeden index môže obsahovať jeden typ údajov s vlastnou dátovou štruktúrou.

### Shard

Shard je podmnožina dokumentov indexu. Elasticsearch používa shards, keď je prekročený limit úložiska. Umožní tak index rozdeliť na menšie časti ktore sa daju rozdeliť medzi viacero serverov.

### Dokument

Dokument je hlavnou a základnou jednotkou informácií v Elasticsearch a je reprezentovaný vo formáte JSON. Tieto dokumenty je možné ukladať a indexovať. Index má jeden alebo viac dokumentov a dokument má jeden alebo viac poli.

### Mapovanie

Mapovanie je definícia schémy pre index.

### Node

Node(uzol) je inštancia procesu Elasticsearch. Je to server, ktorý ukladá údaje a je súčasťou indexových a vyhľadávacích funkcií Klastra.

### Klaster

Klaster sa skladá z jedného alebo viacerých uzlov (serverov), ktoré ukladajú všetky údaje a poskytujú možnosti indexovania a vyhľadávania vo všetkých uzloch. Každý klaster má jeden aktívny hlavný uzol, ktorý sa volí automaticky (napr. pri zlyhaní aktuálneho hlavného uzla).

### Replika

Replika je mechanizmus, ktorý Elasticsearch používa na riešenie zlyhaní, ako je napríklad vypnutie uzla, bez straty údajov. Je to kópia primárneho shardu a môže sa používať na vyhľadávanie rovnako ako pôvodný shard.

## Využitie Elasticsearch

Elasticsearch je schopný indexovat mnoho typov údajov. Môže byť teda využitý vo viacerćh odvetviach, ako napríklad vyhľadávanie v aplikáciách, vyhľadávanie vo webových stránkach, logovanie a analýza logov, monitorovanie kontajnerov, monitorovanie výkonu aplikácií, analýza a vizualizácia geopriestorových údajov, bezpečnostná analýza a mnoho ďalších.

## Ako Elasticsearch funguje

Elasticsearch funguje tak, že vyhľadáva a spravuje dokumentovo orientované a pološtruktúrované údaje. Funguje na architekture “zdielania ničoho”, čo je architektúra v ktorej je každá požiadavka na aktualizáciu obslúžená jedným uzlom.

Uzly tak pristupujú nezavisle na sebe, k tej istej pamäti alebo úložisku. Primárnou dátovou štruktúrou ktorú Elasticsearch používa, je invertovaný index spracovavaný pomocou API Apache Lucene. Invertovaný index je mapovanie každého jedinečného slova (Klucu) na zoznam dokumentov obsahujúci toto slovo, čo umožní velmi rychlo vyhladat dokumenty v ktorých sa nachádzajú jednotlive slova (Kluce). Informacie o indexe sa dalej ukladajú do partícií nazývaných shards. Elasticsearch dokáže dynamicky distribuovať a prideľovať shardy uzlom v klastri alebo ich replikovať. Vďaka tomuto mechanizmu je flexibilný, pokiaľ ide o distribúciu údajov. Redundanciu možno zabezpečiť rozmiestnením replík shardov ("kópie" primárnych shardov) do rôznych uzlov klastra.

<https://sematext.com/guides/elasticsearch/>

<https://www.elastic.co/what-is/elasticsearch>

<https://en.wikipedia.org/wiki/Shared-nothing_architecture>

# 

# 

# 

# 

# Kubernetes

Kubernetes je open-source kontajnerový systém na automatizáciu nasadzovania, škálovania a správy softvéru. Moderný softvér sa čoraz častejšie spúšťa ako flotily kontajnerov, niekedy nazývané mikroslužby. Kompletná aplikácia sa môže skladať z mnohých kontajnerov, ktoré musia spolupracovať špecifickými spôsobmi. Kubernetes je softvér, ktorý mení súbor fyzických alebo virtuálnych hostiteľov (serverov) na platformu, ktorá:

* hosťuje kontajnerové pracovné záťaže, poskytuje im výpočtové, úložné a sieťové zdroje
* automaticky spravuje veľké množstvo kontajnerizovaných aplikácií - udržiava ich v dobrom stave a dostupné tým, že sa prispôsobuje zmenám a výzvam

### Použitie

Pomocou Kubernetes môžeme:

* Orchestrovať kontajnery na viacerých hostiteľoch.
* Lepšie využívať hardvér na maximalizáciu zdrojov potrebných na prevádzku vašich podnikových aplikácií.
* Riadiť a automatizovať nasadzovanie a aktualizácie aplikácií.
* Pripájať a pridávať úložisko na spustenie stavových aplikácií.
* Škálovať kontajnerové aplikácie a ich zdroje za chodu.
* Deklaratívne spravovať služby, čo zaručuje, že nasadené aplikácie budú vždy bežať tak, ako ste ich zamýšľali.
* Kontrolovať stav a samo regenerovať aplikácie pomocou automatického umiestňovania, automatického spúšťania, automatickej replikácie a automatického škálovania.

### Setup

V našom prípade potrebujeme spustiť klaster Kubernetes na svojom notebooku (lokálne). Preto potrebujeme najprv nainštalovať nástroj Minikube.

<https://minikube.sigs.k8s.io/docs/start/>

# Apache Spark

Apache Spark je multi-jazyčný nástroj používaný v dátovom inžinierstve, data science a umelej inteligencii pre jedno uzlové klastre alebo mašiny. Spark poskytuje rýchle iteratívne, funkcionálne riešenia pre veľké data sety, zvyčajne použitím cache dát vo vyrovnávacej pamäti. Apache Spark je výpočtový framework, ktorý nie je nutne prepojený s Map/Reduce Elasticsearch integráciami, ale funguje na podklade Hadoop integrácie HDFS. elasticsearch-hadoop povoľuje Elasticsearch-u používať Spark dvoma spôsobmi: pomocou dedikovanej podpory od Elasticsearch verzie 2.1 alebo pomocou Map/Reduce integrácie od 2.0. Spark 2.0 je podoporovaný v elasticsearch-hadoop od Elasticsearch verzie 5.0.

Navyše je Apache Spark dostupný pre mnoho programovacích jazykov, najmä: Java, Scala, Python, R, C#, F#. No ako integrácia Elasticsearch je Apache Spark API dostupné pre jazyky: Java, Scala a Python. My v rámci implementácie a demoštrácie sme používali Apache Spark API pre jazyk Java.

<https://www.elastic.co/guide/en/elasticsearch/hadoop/current/index.html>

## Komponenty Apache Spark

1. Spark Core
2. Spark SQL
3. Spark Streaming
4. Spark MLlib
5. GraphX
6. Spark R

<https://www.knowledgehut.com/tutorials/apache-spark-tutorial/apache-spark-components>

## Integrácia v Elasticsearch

Pridanie integrácie v rámci Elasticsearch bolo jednoduché a pozostávalo z nasledujúcich krokov.

1. Menu -> Add integration -> vo vyhľadávacom poli hľadáme Apache Spark
2. V čase implementácie, v novembri 2022, bolo Apache Spark API v rámci Elasticsearch v beta verzii 0.3.0
3. Následne bola potrebná konfigurácia Spark integrácie. Integrácia nám poskytla možnosť zbierania Apache metrík pomocou služby **Jolokia.** V tom prípade nutnosť vyplnenia Host endpointu pre Apache spark aplikačné metriky, drive metriky, executor metriky a node metriky.. My sme túto možnosť nevyužili

# 

# Docker

Docker je open-source platforma na vývoj a spúšťanie aplikací. Docker umožňuje oddeliť aplikáciu od infraštruktúry, aby sa dal software rýchlejšie deployovat. Taktiež skracuje čas testovania a oneskorenia medzi vývojom a nasadenim kodu.

Docker poskytuje možnosť zabalit aplikáciu do package a spustiť ju v izolovanom prostredi nazývanom kontajner. Izolácia umožňuje spustiť na hostovy viacero kontajnerov súčasne.

## Architektúra

Docker používa architektúru klient-server. Klient Docker komunikuje s démonom Docker, ktorý spúšťa a distribuuje kontajnery

