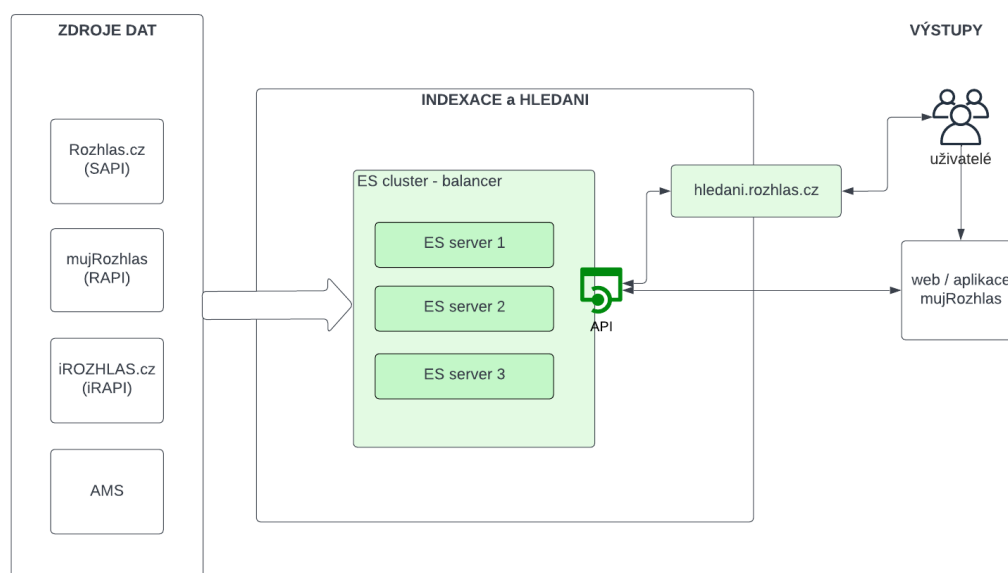


Příloha č.3 – Technická specifikace: Rozvoj indexace obsahu a vyhledávání pro pilíře mujRozhlas, Rozhlas a iRozhlas

Indexace a vyhledávání pro internetové pilíře ČRo je vybudováno na platformě ElasticSearch. Základem celého řešení je Elasticsearch v 7.5 cluster (ES můj rozhlas) instalovaný na 3 serverech v síti ČRo spojených do jednoho clusteru. Backend ElasticSearch je doplněn o frontend aplikaci hledani.rozhlas.cz která komunikuje s backendem přes API, které je současně vystaveno pro využití externích aplikací jako web a mobilní aplikace mujRozhlas. Předmětem zakázky je rozvoj backendu, API a frontendu vyhledávání dle potřeb ČRo (vyznačeno zeleně na obrázku níže).



Zdroje dat pro indexaci:

Zdrojem dat jsou následující API která podávají data z Rozhlas.cz a mujRozhlas.cz

SAPI - data z Rozhlas.cz

- <https://rozhlas.cz/sapi>
- Používané endpointy: nodes, files
- nahrazuje standardní OAI modul

iRAPI - data z iROZHLAS.cz

- <https://www.irozhlas.cz/api/v2>
- používané endpointy: articles-indexer, authors-indexer, pages-indexer, terms-indexer

RAPI - data z mujRozhlas

- <https://mujrozhlas.croapi.cz>
- Používané endpointy: topics, persons, keywords, stations, serials, shows, episodes-indexer

Další zdrojem dat je přepis mluveného slova (systém AMS), kde nabíráme několik typů obsahu:

- přepis živého vysílání - websocket aplikace (brutt přepis)
 - `ws://mrcix01:8080/livestream/ingest/[nazevStanice]`
- offline přepisy dat z disku (od 2003 do roku 2020) (přepis archivu)
 - NFS: `pr-s-fs001.cro.cz/mrams /mnt/mrams01`
- online data přes API (od roku 2020) (deep přepis živého vysílání)
 - `https://api.mujirozhlas.cz/ams/stt/station/`
- online indexace autorských dat přes API (přepis ondemand)
 - `https://www.irozhlas.cz/admin/ams/indexer?page=0`

Indexace dat

Indexace dat probíhá v následujících krocích:

- identifikace poslední aktualizace dat
- stažení nových/změněných dokumentů (epizod, pořadů ...) pomocí REST API, websocket či přímo ze souborů.
- analýza dokumentu, rozšíření o potřebná data (např. detaily pořadů, stanic, lidí)
- normalizace dokumentu
- zápis do Elasticsearch indexů
- detekce smazaných záznamů a jejich odmazání z indexu

Indexery jsou psány v jazyce JAVA vždy pro konkrétní typ dat. Jsou umístěny na jednom serveru v rámci ES clusteru a spouštěny periodicky pomocí cronu.

Aplikace a výstupy pro vyhledávání

Na všech ES serverech je nainstalován webový server Tomcat a je spojen přes loadbalancer. Všechny aplikace následně přistupují pouze přes rozhraní loadbalanceru. Všechny aplikace jsou napsány v jazyce Javy a nasazeny pomocí war souborů.

Seznam aplikace/contextů

Context	Účel
/extranetAll	stará <code>https://hledani.rozhlas.cz/stare/</code>
/hledani	UI stávajícího <code>https://hledani.rozhlas.cz</code>
/hledani-api	API <code>https://hledani.rozhlas.cz/api</code>
/hledani-beta	UI testovacího verze nového hledání <code>https://betasearch.rozhlas.cz</code>
/hledani-beta-api	API <code>https://betasearch.rozhlas.cz/api</code>
/iRadio	UI starého audioarchivu <code>https://hledani.rozhlas.cz/iradio/</code>
/irozhlas	API pro mobilní apku irozhlasu

/livestream	websocket pro přijímání textové analýzy z živého vysílání
/mujrozhlas	API pro web/mobilní aplikaci mujrozhlas
/mujrozhlas_stage	stage API pro web/mobilní aplikaci mujrozhlas
/searchstats	API pro zápis statistik

API pro hledání

API slouží pro vyhledávání a je mezivrstvou mezi UI (mobilní aplikací, webové vyhledávání aj.) a Elasticsearchem. Jednotlivá API se liší zejména definicí indexů, ve kterých se hledá a definicí dotazů, jak se hledá. Komunikuje se s následujícími endpointy:

- /search – vrací seznam výsledků a agregací
- /aggregatesearch – vrací seznam výsledků sloučených dle typu obsahu (epizoda, pořad ...) a agregací
- /suggest – našeptávač. Na základě minimálně 2 písmen jsou vráceny názvy odpovídajících objektů

Webové UI rozhraní

Webové rozhraní je dostupné na adrese <https://hledani.rozhlas.cz>. Jedná se o Angular aplikaci komunikující přes API s vyhledáváním Elasticsearch.

Ukázky indexovaných dat (RAPI, SAPI)

SAPI - <https://portal.rozhlas.cz/sapi/nodes?dateFrom=2023-01-29T20:00:00+01:00&dateTo=2023-01-29T20:10:00+01:00>

RAPI - <https://api.mujrozhlas.cz/episodes-indexer/12c67855-dcc7-4597-a45b-44623094147f>

Odezva systému

Získání výsledku dotazu nesmí přesáhnout 3 vteřiny.