Národní knihovna České republiky

Tvorba intermediary formátu I.

Mgr. Zdenko Vozár

Workshop - VYUŽITÍ DAT WEBOVÝCH ARCHIVŮ: Možnosti a limity

2019 - Praha 21. 10. 2019

Národní knihovna ČR

zdenko.vozar@nkp.cz

Webarchív CZ a jeho data: Obsah

Data Kernels (3) Discussion (1) Activity Metadata

Usability 8.2 License Other (specified in description)

Description

Le Grand Débat National (décembre 2018 à avril 2019)

À l'initiative du Président de la République, le Gouvernement français engage un **grand** des grands enjeux de la nation : *la fiscalité et les dépenses publiques, l'organisation de l' la démocratie et la citoyenneté.*

Webarchív a jeho data

Povaha webarchívu

- cez 400 TB historických dát
  - ročný prírastok 20-40 TB
- hierarchické úložište
- serializované do niekoľkých dobre dokumentovaných formátov podľa ISO
  - WARC (1 000 MB,)
  - ARC (100 MB)
  - dokopy 900 000 jednotiek
  - CDX
- performance - zatiaľ good enough
- user friendly UI?

Riziká

- sprístupňovanie
  - index v CDX
  - čokoľvek nad rámec indexu - masívne
- archivácia
  - staršie data bez metadát
  - bitová ochrana
  - retencia
- špecifikum
  - množstvo podobných alebo rovnakých dát
  - deduplikácia
- úlohy interaktívne a iteratívne

Príležitosti Webarchívu NK ČR: Knižnica 2.0

a. Rozšírenie užívateľských operácií (KP)

- optimalizácia dostupnosti dát a obecne škálovateľnosti operácií WA
- WAT, WET
- prezentácia (BE):
  - OLAP (Online analytical processing)
  - WORM joby (Write once, read many)
- Discovery UI: Fulltext, Filtre, Advanced search, Kolekcie
- Export UI:
  - Výsledky hľadania
  - Preprocessed public datasets online
- Poskytnutie výpočtových prostriedkov:
  - Data scientist module onsite: Jupyter NB
  - Cluster analysis on demand

b. Obohatenia metadát:

- preprocessing
  - historická analýza (Grainery)
  - NER
  - tokenizácia
  - sentiment
  - analýza zvuku a videa
- výskum a TDM:
  - korpusová analýza, NLP
  - linkovanie, síťové grafy
  - analýza komplexních sítí
  - knowledge trees
  - kategorizácia, topic identification

- Grainery
- Analýza hierarchického úložišťa WA a jeho sémantiky
- súpis všetkých WARC, ARC, CDX a logov
- ich analýza a tvorba jedinečného záznamu harvest
  - základná typológia: Serials, Topics, Totals, Requests
  - frekvenčné subtypy: T, V, 1M, 2M, 6M, 12M, V1,

Resilient distributed datasets (RDDs)

Kolekcia premenných

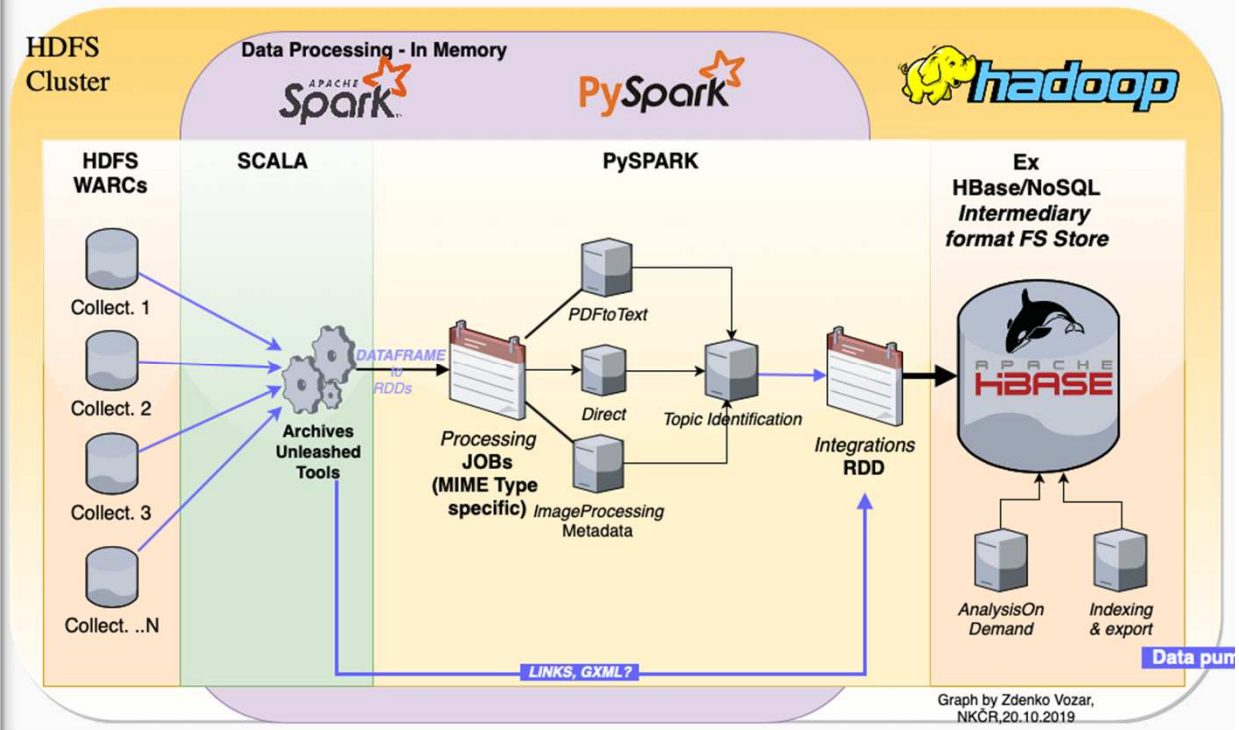
- distribúcia v nodoch, bloky 128 MB bdf., fault tolerant, elem. nemenné
- paralelné spracovanie
  - načítanie dynamicky - pamäť df
  - načítanie ako externý zdroj (Ex) - zdieľaný FS
    - Cloud, AWS S3
    - HDFS, HBase, Cassandra
    - rsp. vysoko dostupné a škálovateľné NoSQL)
- efektívnejšie M-R operácie (replikácia, serializácia, disk I-O)
  - iteratívne a interaktívne ops - sys. overheads (RW op.)
  - data sharing a perzistentnosť (ideálne v distr. pamäti, či Ex)
- obsahuje akýchkoľvek typ objektov z Java, Python, Scala

Vznik a Operácie nad kolekciou

- Transformácia z Hadoop InputFormat (split, recReader)
  - všetko na čo sa dá aplikovať, rsp expost definovať M-R job (fc. vs. agg)
  - textové súbory, sekvenčné s., H.Input Format súbory
- Transformácie
  - map(f, preservation), filter(f), groupBy), join(RDD), pipe
- Akcie
  - reduce, count, collect/take, foreach(f), cache, join

Spracovanie dát do Intermediary formátu

1. Vystavenie kolekcie: WARC Data pump - Storage / Crawl
2. Archives Unleashed Toolkit
   a. Preprocessing
      i. Rem. boilerplate - implementácia jusText
      ii. Rem. Duplic/NearDupl.
   b. Extrakcia linkov
3. Pokročilé spracovanie do IM formátu v PySPARK
   a. Near-duplicates analysis
   b. Extrakcia NER
   c. Kategorizácia s ML
   d. Katalogizácia
4. Kompletácia DF do RDD
5. Indexácia v SOLR
6. Distribúcia



Rizikové faktory

- Udržateľnosť projektu
  - financovanie opráv stávajúcej HW infraštruktúry
  - integrácia s ďalšími elektronickými zdrojmi NK ČR
- Rozšírenie projektu na celý Webarchív
  - Nákup nového HW a aktualizácia všetkých komponent SW
  - Nové vývojové práce
  - Personálne zdroje
- Distribúcia do ďalších knižníc
- Široké očakávania výskumníkov

Národní knihovna České republiky

Tvorba intermediary formátu I.

Mgr. Zdenko Vozár

Workshop - VYUŽITÍ DAT WEBOVÝCH ARCHIVŮ: Možnosti a limity

2019 - Praha 21. 10. 2019

Národní knihovna ČR

zdenko.vozar@nkp.cz