

WACloud: Akvizice dat, formát WARC a procesní a datová integrace do HDFS řešení

NK ČR, Praha 17.10. 2022

Mgr. Zdenko Vozár

National Library of the Czech Republic

Úvod

- Akvizícia dát: Systém sklizní a ich AS IS zpracovanie
- Kontajner WARC 1.0
- Zámer riešenia WACloud
- TRS + TOBE infraštruktúra zpracovania dát

Akvizícia dát

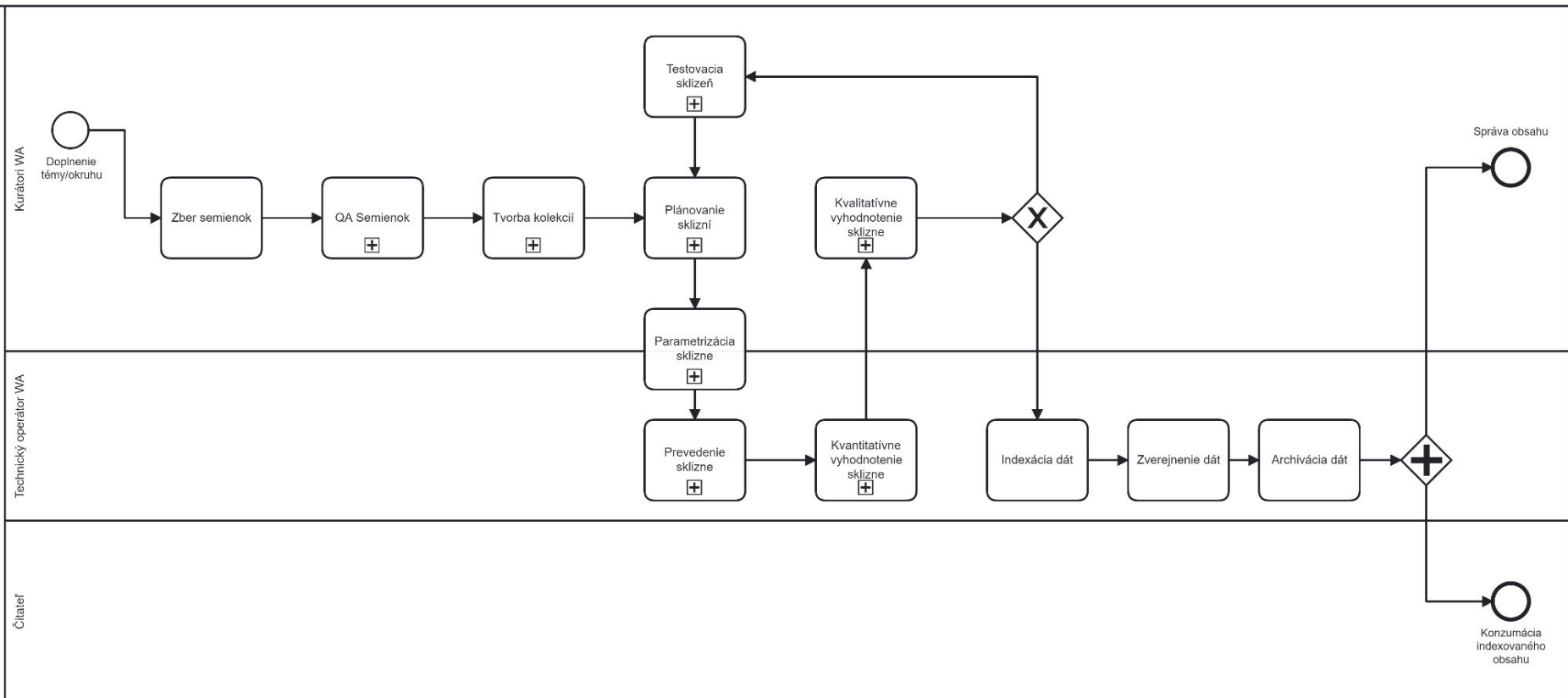
System sklizní a ich AS IS zpracovanie

Akvizícia dát Webarchívu ČR (2000 - teraz)

Relevantný bohémikálny obsah (územie, jazyk, autorstvo or téma/obsah), nielen “.cz”

- **Objem:** 435 TB zazipovaných dát
- **Ročná akvizícia:** 25-50 TB zazipovaných dát
 - **Formáty:** Veľké množstvo (/long tail)
 - **Frekvencia:** Po hodinách - denne - x mesačne - ročne
 - **Trvanie:** Hodiny - cca pol mesiaca
- **Bežné typy:** Serials, Topics, Tests, Continuous / Requests
- **Celoplošná sklizeň:** Totals

Hlavný proces akvizície a publikácie dát



Technologický stack:

OS:	OpenSUSE, Centos
OS Docker:	Rôzne
WebServ:	NGINX
Middleware:	Tomcat
DB:	MariaDB / BDB
index:	CDX

Úložiště:

- Hierarchické - GPFS
- LTO - bit prot.

Plánovač sklizní a
správa semienok:

- Seeder

Harvestre:

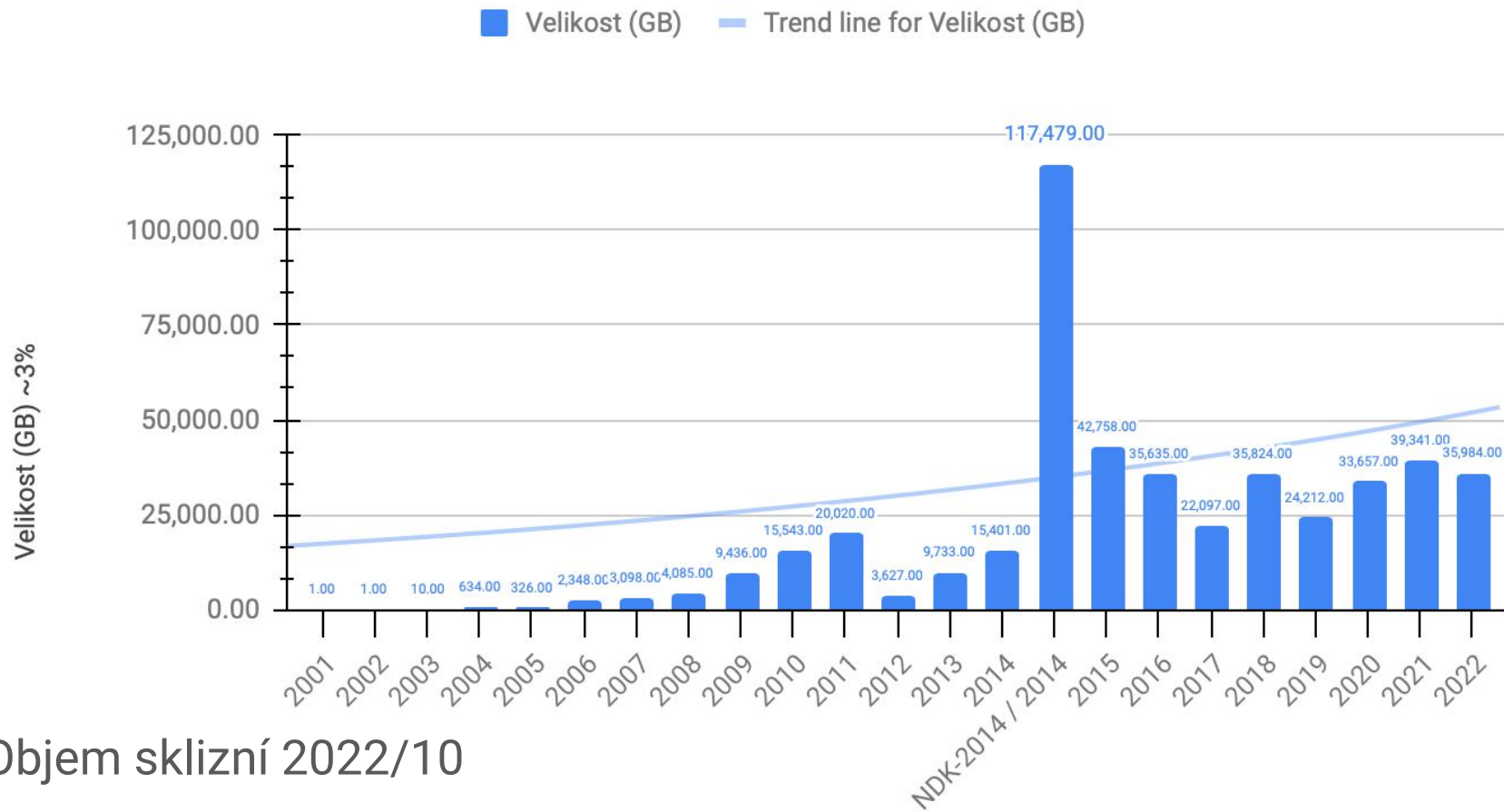
- Heritrix
- Webrecorder

Zobrazovač:

- OpenWayback
- pyWayback (TB)

Velikost (GB) vs. Rok

NK ČR: Z.Vozár, 2022/10/17



Objem sklizní 2022/10

Kontajner WARC 1.0

Ukážka obsahu

WARC/1.0

Hlavička a metadata

WARC/1.0

WARC-Type: warcinfo

WARC-Date: 2022-02-25T17:11:44Z

WARC-Filename:

Continuous-UkraineWar-2022-02-25-UkraineWar2022_crawler18--20220225171144760-00000-13139~crawler18.webarchiv.cz~7778.warc.gz

WARC-Record-ID:

<urn:uuid:db3cd15e-e3b4-4f4e-895d-765d7b08b4eb>

Content-Type: application/warc-fields

Content-Length: 635

== Harvest metadata part ==

software: Heritrix/3.4.0-20210923 <http://crawler.archive.org>

ip:

hostname: crawler18.webarchiv.cz

format: WARC File Format 1.0

conformsTo:

http://bibnum.bnf.fr/WARC/WARC_ISO_28500_version1_latestdraft.pdf

operator: Zdenko Vozar

publisher: National Library of the Czech Republic - Webarchiv.cz

audience: Webarchiv.cz Users

isPartOf: Continuous-UkraineWar 2022-02-25-UkraineWar2022

description: Denni sklizen pri prilezitosti Ruskeho vpadu na Ukrainu.

robots: ignorepolicy

http-header-user-agent: Mozilla/5.0 (compatible; heritrix/3.4.0-20210923 +<http://webarchiv.cz/kontakty>)

http-header-from: webarchiv@nkp.cz

WARC/1.0

Initial text/DNS responses

WARC/1.0
WARC-Type: response
WARC-Target-URI: dns:denikn.cz
WARC-Date: 2022-02-25T17:11:44Z
WARC-IP-Address: 8.8.8.8
WARC-Record-ID:
<urn:uuid:840afcc0-fb8f-4da7-99db-3f70d11e27ab>
Content-Type: text/dns
Content-Length: 46

```
20220225171144
denikn.cz.           78      IN      A
92.60.51.9
```

WARC / 1.0: L. 6152

HTML Content
Content-Type: application/http

WARC/1.0

WARC-Type: response

WARC-Target-URI:

<https://www.europarl.europa.eu/news/cs/press-room/20220210IPR23007/vyhruzky-ruska-vu-ci-ukrajine-jsou-pro-evropu-varovnym-signalem-tvrdi-poslanci>

WARC-Date: 2022-02-25T17:11:51Z

WARC-IP-Address: 136.173.69.97

WARC-Payload-Digest: sha1:YNN EGLZAM3AD57FOWX5SS3HN2B4HRHPD

WARC-Record-ID: <urn:uuid:75b74d2d-8b66-4872-9c56-6e9a43f265f6>

Content-Type: application/http; msgtype=response

Content-Length: 132371

HTTP/1.1 200 OK

Server: nginx

Date: Fri, 25 Feb 2022 17:11:47 GMT

Content-Type: text/html; charset=UTF-8

Content-Length: 131766

Set-Cookie: engnewsroomroute=node1; path=/

Last-Modified: Fri, 25 Feb 2022 17:11:47 UTC

Cache-Control: max-age=60, must-revalidate

Set-Cookie: JSESSIONID=CF2CB41F162DA9EE06A06A36D6AB0478; Path=/news;

HttpOnly

ETag: "006799b806fafe519b6e3a9d4fc8e37df"

Content-Language: cs-CZ

X-Content-Type-Options: nosniff

X-XSS-Protection: 1; mode=block

X-Frame-Options: DENY

Referrer-Policy: same-origin

Content-Security-Policy: upgrade-insecure-requests

Connection: close

<!DOCTYPE html>

<html lang="cs">

WARC / 1.0: l. ff

HTML head

```
<!DOCTYPE html>  
<html lang="cs">  
  <head>
```

```
    <title>Výhrůžky Ruska vůči Ukrajině jsou pro Evropu varovným signálem, tvrdí poslanci | Zpravodajství | Evropský parlament</title>
```

```
    <meta property="og:title" content="Výhrůžky Ruska vůči Ukrajině jsou pro Evropu varovným signálem, tvrdí poslanci | Zpravodajství | Evropský parlament"/>
```

```
    <meta property="twitter:title" content="Výhrůžky Ruska vůči Ukrajině jsou pro Evropu varovným signálem, tvrdí poslanci | Zpravodajství | Evropský parlament"/>
```

```
    <meta itemprop="name" content="Výhrůžky Ruska vůči Ukrajině jsou pro Evropu varovným signálem, tvrdí poslanci | Zpravodajství | Evropský parlament"/>
```

```
    <meta name="description" content="V rozpravě o vztazích mezi EU a Ruskem, evropské bezpečnosti a ruské vojenské hrozbě vůči Ukrajině poslanci vyzvali k jednotné reakci a vyjádřili podporu Ukrajině."/>
```

```
    <meta property="og:description" content="V rozpravě o vztazích mezi EU a Ruskem, evropské bezpečnosti a ruské vojenské hrozbě vůči Ukrajině poslanci vyzvali k jednotné reakci a vyjádřili podporu Ukrajině."/>
```

```
    <meta property="twitter:description" content="V rozpravě o vztazích mezi EU a Ruskem, evropské bezpečnosti a ruské vojenské hrozbě vůči Ukrajině poslanci vyzvali k jednotné reakci a vyjádřili podporu Ukrajině."/>
```

```
    <meta itemprop="description" content="V rozpravě o vztazích mezi EU a Ruskem, evropské bezpečnosti a ruské vojenské hrozbě vůči Ukrajině poslanci vyzvali k jednotné reakci a vyjádřili podporu Ukrajině."/>
```

WARC/1.0: L. 7391

Actual Content

<!-- Text content -->

<div class="ep_gridcolumn ep-m_product" data-view1200="6" data-view1020="6" data-view750="10" data-view640="6" data-view480="8" data-view320="4">

<div class="ep_gridcolumn-content" >

<div class="ep-a_text"><p class="ep-wysiwig_paragraph">Poslanci ve středu dopoledne v rámci plenární rozpravy s předsedou Evropské rady Charlesem Michelem, předsedkyní Evropské komise Ursulou von der Leyenovou a vysokým představitelem Unie pro zahraniční věci a bezpečnostní politiku Josepem Borrellem zhodnotili nejnovější vývoj v souvislosti s ruskými vojenskými hrozbami vůči Ukrajině.</p>

<p class="ep-wysiwig_paragraph">Předsedkyně Parlamentu Roberta Metsolová při zahájení rozpravy zdůraznila, že Evropský parlament opakovaně vyjádřil solidaritu s ukrajinským lidem, který stále čelí nejistotě a hrozbám ruské vojenské agrese.</p>

=====
=

<p class="ep-wysiwig_paragraph">Mnozí poslanci vyjádřili trvalou podporu a obdiv ukrajinskému lidu, který již léta čelí hrozbě ruské agrese, a zároveň zopakovali, že je třeba pokračovat v diplomatických jednáních s Moskvou a připravit proti Rusku tvrdé sankce. Mezi nimi musí být řada opatření, včetně těch týkajících se plynovodu Nord Stream 2 vedoucího z Ruska do Německa, uvedli někteří.</p>

<p class="ep-wysiwig_paragraph">Poslanci také poukázali na to, že důvodem ruské agresivity není rozšiřování NATO, ale spíše síla hodnot a přitažlivost demokratických společností, která děsí ruského prezidenta Vladimira Putina a Kreml. Někteří poslanci také kritizovali EU za příliš nejednoznačnou reakci na kroky Ruska. Další zdůrazňovali, že při potlačování ruské agrese musí po slovech následovat činy.</p>

<p class="ep-wysiwig_paragraph">Záznam plenární rozpravy je k dispozici zde.</p></div>

</div>

WARC/1.0: eor l. 8762

End of record

</html>

WARC/1.0

WARC-Type: request

WARC-Target-URI:

<https://www.europarl.europa.eu/news/cs/press-room/20220210IPR23007/vyhruzky-ruska-vuci-ukrajine-jsou-pro-evropu-varovnym-signalem-tvrdi-poslanci>

WARC-Date: 2022-02-25T17:11:51Z

WARC-Concurrent-To: <urn:uuid:75b74d2d-8b66-4872-9c56-6e9a43f265f6>

WARC-Record-ID: <urn:uuid:b3d2a226-e9fc-4937-8331-64e1861d1f3c>

Content-Type: application/http; msgtype=request

Content-Length: 370

GET

[/news/cs/press-room/20220210IPR23007/vyhruzky-ruska-vuci-ukrajine-jsou-pro-evropu-varovnym-signalem-tvrdi-poslanci](https://www.europarl.europa.eu/news/cs/press-room/20220210IPR23007/vyhruzky-ruska-vuci-ukrajine-jsou-pro-evropu-varovnym-signalem-tvrdi-poslanci) HTTP/1.0

From: webarchiv@nkp.cz

Connection: Close

Accept: text/html,application/xhtml+xml,application/xml;q=0.9,*/*;q=0.8

Host: www.europarl.europa.eu

User-Agent: Mozilla/5.0 (compatible; heritrix/3.4.0-20210923 +http://webarchiv.cz/kontakty)

WARC/1.0

WARC-Type: metadata

WARC-Target-URI:

<https://www.europarl.europa.eu/news/cs/press-room/20220210IPR23007/vyhruzky-ruska-vuci-ukrajine-jsou-pro-evropu-varovnym-signalem-tvrdi-poslanci>

WARC-Date: 2022-02-25T17:11:51Z

WARC-Concurrent-To: <urn:uuid:75b74d2d-8b66-4872-9c56-6e9a43f265f6>

WARC-Record-ID: <urn:uuid:ead5bb5c-b4eb-4ea9-91c1-b95041cb50b5>

Content-Type: application/warc-fields

Content-Length: 18199

seed:

fetchTimeMs: 304

charsetForLinkExtraction: UTF-8

outlink: <https://www.europarl.europa.eu/favicon.ico> | =INFERRED_MISC

outlink: <mailto:hana.raissi@europarl.europa.eu> | a/@href

outlink: [tel:\(+420\)225520708202671](tel:+420225520708202671) | a/@href

Zámer WACloud

Práca s volatilnými digital born datami

Data mining

- Stanovenie workflow datového spracovania
- Pomocou využitia modelov strojového učenia
- Extrakcia textu
- Extrakcia zvuku
- Sémantická analýza
- Analýza tém

Analýza dát webarchívu

- Legislatívne posúdenie
- Kvantitatívna štatistická analýza
- Sémantická analýza
- Analýza tém
- Sieťová analýza
- Sociologická analýza

Umožnenie exportov dát nad užívateľským výberom

- Formáty: JSON, TXT, CSV sec.
- Prostredie: UI, secure REST API
- Princíp:

I. Agregácia on demand via metadata

II. Plánovanie extrakcie vzoriek / veľkých dát

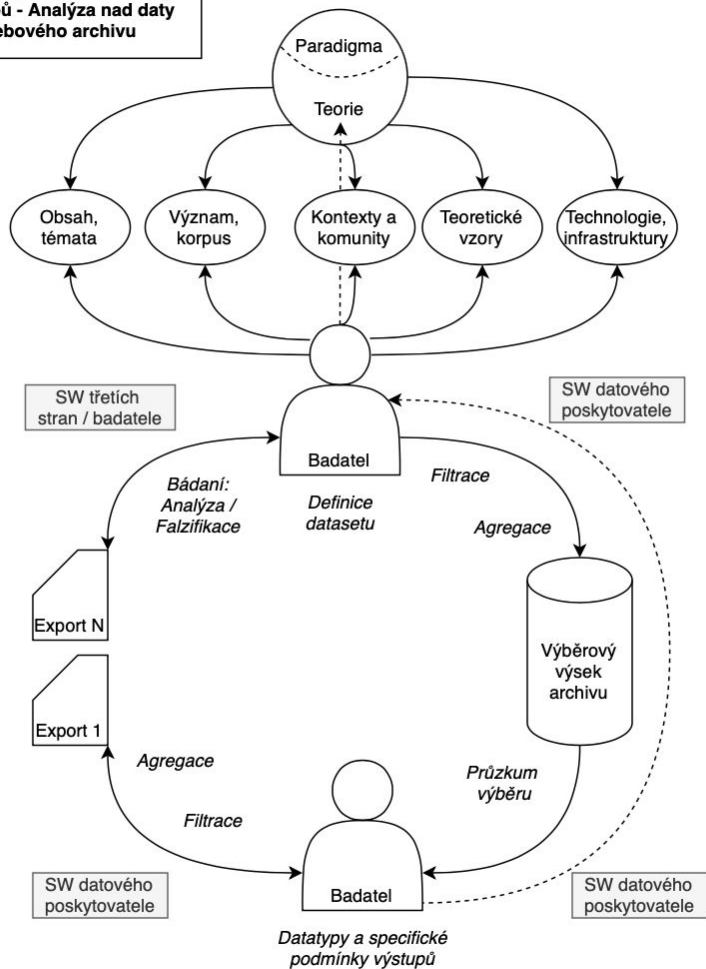
III. Iteratívne úpravy požiadavkov

Exporty dát

Prostřednictvím filtračních mechanismů:

- Administrativně metadáta sklizní
- Modelovanie:
 - Typ stránky
 - Téma, Sentiment
- Lexikální analýza:
 - Tokeny, Tokenizované vety
 - Exp: Kolokácie, Plné Texty
- Sieťová analýza
 - Podľa referencovania domén
 - Exp: CSV, JSON



Ideotypický návrh badatelských postupů - Analýza nad daty webového archivu



Analýza nad daty: Postup WACloudom

Infraštruktúra a dátová integrácia

Základy WACloudu

  [Webarchiv](#) [New query](#) [Favorite](#) [My queries](#) [FAQ](#) [User](#)

FILTERS <

Theme

Theme

Page type

Page type

Date of harvest

From

To

URL

Operator

URL

Sentiment

Sentiment

SETTINGS OF LIMITS

Stop words

a, aby, aj, ale, ani, asi, atd, atp,...

Number of entries

Number of entries

Random records?

QUERY

Logical operators: ()

HARVESTS

Te	Se	To
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Harvest's name: Topics_2022-09-T-TestiRozhrani	Harvest's name: Serials-2021-02-1M-2M_OneShot-crawler00	Harvest's name: Topics-2018-01-T2018VolPrezII-crawler01_
Size: -	Size: -	Size: -
Number of WARC's: 126,646	Number of WARC's: 4,735,803	Number of WARC's: 11,852
Date of start: 12/09/2022	Date of start: 26/02/2021	Date of start: 12/01/2018

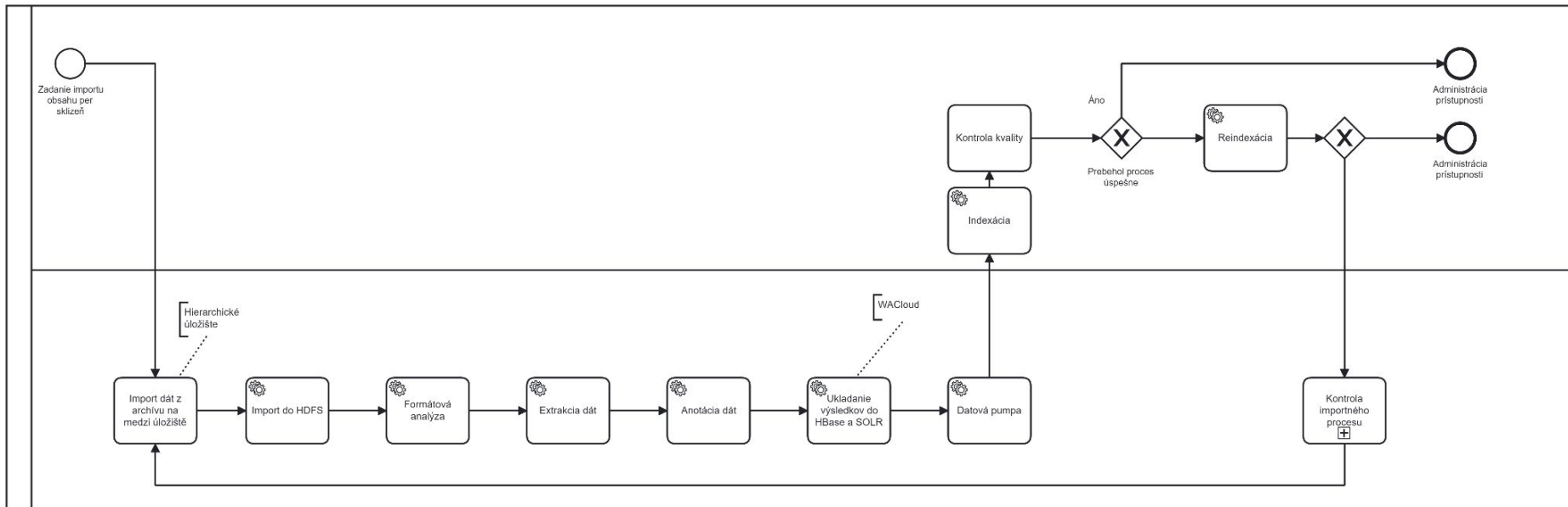
[Continue](#)

UAT vs PROD

Centos 7:	Ambari
Ubuntu 20.4 LTS:	BigTop
DB:	HBase
Index:	SOLR

		TYP	RAM (GB)	CPU	SAS/SSD OS (GB)	SAS/SSD DATA (GB)	SATA DATA
APP	UAT	VMWARE	20	4 vCPU	080	2,000	-
	PROD	Fyz. (HUAWEI RH2288 V3)	128	8 core/CPU	480	2,880	-
MN	UAT	VMWARE	32	4 vCPU	080	2,000	10,000
	PROD	Fyz. (HUAWEI RH2288 V3)	188	8 core/CPU	100	380	-
DN	UAT	VMWARE	16	4 vCPU	080	-	10,000
	PROD	Fyz. (HUAWEI RH2288 V3)	128	8 core/CPU	480	-	048

Proces integrácie dát z GPFS do WACloudu



Ďakujem za pozornosť !

Mgr. Zdenko Vozár

zdenko.vozar@gmail.com

National Library of the Czech Republic