OBRADA LOGOVA WEB SERVERA

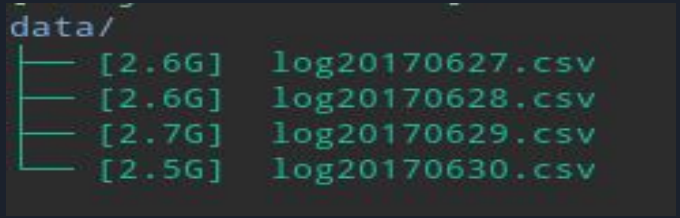
Danijel Radaković, R1 20/2019

DATA SET

- **EDGAR Log File Data Set** - sadrži informacije o pretragama koje su izvršene pomoću *SEC.gov* web sajta u periodu od 14.02.2013. do 30.06.2017. godine
- *SEC - Securities And Exchange Commission*
- Veličina *Data Set*-a je otprilike 13 TB, od čega je 10 GB obrađivano u projektu
- Data Set se sastoji od **csv** fajlova u kojem se nalaze *access* logovi **Apache Web Server**-a u anonimizovanom formatu

DATA SET

- Obrada se vrši nad logovima nastali u periodu od 26.06.2017. do 30.06.2017.
- Format loga je opisan ovde

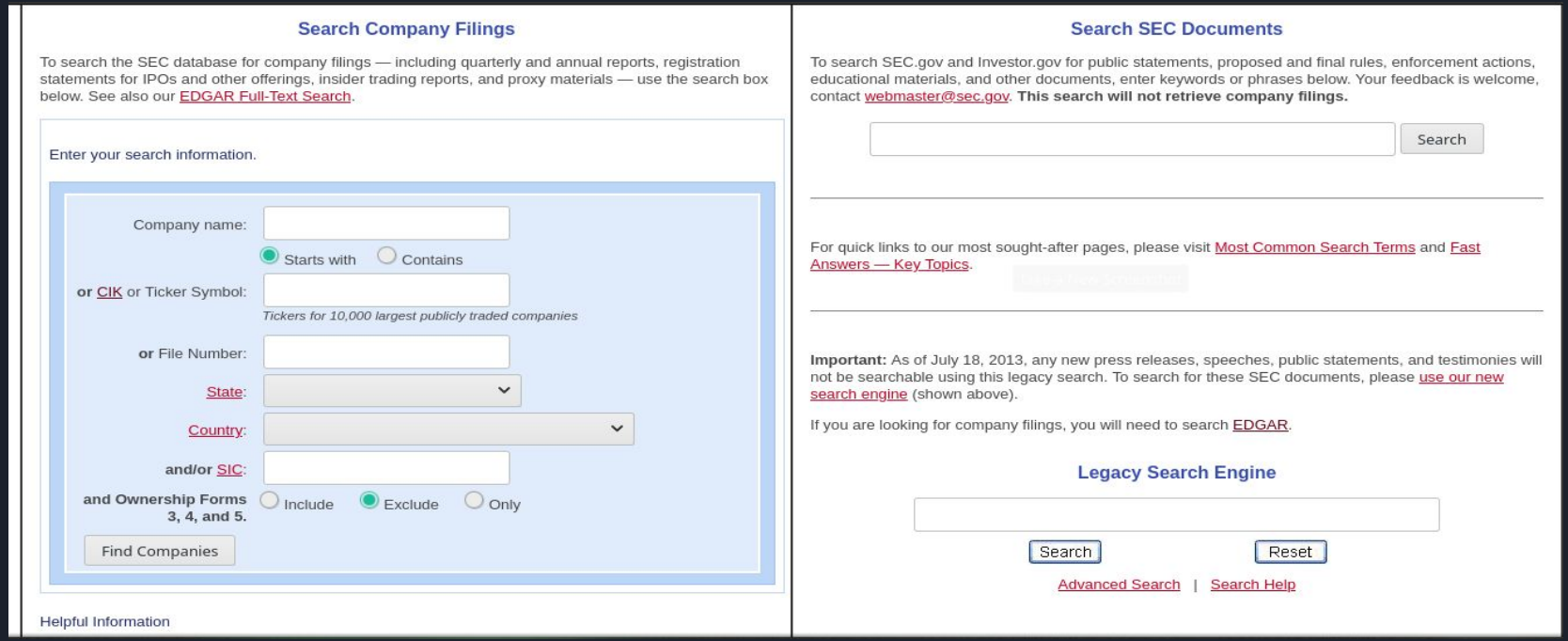


DATA SET - FORMAT LOGA

```
ip,date,time,zone,cik,accession,extention,code,size,idx,norefer,noagent,find,crawler,browser
101.81.229.jeb,2017-06-27,00:00:00,0.0,83402.0,0000083402-97-000006,-index.html,200.0,7492.0,1.0,0.0,0.0,10.0,0.0,
101.81.229.jeb,2017-06-27,00:00:00,0.0,748592.0,0001019687-09-002913,-index.htm,200.0,7931.0,1.0,0.0,0.0,10.0,0.0,
101.81.229.jeb,2017-06-27,00:00:00,0.0,768835.0,0000950152-00-003270,-index.html,200.0,7809.0,1.0,0.0,0.0,10.0,0.0,
101.81.229.jeb,2017-06-27,00:00:00,0.0,769397.0,0000929624-99-000165,-index.html,200.0,6624.0,1.0,0.0,0.0,10.0,0.0,
101.81.229.jeb,2017-06-27,00:00:00,0.0,769397.0,0000769397-17-000031,-index.htm,200.0,11561.0,1.0,0.0,0.0,10.0,0.0,
101.81.229.jeb,2017-06-27,00:00:00,0.0,354950.0,0000354950-99-000003,-index.html,200.0,7224.0,1.0,0.0,0.0,10.0,0.0,
101.81.229.jeb,2017-06-27,00:00:00,0.0,886744.0,0001095811-01-504221,-index.htm,200.0,6420.0,1.0,0.0,0.0,10.0,0.0,
101.81.229.jeb,2017-06-27,00:00:00,0.0,842183.0,0000950123-94-000909,-index.html,200.0,6356.0,1.0,0.0,0.0,10.0,0.0,
101.81.229.jeb,2017-06-27,00:00:00,0.0,748592.0,0001019687-13-004331,-index.htm,200.0,11234.0,1.0,0.0,0.0,10.0,0.0,
```

PRIMER PRETRAGE

- https://www.sec.gov/search/search.htm



Search Company Filings

To search the SEC database for company filings — including quarterly and annual reports, registration statements for IPOs and other offerings, insider trading reports, and proxy materials — use the search box below. See also our EDGAR Full-Text Search.

Enter your search information.

Company name:
Starts with Contains
or CIK or Ticker Symbol:
Tickers for 10,000 largest publicly traded companies
or File Number:
State:
Country:
and/or SIC:
and Ownership Forms 3, 4, and 5. Include Exclude Only
Find Companies

Helpful Information

Search SEC Documents

To search SEC.gov and Investor.gov for public statements, proposed and final rules, enforcement actions, educational materials, and other documents, enter keywords or phrases below. Your feedback is welcome, contact webmaster@sec.gov. **This search will not retrieve company filings.**

Search

For quick links to our most sought-after pages, please visit Most Common Search Terms and Fast Answers — Key Topics.

Important: As of July 18, 2013, any new press releases, speeches, public statements, and testimonies will not be searchable using this legacy search. To search for these SEC documents, please use our new search engine (shown above).

If you are looking for company filings, you will need to search EDGAR.

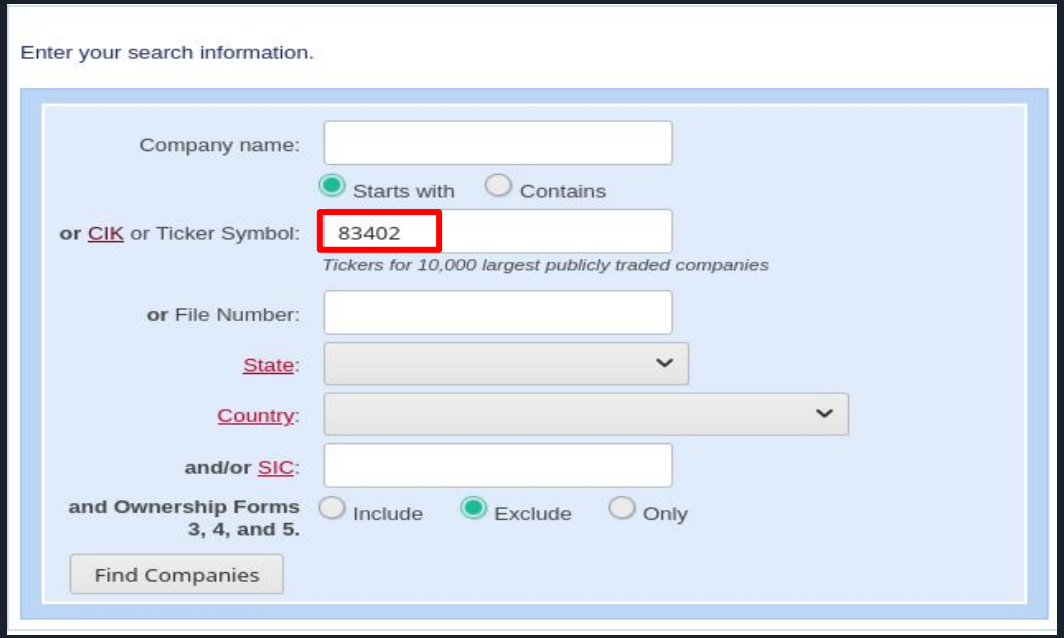
Legacy Search Engine

Search Reset

Advanced Search | Search Help

PRIMER PRETRAGE - UNOS CIK

```
ip,date,time,zone,cik,accession,extention,code,size,idx,norefer,noagent,find,crawler,browser
101.81.229.jeb,2017-06-27,00:00:00,0.0,83402.0,0000083402-97-000006,-index.html,200.0,7492.0,1.0,0.0,0.0,10.0,0.0,
```



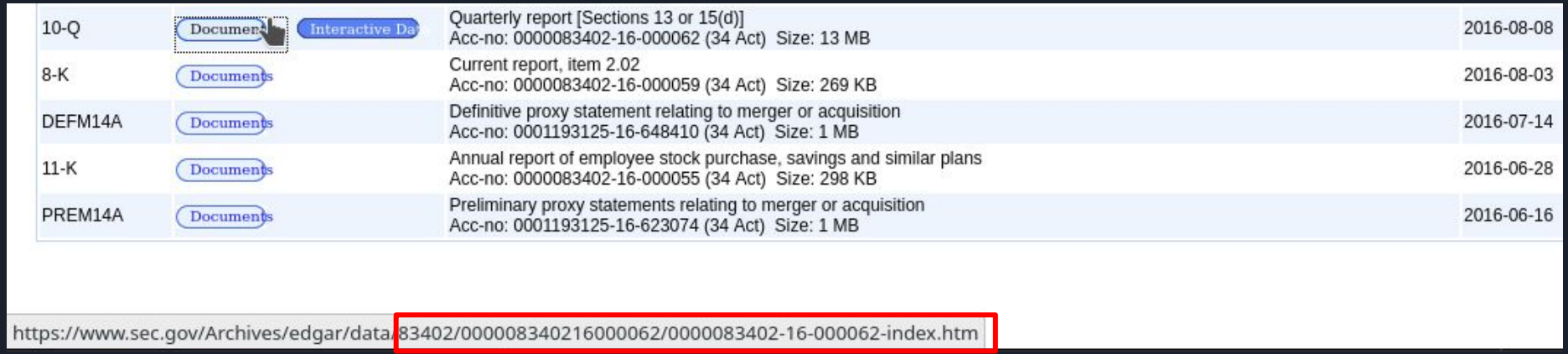
PRIMER PRETRAGE - REZULTAT PRETRAGE

```
ip,date,time,zone,cik,accession,extention,code,size,idx,norefer,noagent,find,crawler,browser
101.81.229.jeb,2017-06-27,00:00:00,0.0,83402.0,0000083402-97-000006,-index.html,200.0,7492.0,1.0,0.0,0.0,10.0,0.0,
```

8-K	Documents		Current report, items 5.07, 8.01, and 9.01 Acc-no: 0001193125-16-695325 (34 Act) Size: 34 KB	2016-08-29
8-K	Documents		Current report, item 8.01 Acc-no: 0001193125-16-680079 (34 Act) Size: 52 KB	2016-08-12
DEFA14A	Documents		Additional definitive proxy soliciting materials and Rule 14(a)(12) material Acc-no: 0001193125-16-680080 (34 Act) Size: 52 KB	2016-08-12
10-Q	Documents	Interactive Da	Quarterly report [Sections 13 or 15(d)] Acc-no: 0000083402-16-000062 (34 Act) Size: 13 MB	2016-08-08
8-K	Documents		Current report, item 2.02 Acc-no: 0000083402-16-000059 (34 Act) Size: 269 KB	2016-08-03
DEFM14A	Documents		Definitive proxy statement relating to merger or acquisition Acc-no: 0001193125-16-648410 (34 Act) Size: 1 MB	2016-07-14
11-K	Documents		Annual report of employee stock purchase, savings and similar plans Acc-no: 0000083402-16-000055 (34 Act) Size: 298 KB	2016-06-28
PREM14A	Documents		Preliminary proxy statements relating to merger or acquisition Acc-no: 0001193125-16-623074 (34 Act) Size: 1 MB	2016-06-16

PRIMER PRETRAGE

```
ip,date,time,zone,cik,accession,extention,code,size,idx,norefer,noagent,find,crawler,browser
101.81.229.jeb,2017-06-27,00:00:00,0.0,83402.0,0000083402-97-000006,-index.html,200.0,7492.0,1.0,0.0,0.0,10.0,0.0,
```



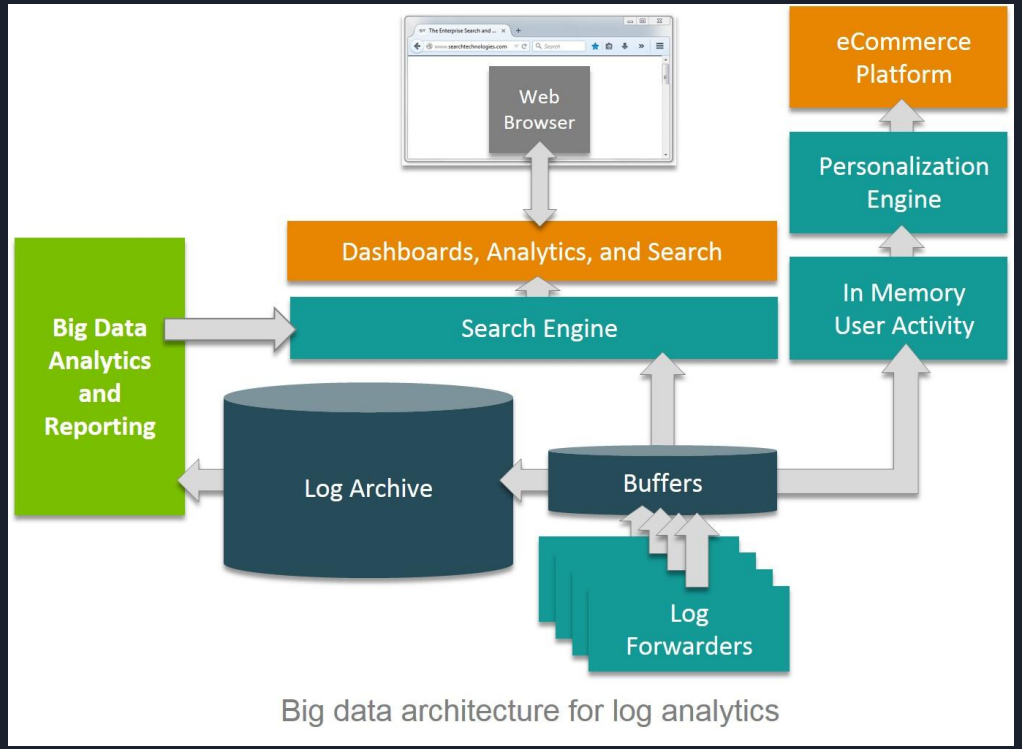
10-Q	Documents Interactive Da	Quarterly report [Sections 13 or 15(d)] Acc-no: 0000083402-16-000062 (34 Act) Size: 13 MB	2016-08-08
8-K	Documents	Current report, item 2.02 Acc-no: 0000083402-16-000059 (34 Act) Size: 269 KB	2016-08-03
DEFM14A	Documents	Definitive proxy statement relating to merger or acquisition Acc-no: 0001193125-16-648410 (34 Act) Size: 1 MB	2016-07-14
11-K	Documents	Annual report of employee stock purchase, savings and similar plans Acc-no: 0000083402-16-000055 (34 Act) Size: 298 KB	2016-06-28
PREM14A	Documents	Preliminary proxy statements relating to merger or acquisition Acc-no: 0001193125-16-623074 (34 Act) Size: 1 MB	2016-06-16

ARHITEKTURA SISTEMA

Enterprise rešenja koriste sledeći stack tehnologija:

- Elasticsearch - brza pretraga logova
- Logstash - parsiranje i skladištenje logova
- Kibana - vizualizacija u browser-u

ARHITEKTURA SISTEMA

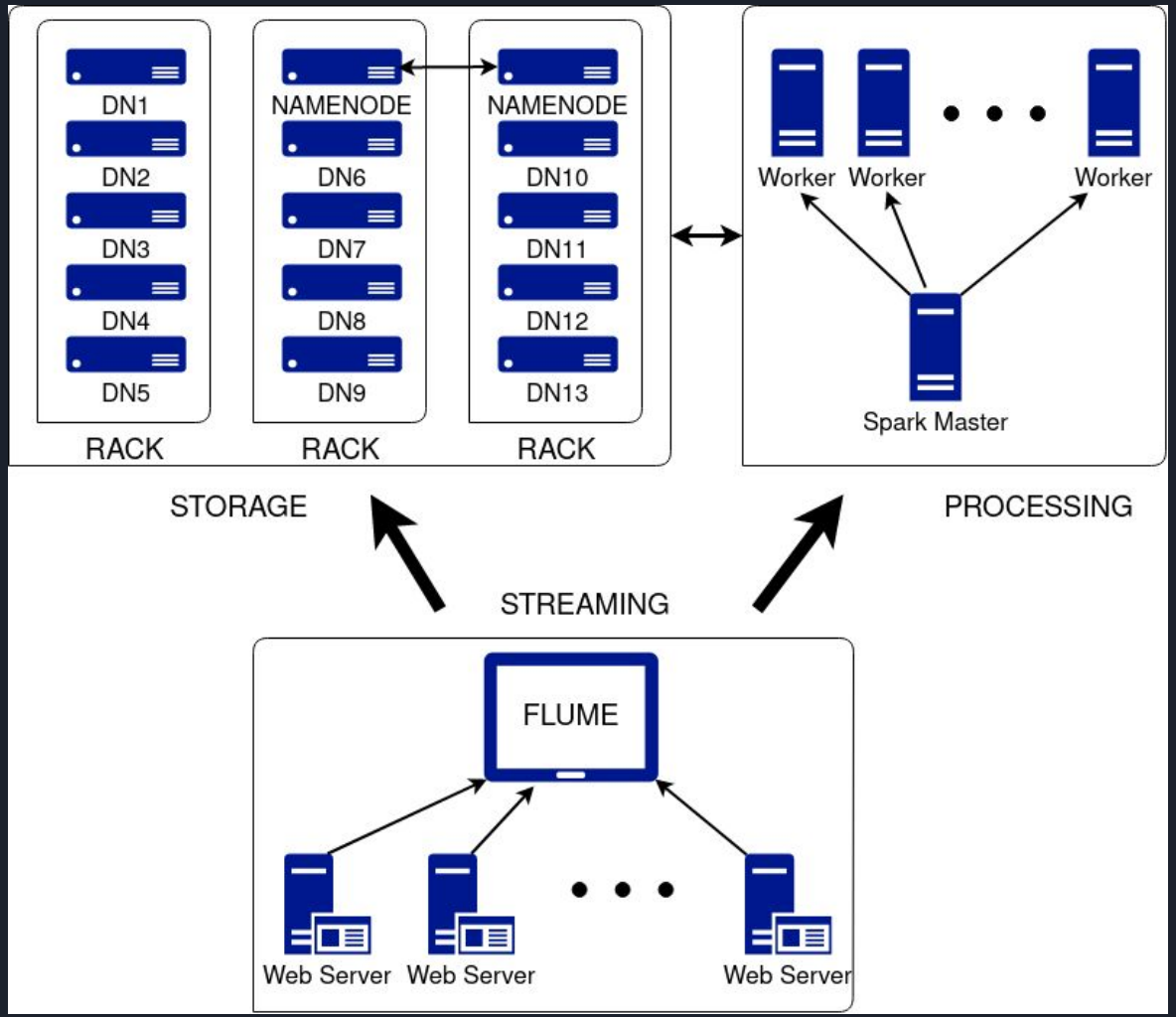


Big data architecture for log analytics

ARHITETKTURA SISTEMA

Open source alternative:

- skladištenje i obrada logova: Apache Flume, Apache Spark, Search Technologies' Aspire
- pretraga logova: Solr, Lucidworks
- vizualizacija: Apache Hue, Pentaho Analytics and Data Integration, HighCharts, D3 Charts


DN1
DN2
DN3
DN4
DN5
RACK
NAMENODE
DN6
DN7
DN8
DN9
RACK
NAMENODE
DN10
DN11
DN12
DN13
RACK
Worker
Worker
Worker
Spark Master
STORAGE
PROCESSING
STREAMING
FLUME
Web Server
Web Server
Web Server

BATCH OBRADA - PREDPROCESIRANJE - SIROVI PODACI

ip	date	time	zone	cik	accession	extension	code	size	idx	noRefer	noAgent	find	crawler	browser
101.81.133.jja	2017-06-29	00:00:00	0.0	1515671.0	0000940400-17-000412	-index.htm	200.0	6832.0	1.0	0.0	0.0	9.0	0.0	null
101.81.133.jja	2017-06-29	00:00:00	0.0	1105685.0	0001209191-17-042148	-index.htm	200.0	9902.0	1.0	0.0	0.0	9.0	0.0	null
101.81.77.ach	2017-06-29	00:00:00	0.0	104894.0	0001193125-10-215678	-index.htm	200.0	7857.0	1.0	0.0	0.0	10.0	0.0	null
107.23.85.jfd	2017-06-29	00:00:00	0.0	1305479.0	0000905148-06-007070	-index.htm	200.0	2944.0	1.0	0.0	0.0	10.0	0.0	null
107.23.85.jfd	2017-06-29	00:00:00	0.0	1336279.0	0001193125-07-254842	-index.htm	200.0	2845.0	1.0	0.0	0.0	10.0	0.0	null
107.23.85.jfd	2017-06-29	00:00:00	0.0	1353349.0	0001056404-06-002693	-index.htm	200.0	2665.0	1.0	0.0	0.0	10.0	0.0	null
107.23.85.jfd	2017-06-29	00:00:00	0.0	1353349.0	0001056404-06-002937	-index.htm	200.0	2665.0	1.0	0.0	0.0	10.0	0.0	null
107.23.85.jfd	2017-06-29	00:00:00	0.0	1336279.0	0001193125-07-214209	-index.htm	200.0	2840.0	1.0	0.0	0.0	10.0	0.0	null
107.23.85.jfd	2017-06-29	00:00:00	0.0	1353349.0	0001056404-06-003388	-index.htm	200.0	2666.0	1.0	0.0	0.0	10.0	0.0	null
107.23.85.jfd	2017-06-29	00:00:00	0.0	1305479.0	0000905148-07-000003	-index.htm	200.0	2940.0	1.0	0.0	0.0	10.0	0.0	null
107.23.85.jfd	2017-06-29	00:00:00	0.0	1353349.0	0001056404-07-001256	-index.htm	200.0	2658.0	1.0	0.0	0.0	10.0	0.0	null
107.23.85.jfd	2017-06-29	00:00:00	0.0	1390505.0	0001144204-07-008989	-index.htm	200.0	2779.0	1.0	0.0	0.0	10.0	0.0	null
107.23.85.jfd	2017-06-29	00:00:00	0.0	1353349.0	0001056404-06-004082	-index.htm	200.0	2668.0	1.0	0.0	0.0	10.0	0.0	null
107.23.85.jfd	2017-06-29	00:00:00	0.0	1353349.0	0001056404-07-000444	-index.htm	200.0	2670.0	1.0	0.0	0.0	10.0	0.0	null
107.23.85.jfd	2017-06-29	00:00:00	0.0	1305479.0	0000905148-07-000083	-index.htm	200.0	2855.0	1.0	0.0	0.0	10.0	0.0	null
107.23.85.jfd	2017-06-29	00:00:00	0.0	1390505.0	0001144204-07-009290	-index.htm	200.0	2713.0	1.0	0.0	0.0	10.0	0.0	null
107.23.85.jfd	2017-06-29	00:00:00	0.0	1336279.0	0001193125-07-257015	-index.htm	200.0	2875.0	1.0	0.0	0.0	10.0	0.0	null
107.23.85.jfd	2017-06-29	00:00:00	0.0	1390505.0	0001144204-07-009292	-index.htm	200.0	2709.0	1.0	0.0	0.0	10.0	0.0	null
107.23.85.jfd	2017-06-29	00:00:00	0.0	1336279.0	0001193125-07-233704	-index.htm	200.0	2830.0	1.0	0.0	0.0	10.0	0.0	null
107.23.85.jfd	2017-06-29	00:00:00	0.0	1336279.0	0001193125-07-193820	-index.htm	200.0	2842.0	1.0	0.0	0.0	10.0	0.0	null

BATCH OBRADA - PREDPROCESIRANJE - TRANSFORMISANI PODACI

ip	cik	accession	extension	code	size	idx	noRefer	noAgent	find	crawler	datetime
101.81.133.jja	1515671.0	0000940400-17-000412	-index.htm	200	6832	1	0	0	9	0	2017-06-29 00:00:00
101.81.133.jja	1105685.0	0001209191-17-042148	-index.htm	200	9902	1	0	0	9	0	2017-06-29 00:00:00
101.81.77.ach	104894.0	0001193125-10-215678	-index.htm	200	7857	1	0	0	10	0	2017-06-29 00:00:00
107.23.85.jfd	1305479.0	0000905148-06-007070	-index.htm	200	2944	1	0	0	10	0	2017-06-29 00:00:00
107.23.85.jfd	1336279.0	0001193125-07-254842	-index.htm	200	2845	1	0	0	10	0	2017-06-29 00:00:00
107.23.85.jfd	1353349.0	0001056404-06-002693	-index.htm	200	2665	1	0	0	10	0	2017-06-29 00:00:00
107.23.85.jfd	1353349.0	0001056404-06-002937	-index.htm	200	2665	1	0	0	10	0	2017-06-29 00:00:00
107.23.85.jfd	1336279.0	0001193125-07-214209	-index.htm	200	2840	1	0	0	10	0	2017-06-29 00:00:00
107.23.85.jfd	1353349.0	0001056404-06-003388	-index.htm	200	2666	1	0	0	10	0	2017-06-29 00:00:00
107.23.85.jfd	1305479.0	0000905148-07-000003	-index.htm	200	2940	1	0	0	10	0	2017-06-29 00:00:00
107.23.85.jfd	1353349.0	0001056404-07-001256	-index.htm	200	2658	1	0	0	10	0	2017-06-29 00:00:00
107.23.85.jfd	1390505.0	0001144204-07-008989	-index.htm	200	2779	1	0	0	10	0	2017-06-29 00:00:00
107.23.85.jfd	1353349.0	0001056404-06-004082	-index.htm	200	2668	1	0	0	10	0	2017-06-29 00:00:00
107.23.85.jfd	1353349.0	0001056404-07-000444	-index.htm	200	2670	1	0	0	10	0	2017-06-29 00:00:00
107.23.85.jfd	1305479.0	0000905148-07-000083	-index.htm	200	2855	1	0	0	10	0	2017-06-29 00:00:00
107.23.85.jfd	1390505.0	0001144204-07-009290	-index.htm	200	2713	1	0	0	10	0	2017-06-29 00:00:00
107.23.85.jfd	1336279.0	0001193125-07-257015	-index.htm	200	2875	1	0	0	10	0	2017-06-29 00:00:00
107.23.85.jfd	1390505.0	0001144204-07-009292	-index.htm	200	2709	1	0	0	10	0	2017-06-29 00:00:00
107.23.85.jfd	1336279.0	0001193125-07-233704	-index.htm	200	2830	1	0	0	10	0	2017-06-29 00:00:00
107.23.85.jfd	1336279.0	0001193125-07-193820	-index.htm	200	2842	1	0	0	10	0	2017-06-29 00:00:00

BATCH OBRADA

- *Batch* obrda je implementirana nešto nalik *Command* softverskog obrasca.
- Komanda treba da procesira podatke na određeni način i da izgeneriše izveštaj.
- Ovakav dizajn omogućava fleksibilan i lagan način za dodovanje novih komadni koji će generisati izveštaje.

BATCH OBRADA

- Generisanje opšteg izveštaja koji sadrži:
  - ukupan broj zahteva
  - broj uspešnih zahteva
  - broj neuspešnih zahteva
  - broj jedinstvenih posetilaca (ista IP adresa i *timestamp* zahteva, kada bi postojale informacije o *web browser*-u uključili bismo i to)
  - broj neuspešnih zahteva sa statusom 404
  - broj zahteva koji sadrže Referrer polje
  - ukupan protok saobraćaja izražen u GB

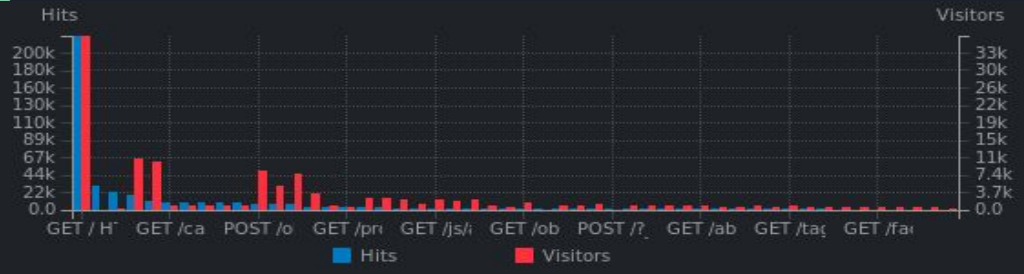
BATCH OBRADA

#	Hits	Visitors	Tx. Amount	Avg. T.S.	Data
	1,351,664 Max: 5,740 Min: 163	278,357 Max: 813 Min: 48	45.35 GiB Max: 864.88 MiB Min: 1.15 MiB	519.89 ms	2,002 Total
1	292 (0.02%)	65 (0.02%)	2.99 MiB (0.01%)	145.26 ms	17/Nov/2019:16
2	641 (0.05%)	143 (0.05%)	8.89 MiB (0.02%)	224.27 ms	17/Nov/2019:15
3	628 (0.05%)	143 (0.05%)	4.92 MiB (0.01%)	216.91 ms	17/Nov/2019:14
4	547 (0.04%)	152 (0.05%)	4.64 MiB (0.01%)	251.42 ms	17/Nov/2019:13
5	723 (0.05%)	138 (0.05%)	6.41 MiB (0.01%)	164.09 ms	17/Nov/2019:12
6	806 (0.06%)	129 (0.05%)	7.8 MiB (0.02%)	155.73 ms	17/Nov/2019:11
7	677 (0.05%)	129 (0.05%)	8.18 MiB (0.02%)	183.47 ms	17/Nov/2019:10

Izveštaj obuhvata:

- broj posetilaca, uključujući web crawler-e, (**Visitors**) i zahteva (**Hits**) koji su se desili u roku od sat vremena
- ukupan protok saobraćaja (**Tx. Amount**) kao i prosečno vreme izvršavanja zahteva (**Avg T.S.**) grupisani na nivou sata

BATCH OBRADA

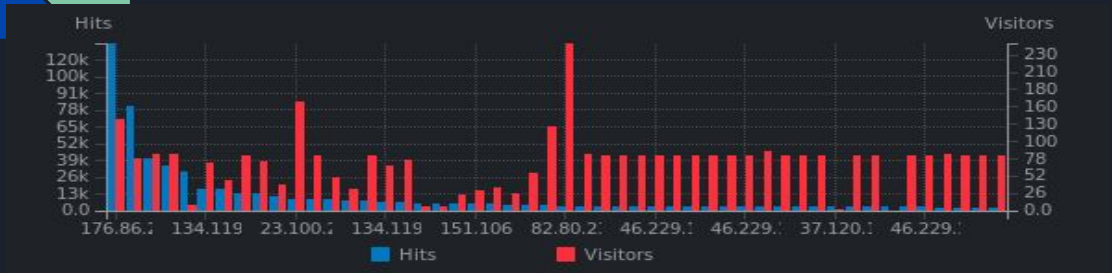


#	Hits	Visitors	Tx. Amount	Avg. T.S.	Data
	874,044 Max: 222,199 Min: 1	510,884 Max: 37,176 Min: 1	3.18 GiB Max: 680.95 MiB Min: 0 Byte	252.04 ms	141,174 Total
1	222,199 (25.42%)	37,176 (7.28%)	680.95 MiB (20.89%)	9.84 ms	/
2	30,892 (3.53%)	8 (0.00%)	3.24 MiB (0.10%)	0.00 us	bind.sh firewall-falcon quetz
3	22,499 (2.57%)	402 (0.08%)	7.66 MiB (0.24%)	1.31 ms	/
4	19,904 (2.28%)	11,000 (2.15%)	63.3 MiB (1.94%)	48.23 ms	/obituaries
5	11,752 (1.34%)	10,266 (2.01%)	148.58 MiB (4.56%)	85.17 ms	/obituaries/439/elaine-k-bra
6	9,120 (1.04%)	919 (0.18%)	55.74 MiB (1.71%)	13.55 ms	/cart
7	9,025 (1.03%)	895 (0.18%)	56.73 MiB (1.74%)	19.19 ms	/products/gerberas

Izveštaj obuhvata:

- broj posetilaca (**Visitors**) i zahteva (**Hits**) grupisani po *SEC Central Index Key* polju
- ukupan protok saobraćaja (**Tx. Amount**) kao i prosečno vreme izvršavanja zahteva (**Avg T.S.**) grupisani *SEC Central Index Key* polju

BATCH OBRADA

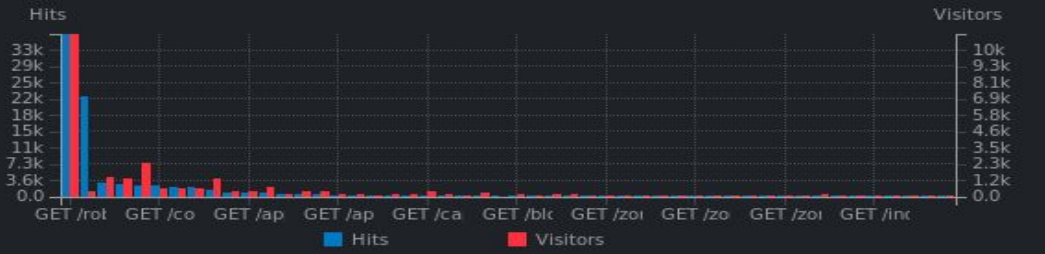


#	Hits	Visitors	Tx. Amount	Avg. T.S.	Data
	1,352,148 Max: 130,526 Min: 1	167,155 Max: 329 Min: 1	45.36 GiB Max: 664.92 MiB Min: 0 Byte	519.76 ms	55,870 Total
1	130,526 (9.65%)	137 (0.08%)	491.61 MiB (1.06%)	1.33 s	176.86.241.109
2	81,966 (6.06%)	78 (0.05%)	222.13 MiB (0.48%)	7.05 ms	69.162.124.233
3	41,294 (3.05%)	85 (0.05%)	153.48 MiB (0.33%)	9.32 ms	134.119.218.243
4	35,011 (2.59%)	86 (0.05%)	136.28 MiB (0.29%)	9.32 ms	134.119.216.167
5	30,905 (2.29%)	8 (0.00%)	3.24 MiB (0.01%)	0.00 us	::1
6	16,753 (1.24%)	72 (0.04%)	61.73 MiB (0.13%)	9.16 ms	134.119.219.93
7	16,586 (1.23%)	45 (0.03%)	62.53 MiB (0.13%)	9.40 ms	134.119.193.63

Izveštaj obuhvata:

- broj posetilaca (**Visitors**) i zahteva (**Hits**) grupisani po IP adresama
- ukupan protok saobraćaja (**Tx. Amount**) kao i prosečno vreme izvršavanja zahteva (**Avg T.S.**) grupisani po IP adresama

BATCH OBRADA



#	Hits	Visitors	Tx. Amount	Avg. T.S.	Data
	135,727 Max: 36,269 Min: 1	66,582 Max: 11,570 Min: 1	448.19 MiB Max: 83.99 MiB Min: 158 B	13.97 ms	8,984 Total
1	36,269 (26.72%)	11,570 (17.38%)	83.99 MiB (18.74%)	3.69 ms	/robots.txt
2	22,414 (16.51%)	430 (0.65%)	80.41 MiB (17.94%)	10.59 ms	/
3	3,254 (2.40%)	1,431 (2.15%)	3.58 MiB (0.80%)	1.62 ms	/apple-touch-icon.png
4	2,889 (2.13%)	1,293 (1.94%)	3 MiB (0.67%)	1.60 ms	/apple-touch-icon-precompose
5	2,443 (1.80%)	2,441 (3.67%)	25.75 MiB (5.74%)	10.75 ms	/blog/10/aloe-vera
6	2,395 (1.76%)	643 (0.97%)	13.84 MiB (3.09%)	10.53 ms	/contactenos
7	2,357 (1.74%)	617 (0.93%)	13.72 MiB (3.06%)	11.00 ms	/login:procesar

Izveštaj obuhvata:

- broj posetilaca,, (**Visitors**) i zahteva (**Hits**) grupisani koji su dobili 404 status
- ukupan protok saobraćaja (**Tx. Amount**)
  kao i prosečno vreme izvršavanja zahteva (**Avg T.S.**) koji imaju status 404

BATCH OBRADA

#	Hits	Visitors	Tx. Amount	Avg. T.S.	Data
	1,352,466 Max: 872,670 Min: 8	222,485 Max: 127,674 Min: 8	45.37 GiB Max: 32.74 GiB Min: 0 Byte	519.69 ms	13 Total
1	917,866 (67.87%)	138,151 (62.09%)	44.7 GiB (98.52%)	761.01 ms	2xx Success
2	278,763 (20.61%)	48,088 (21.61%)	195.45 MiB (0.42%)	3.32 ms	3xx Redirection
3	138,770 (10.26%)	32,132 (14.44%)	450.61 MiB (0.97%)	13.70 ms	4xx Client Errors
4	17,067 (1.26%)	4,114 (1.85%)	39.84 MiB (0.09%)	89.61 ms	5xx Server Errors

Izveštaj obuhvata:

- broj posetilaca,, (**Visitors**) i zahteva (**Hits**) grupisani po statusu
- ukupan protok saobraćaja (**Tx. Amount**) kao i prosečno vreme izvršavanja zahteva (**Avg T.S.**) grupisani po statusu

STREAM OBRADA

- *Stream* obrada radi analizu prispelih logova u realnom vremenu.
- Analiza logova se vrši pomoću pravila koja se mogu lako i fleksibilno dodavati po potrebi.
- Određena pravila se okidaju u pojedinim situacijama i kreiraju alarm sa porukom o uzroku nastanka
- Primer:
  - Pravilo koje detektuje *DoS* napad
  - Pravilo koje detektuje *Brute-force* napad

STREAM OBRADA

- Pojedina pravila nemaju uslov za njihovo okidanje i takva previla predstavljaju izveštaje koji se generišu u realnom vremenu.
- Izvešteji su identični kao u *batch* obradi.