Medius®
With passion.

Enostavna gradnja podatkovnih tokov za usmerjanje, analizo ter učenje v “BigData” svetu

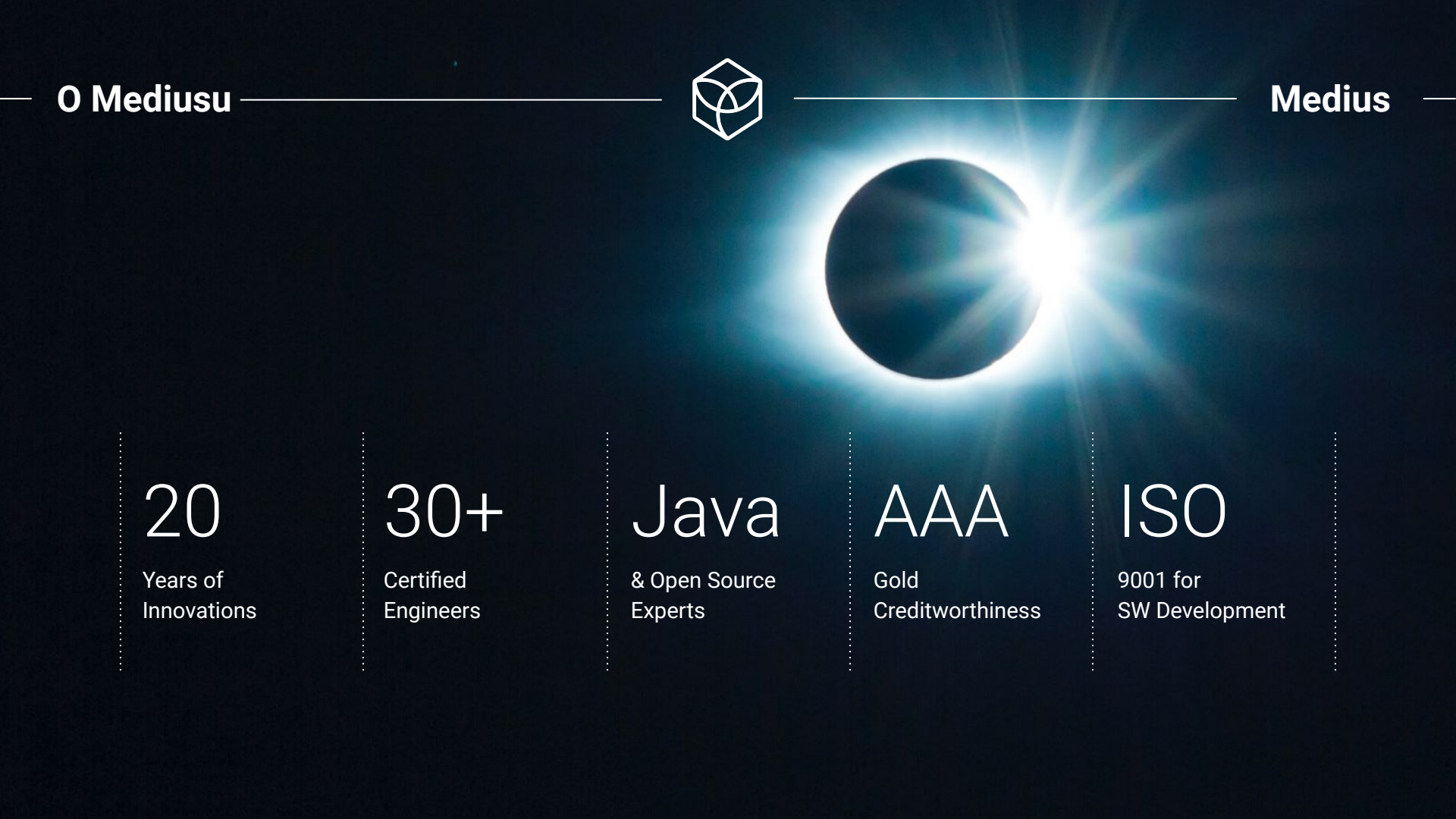
MakeIT 2022, 3.6.2022

David Šenica

- Podatkovni znanstvenik
- Inženir za strojno učenje

Marko Polak

- Arhitekt
- Vodja projektov



20
Years of
Innovations

30+
Certified
Engineers

Java
& Open Source
Experts

AAA
Gold
Creditworthiness

ISO
9001 for
SW Development

Internacionalne nagrade

Medius

United Nations Public Service Award

TM Forum Excellence Award

EuroCloud Best Cloud Service by Startup

EUREKA-ITEA Achievement Award

Best Blockchain Startup at CV Competition

Uvod - agenda

- Big Data in strojno učenje
- Kontekstno odvisni podatkovni tokovi
- Opis demo primera Spletne trgovine in komponente sistema
- Live Demo

Big Data

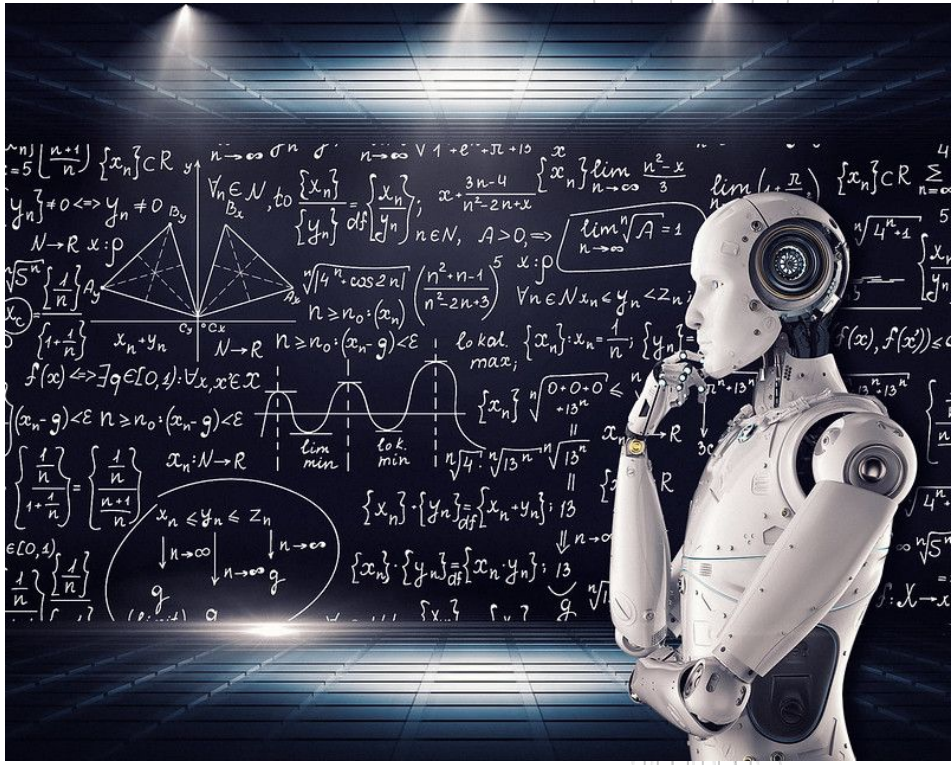
(Velika količina podatkov)

- Tradicionalno: velika količina podatkov (**pre**velika)
- Izzivi
  - zajem
  - shranjevanje
  - analiza
  - iskanje
  - deljenje
  - vizualizacija
  - …

- Bolj moderno:
  - prediktivna analiza
  - ugotavljanje vedenjskih vzorcev uporabnikov(UBA - User behavior analytics)
  - ali kakršnakoli napredna analiza, ki prinaša dodano vrednost iz podatkov

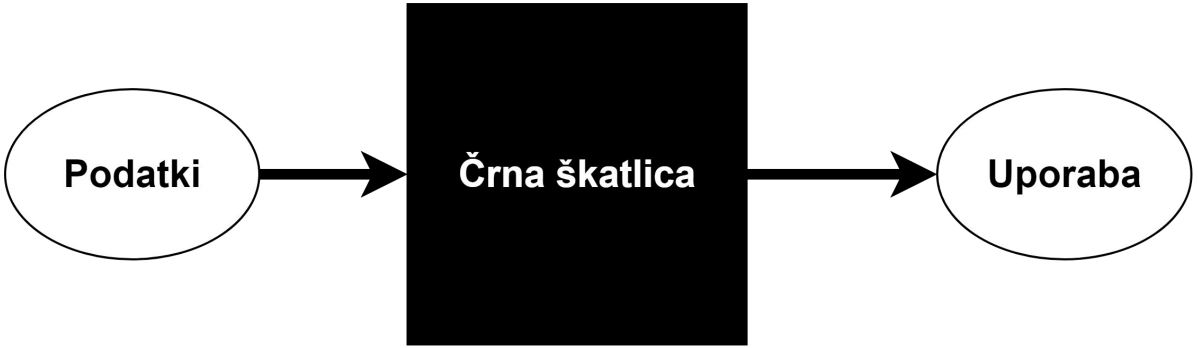
Big Data in strojno učenje - uporabnost

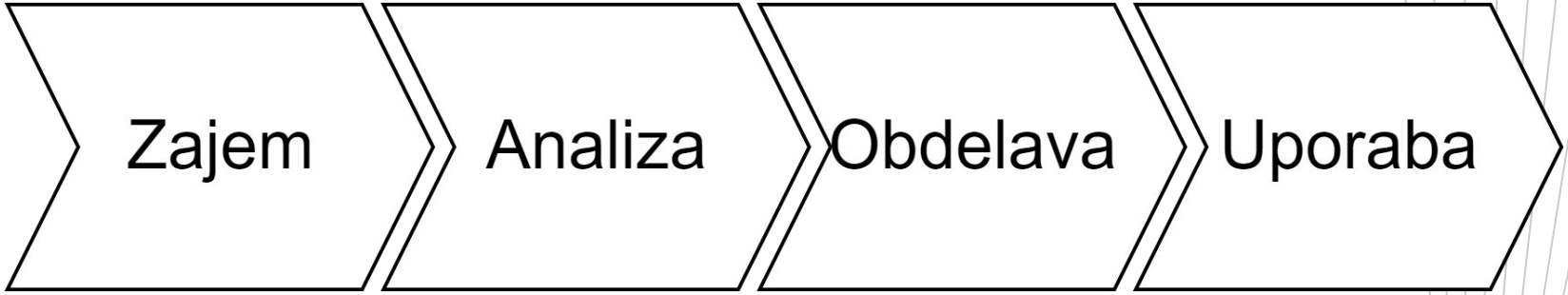
- reševanje problemov
  - napovedi odpovedi
  - detekcija prevar
  - …
- podpora našim poslovnim procesom
  - segmentacija kupcev
  - sentimentalna analiza
  - …

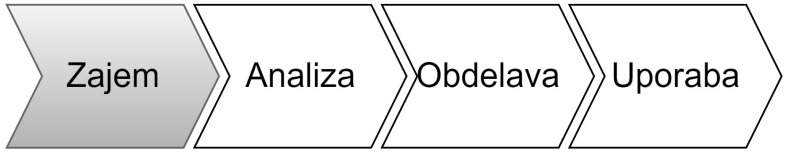


Big Data in strojno učenje - uporabnost

- optimizacija poslovanja
- hitrejše in boljše strateško odločanje
- personalizacija uporabniške izkušnje
- dvig vrednosti podjetja (revenue)
- itd.

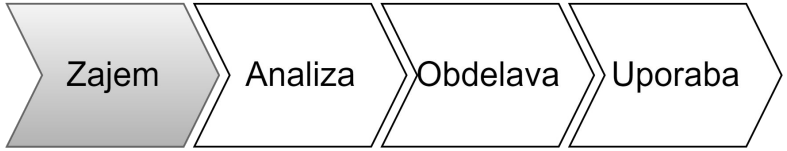



Zajem
Analiza
Obdelava
Uporaba

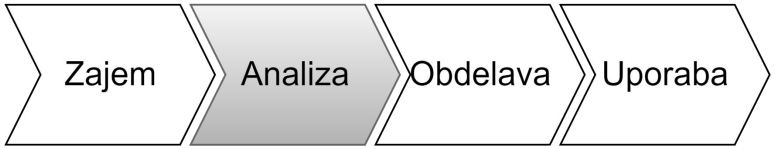


- statični podatki v relacijskih bazah/DWH/NoSQL baze
- komunikacija med aplikacijami(ESB, apache proxy, Istio,...)
- js za sledenje uporabnikom na spletni strani
- namensko spremljanje predefiniranih dogodkov in pošiljanje na zaledni sistem(vgraditev v aplikacije)
- shranjevanje v datoteko
- …

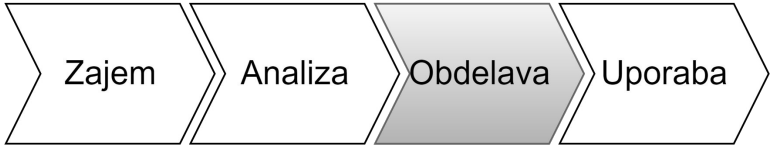
Big Data - zajem podatkov - protokol



- standard out
- datoteka
- rest/soap
- namenski vmesnik
- …

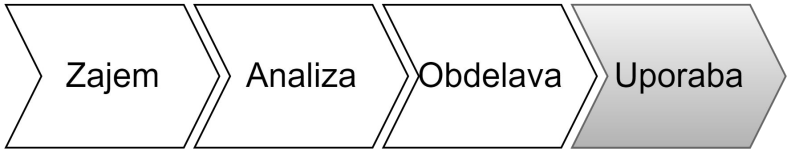


- označevanje(tagging)
- klasifikacija
- kategorizacija
- kontekstualizacija
- …



- standardizacija
- transformacija
- povezovanje nestrukturiranih podatkov in strukturiranih podatkov
- učenje in priprava modelov
- …

Big Data - izpostavljanje rezultatov



- shranjevanje podatkov v bazo
- paketne obdelave(v nočnem času)
- …
- namenski servis (predikcijski, evalvacijski,...)
  - vsem dostopni
  - evalvacija/predikcija v realnem času

Sedaj smo pripravljeni, da se podamo v Big Data svet in v uporabo strojnega učenja :D

A pa smo res?

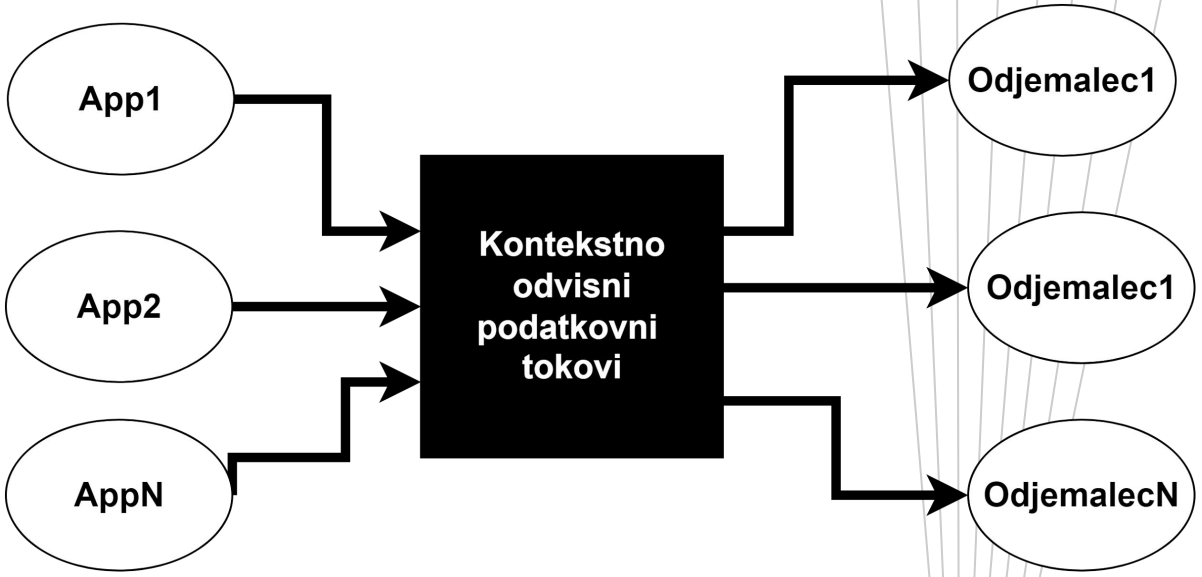
Gartner predvideva, da bo 85% projektov strojnega učenja propadlo.

Nočemo biti na tem vlaku!

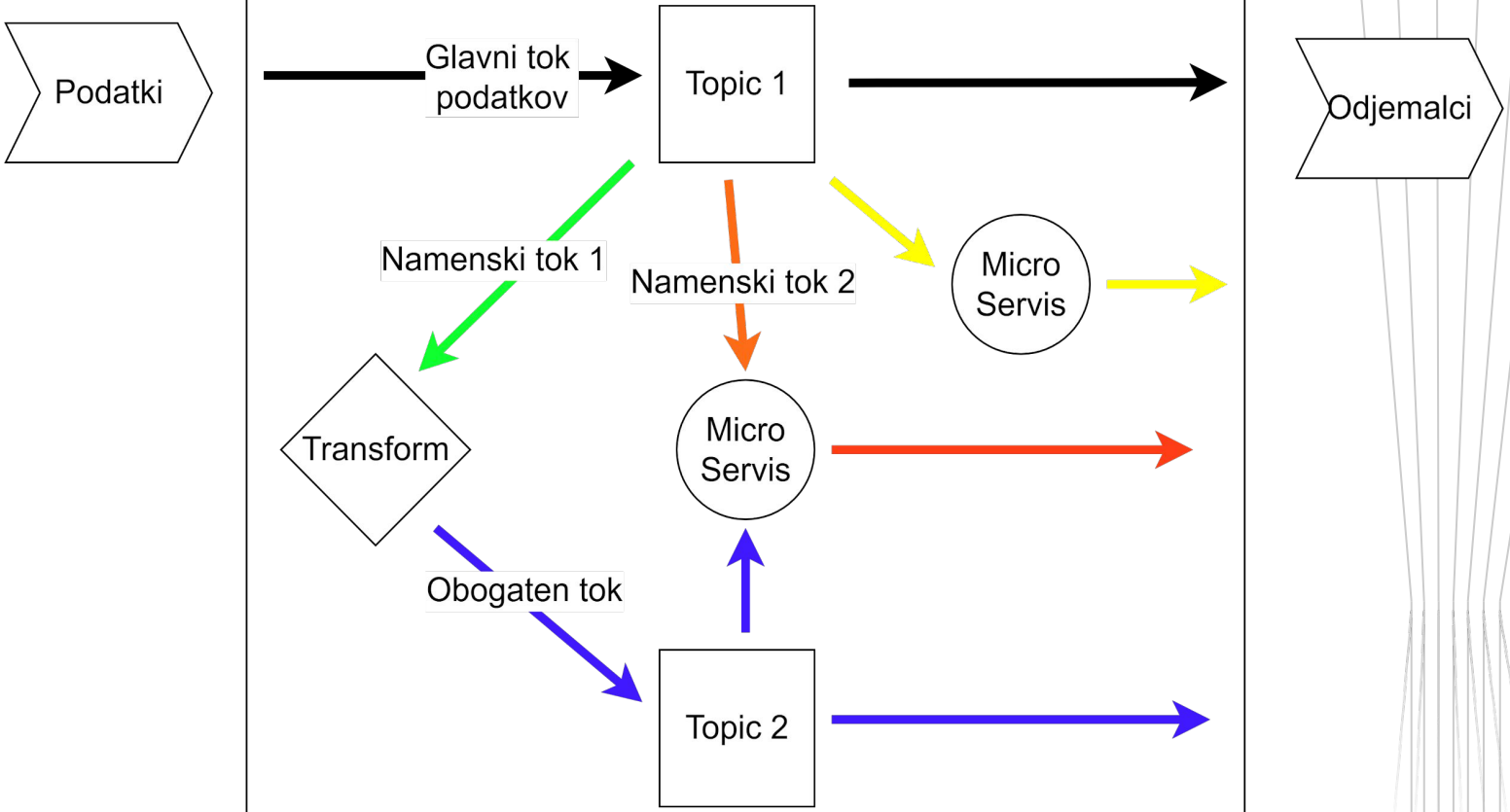
Kontekstno odvisni tokovi podatkov

Kontekstno odvisno tokovi podatkov

- Zajem podatkov s čim večih virov
- Uporabnost podatkov za čim več odjemalcev



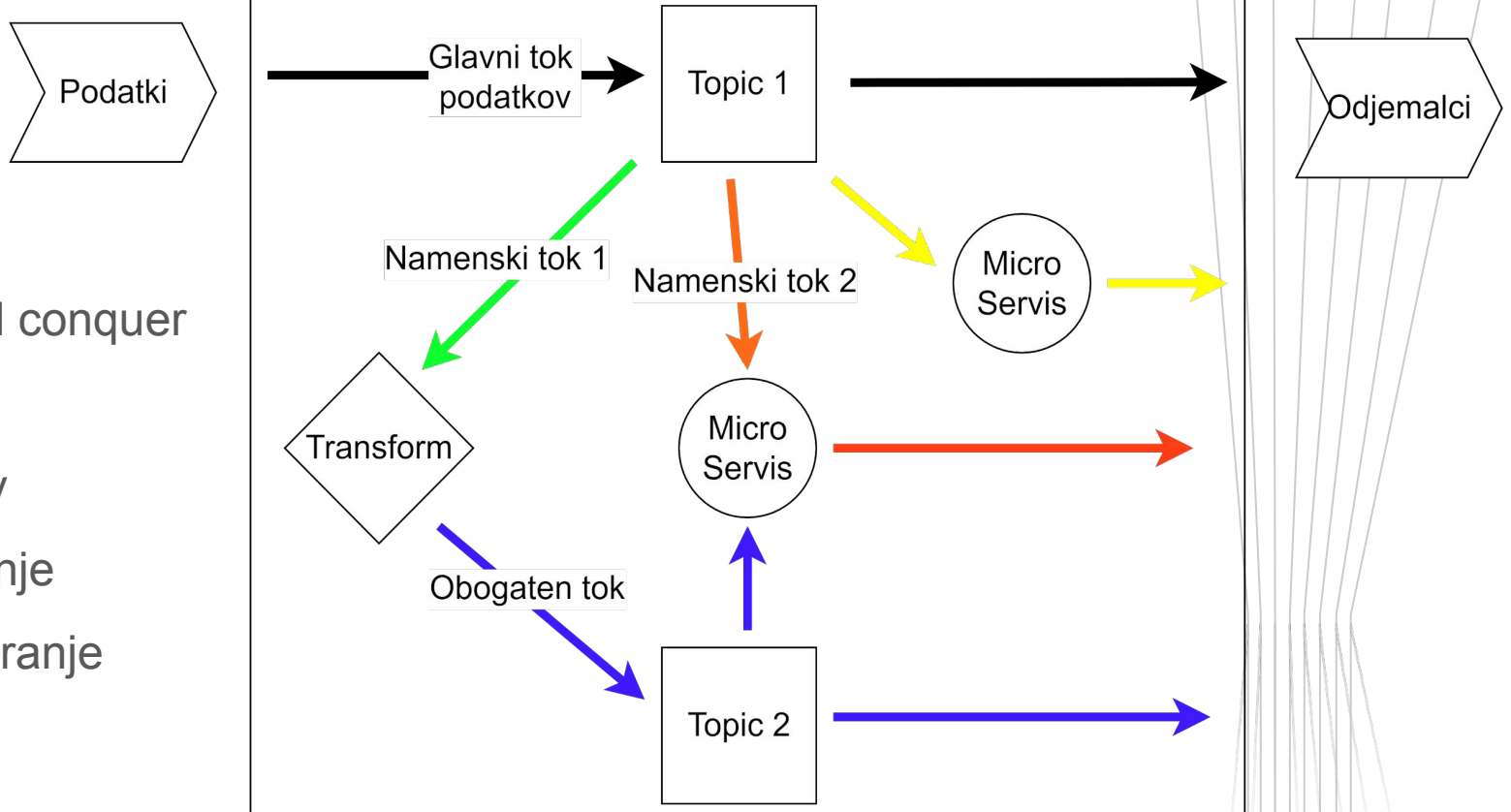
Kontekstno odvisno tokovi podatkov



Kontekstno odvisno tokovi podatkov

Medius

- divide and conquer
- analiza
- obogatitev
- povezovanje
- transformiranje
- …

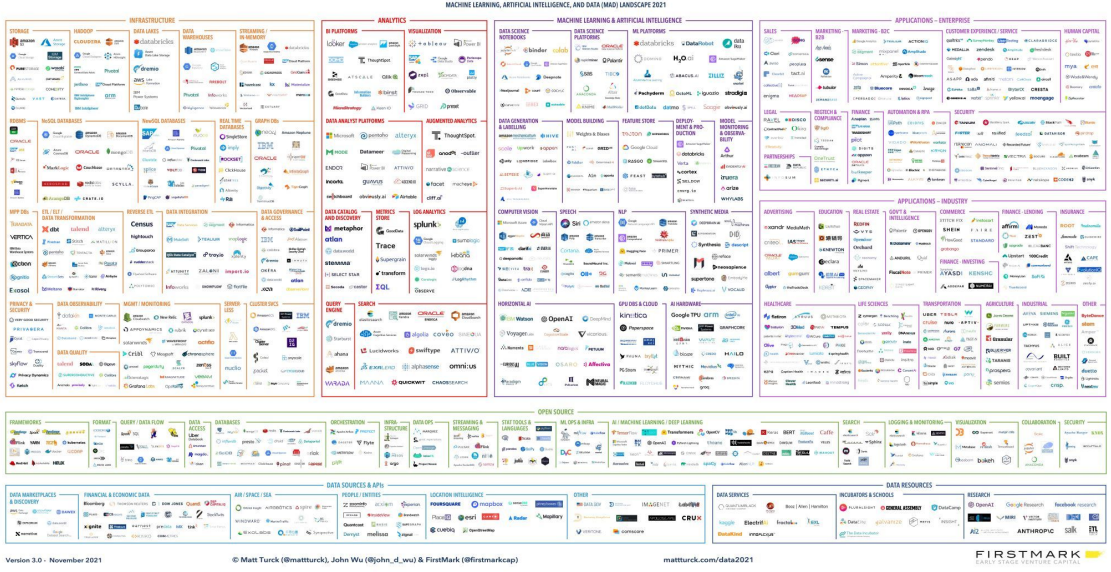


Zakaj?

- ker je bistveno, da damo podatkom uporabno vrednost (z ali brez strojnega učenja)
- da imamo podatke pripravljene za trenutek, ko jih bomo rabili - podatki so osnova za izvajanje strojnega učenja
- da imamo možnost obogatiti podatke - glede na kontekst
    - kontekst podatkov
    - kontekst uporabe (kaj želimo)

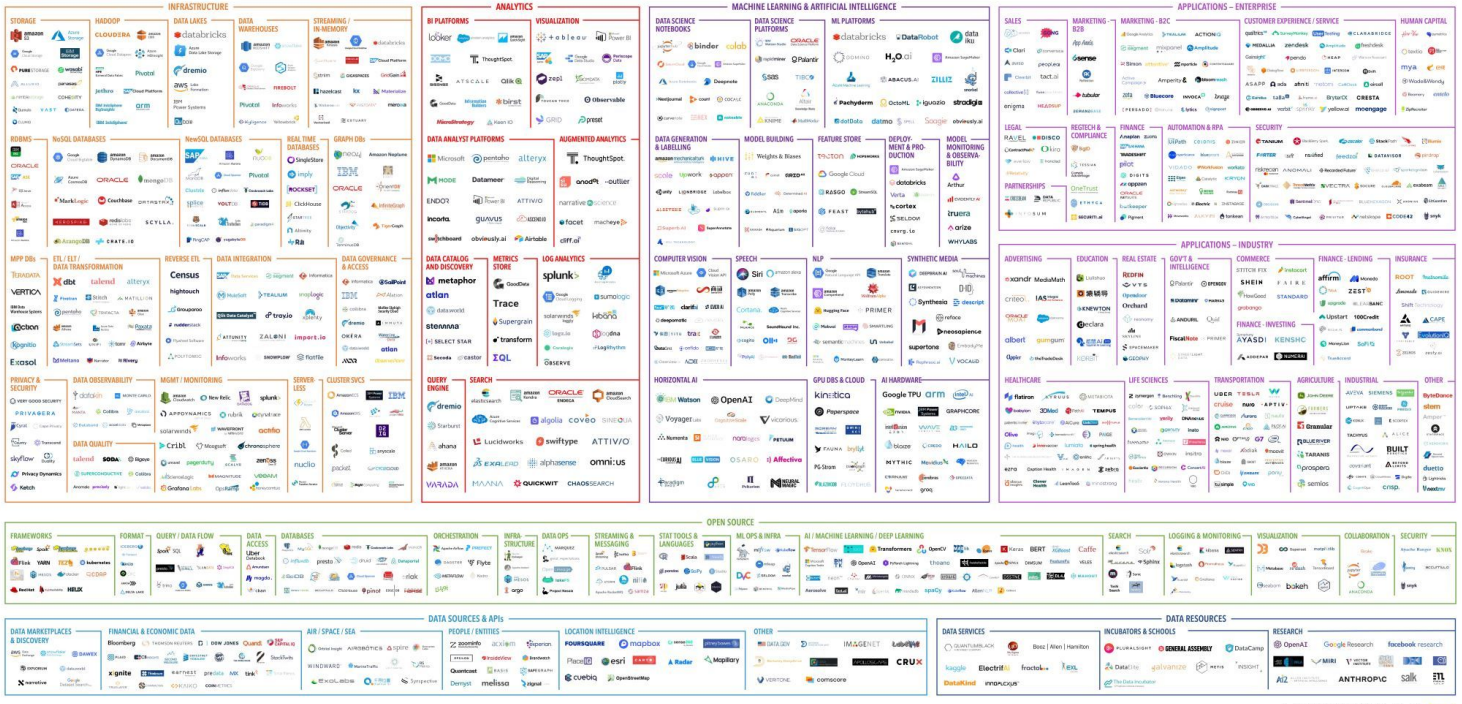
Kontekstno odvisni tokovi podatkov - komponente sistema

- Katere komponente izbrati?
- Kaj upoštevati pri izbiri?
- Naj bo zabavno in enostavno :D



Kontekstno odvisni tokovi podatkov - komponente sistema

Medius



Enostaven primer tokov v spletni trgovini

- Spletna trgovina - nakupovanje
  - želimo uporabnikom ponuditi boljšo uporabniško izkušnjo s tem, da jim ponudimo izdelke, ki bi si jih želeli kupiti
  - podpreti želimo obstoječo komunikacijo s pripravo naročila in dostavo

Kontekstno odvisno tokovi podatkov - zahteve/želje

- enostavnost zajema podatkov
- veliko virov
- visoka prepustnost
- skalabilnost
- veliko odjemalcev
- možnost obogatitve
- NRT procesirnje
- hiter odziv na izbrane podatke/rezultate obdelav

Kontekstno odvisno tokovi podatkov - uporabnost

- iz ogromne količine podatkov vzamemo samo pomembne za naš problem(kontekst)
- tokovi lahko nadomestijo trenutno komunikacijo med aplikacijami
- če načrtujemo nove aplikacije - lahko arhitekturo enostavno prilagodimo tako, da upošteva ta koncept
- uporabna vrednost je takojšnja - ne šele po končanju ML projekta, ki lahko traja več mesecev

Kontekstno odvisno tokovi podatkov - komponente

- Kafka
- Kafka streams
- Quarkus
- Apache Spark
- Elastic
- Logstash
- Kibana

kafka

kibana

Logstash

QUARKUS

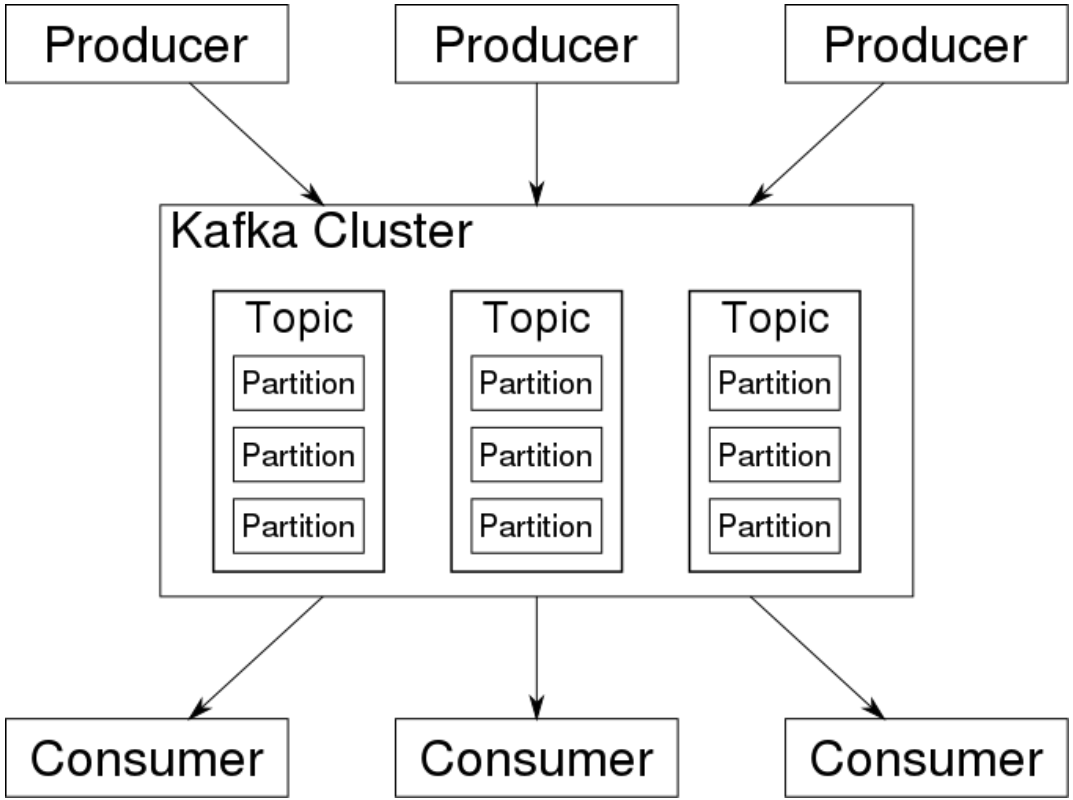
elastic

APACHE Spark™

Kontekstno odvisni tokovi podatkov - Kafka

Kontekstno odvisni tokovi podatkov - Kafka

- distribuirana platforma za pretok dogodkov
- visoka prepustnost
- skalabilno (stopnjevalni sistem - scalable system)
- trdoživo hranjenje podatkov
- visok nivo dostopnosti



Kontekstno odvisni tokovi podatkov - Kafka streams

Kontekstno odvisni tokovi podatkov - Kafka Streams

- javanska knjižnica za analiziranje in procesiranje podatkov shranjenih v kafka topicu
- enostavno delo s tokovi podatkov
- domač java API
- procesiranje zaporednih dogodkov
- windowing
- join-i
- agregacije

Kontekstno odvisni tokovi podatkov - Quarkus

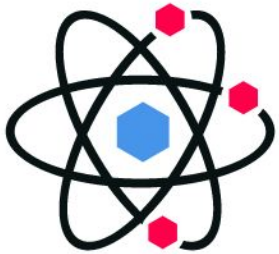
- programsko ogrodje za razvijanje Java aplikacij(Cloud Native, Container First)

What Makes Quarkus Different?

Developer Joy

Kubernetes-native

Best of Breed Libraries and Standards

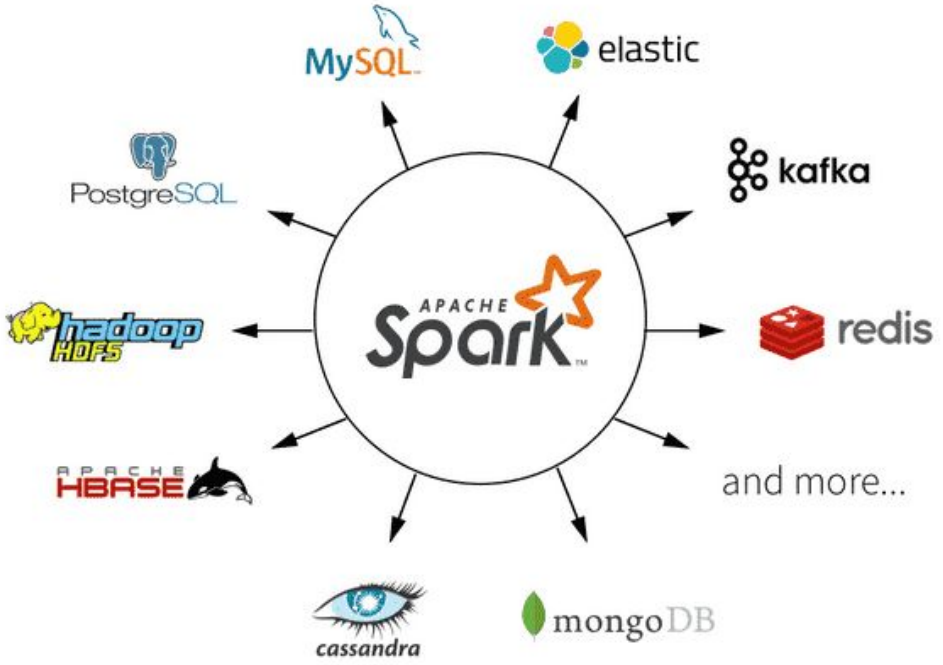


Imperative and reactive code

Kontekstno odvisni tokovi podatkov - Apache Spark

Kontekstno odvisni tokovi podatkov - Apache Spark

- analiza podatkov in strojno učenje na tokovih
- porazdeljeno obdelovanje
- SQL analiza



Kontekstno odvisni tokovi podatkov - Elastic

- Distribuirana zastonjska iskalna in analitična platforma.

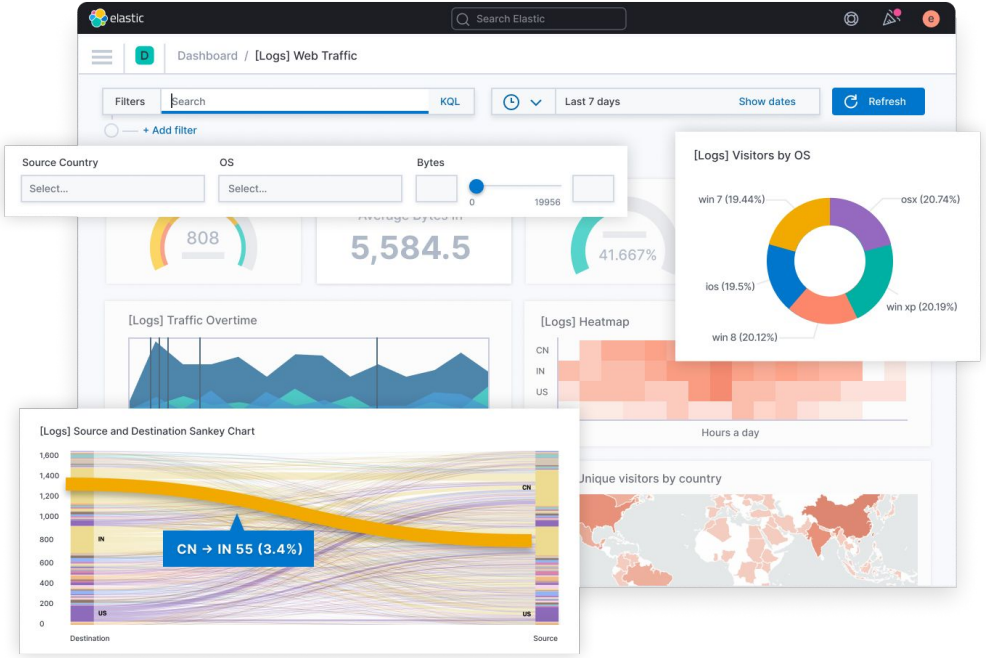
  Namenjena za uporabo s katerimkoli tipom podatkov(tekstovni, numerični, geoprostorskim, strukturiranim in nestrukturiranim).

Kontekstno odvisni tokovi podatkov - Kibana

Kontekstno odvisni tokovi podatkov - Kibana

Medius

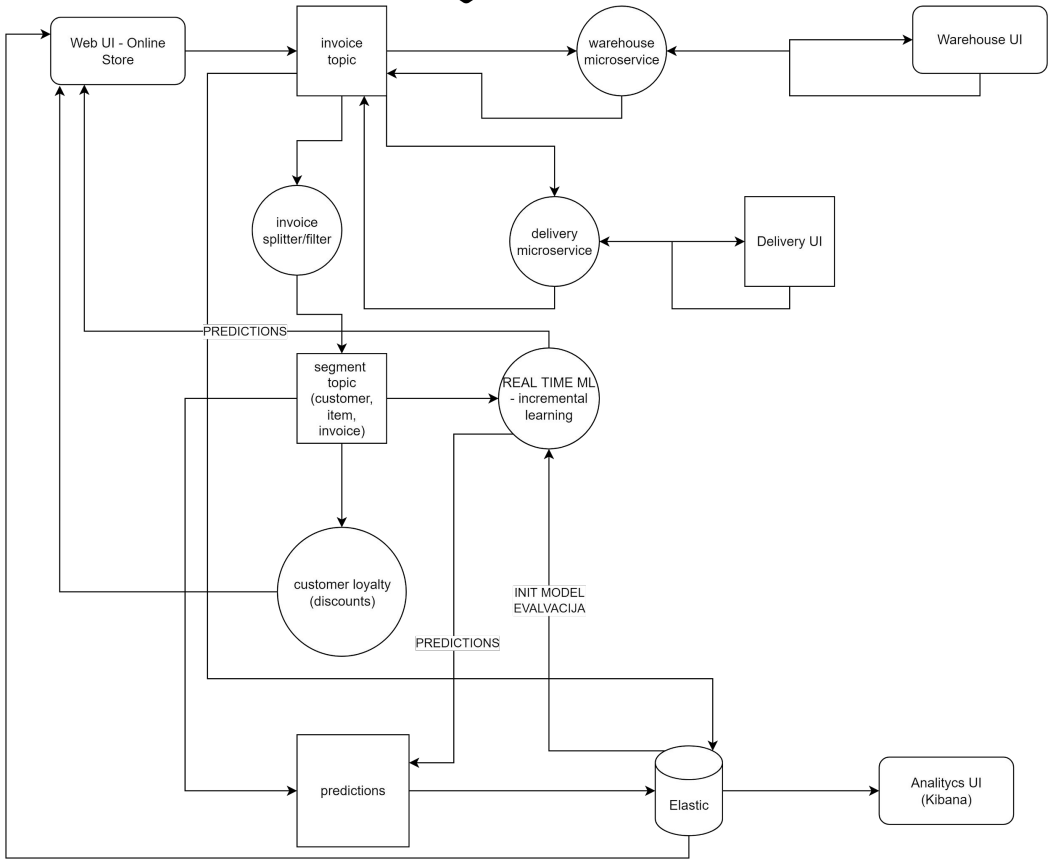
Zastonjska odprtokodna aplikacija za iskanje in vizualizacijo



Enostaven primer tokov v spletni trgovini - arhitektura

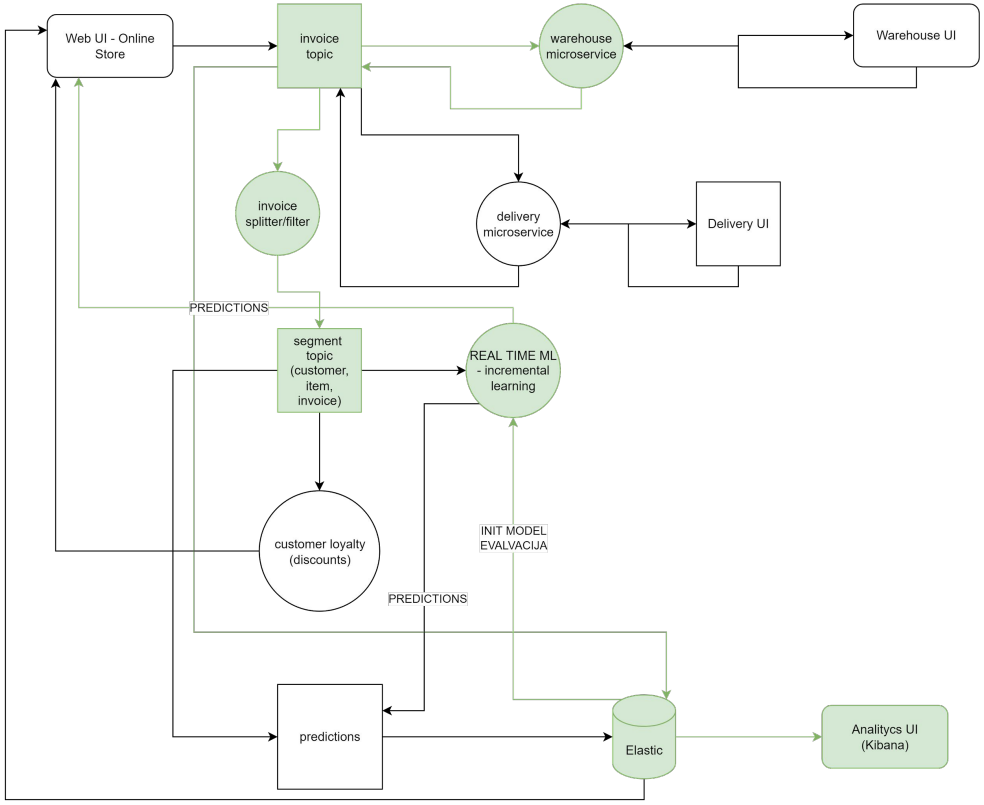
Spletna trgovina - arhitektura

Medius



Spletna trgovina - arhitektura demo

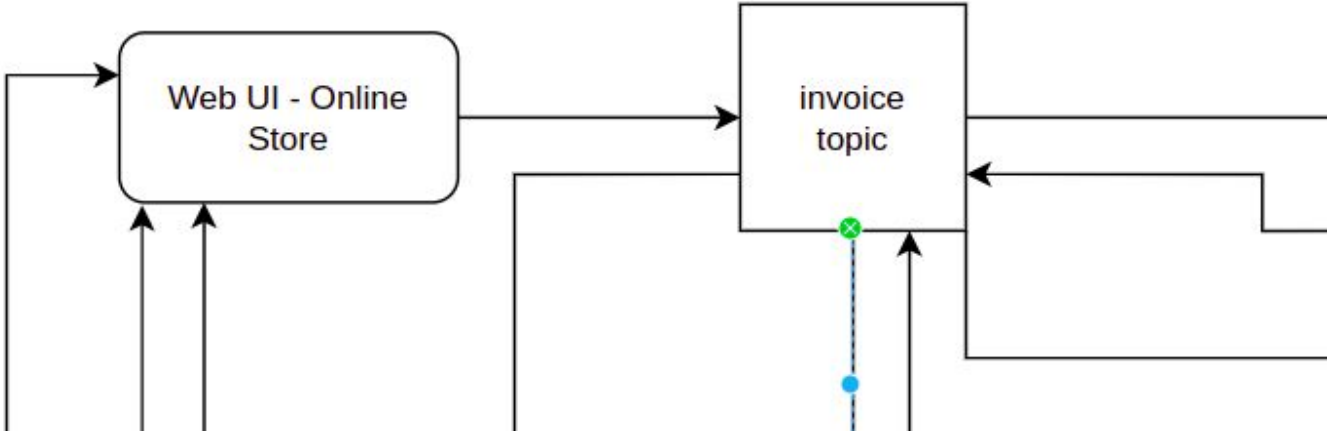
Medius

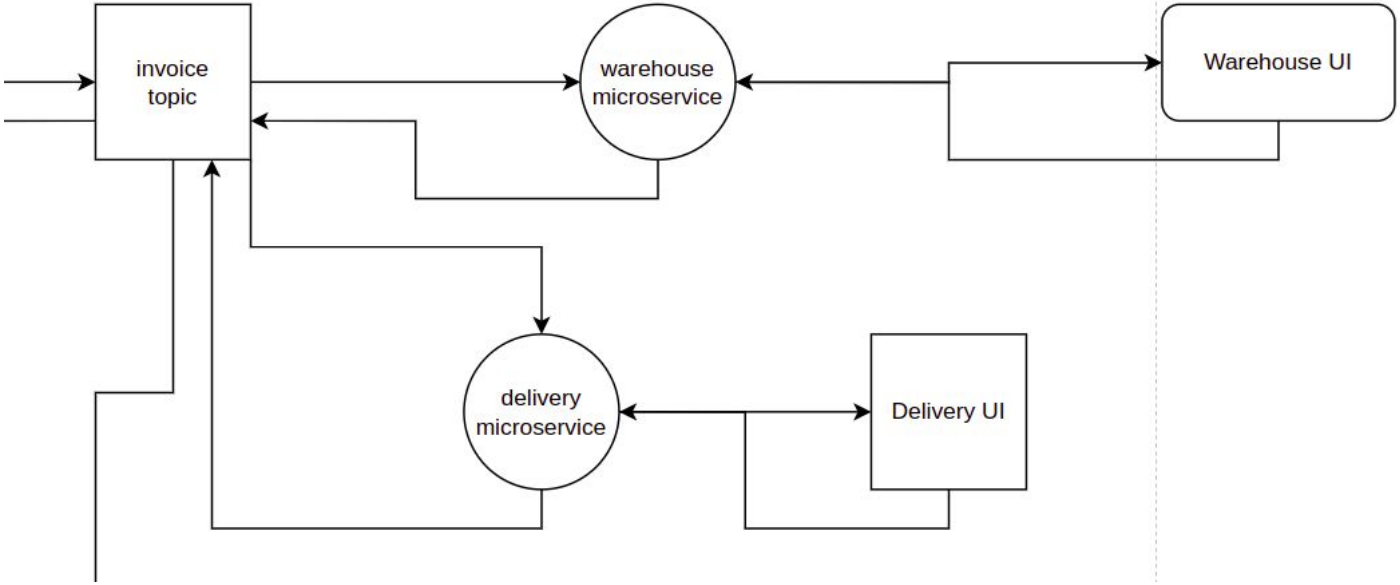


Invoice (račun)

- customerID
- invoiceDate
- country
- invoicePrice
- invoiceNo
- status
- items:
    - description
    - quantity
    - unitPrice
    - stockCode

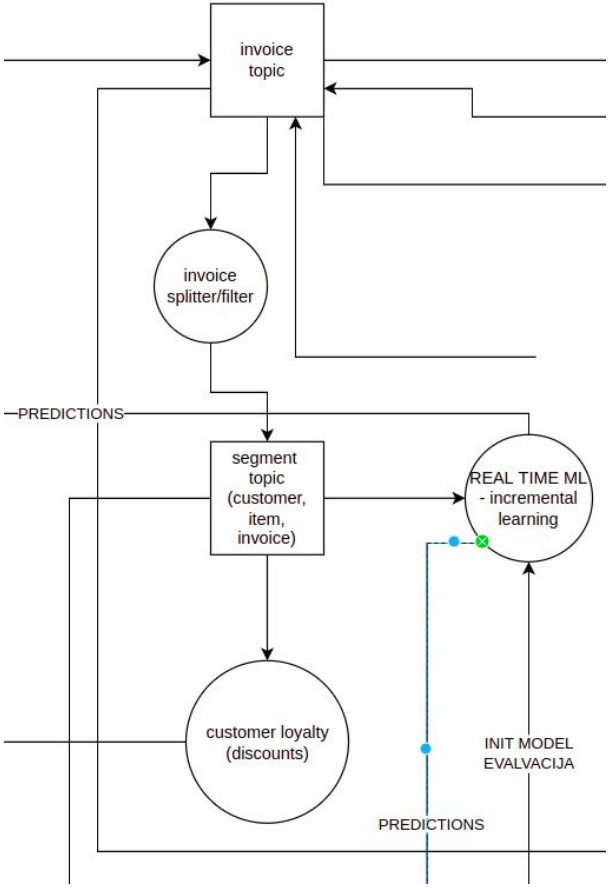
DEMO

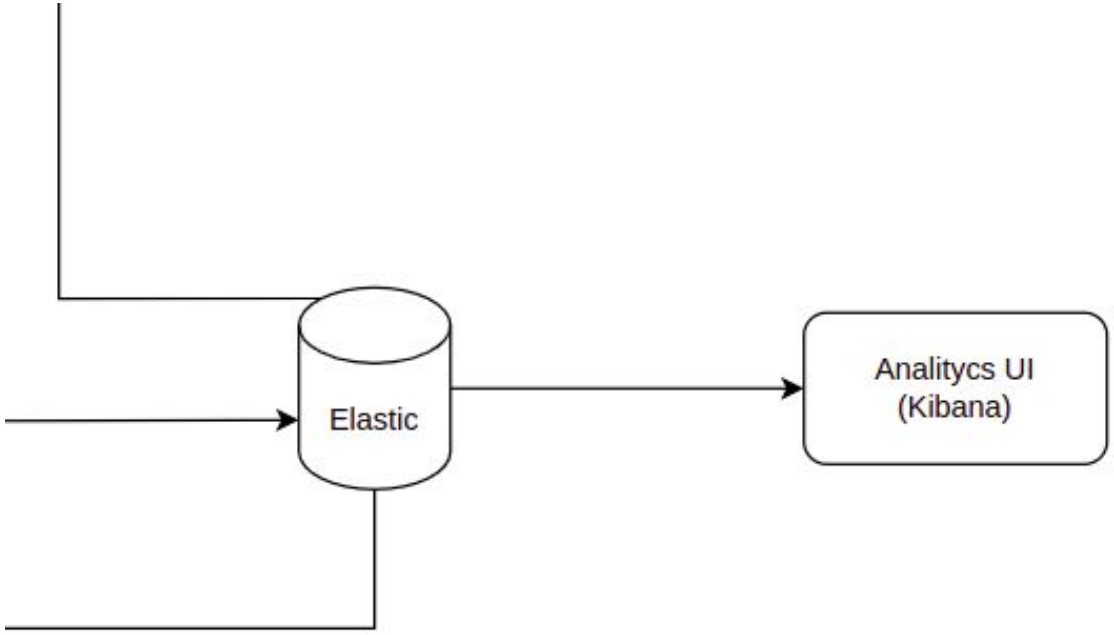

Web UI - Online Store
invoice topic


invoice
topic
warehouse
microservice
Warehouse UI
delivery
microservice
Delivery UI

Spletna trgovina

Medius

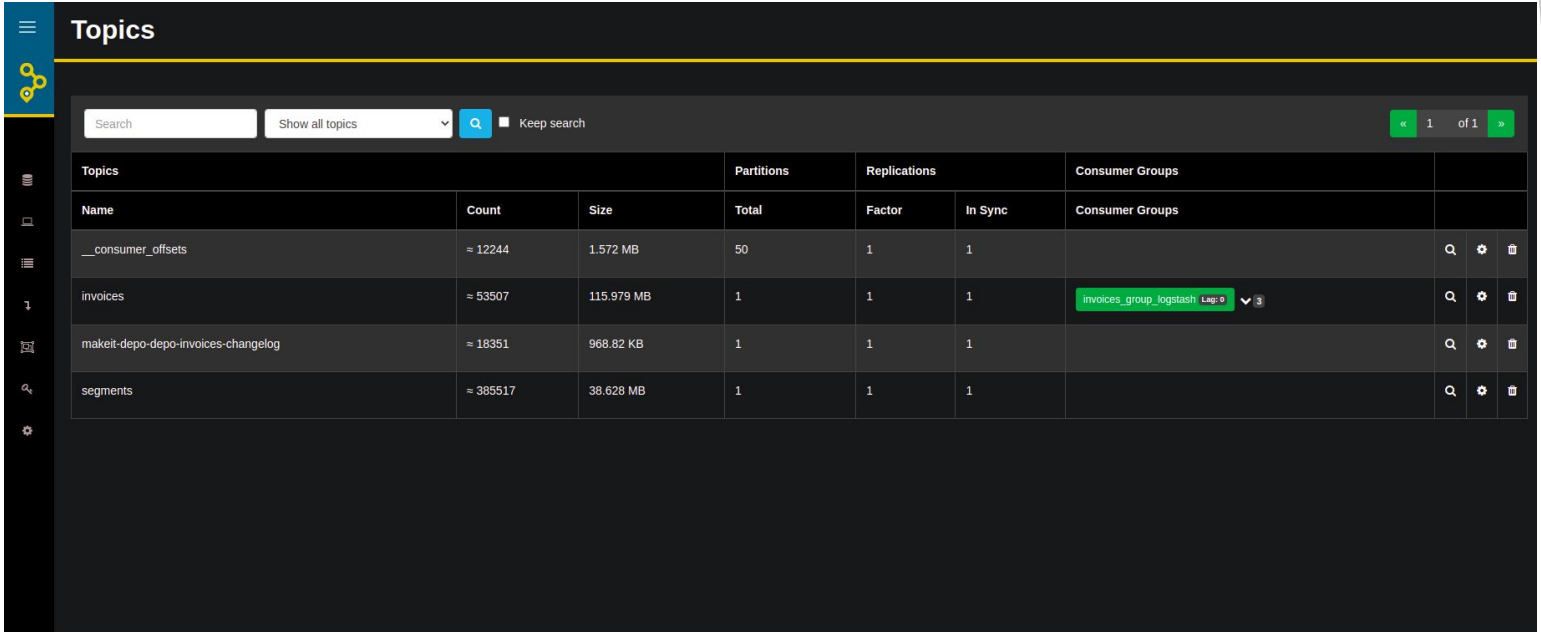



Elastic
Analitycs UI
(Kibana)

Spletna trgovina - demo

Medius

Pregled delovanja kafke



Warehouse

```
@Produces
public Topology depoStream() {
    StreamsBuilder builder = new StreamsBuilder();
    ObjectMapperSerde<Invoice> invoiceSerde = new ObjectMapperSerde<>(Invoice.class);

    Map<String, String> changeLogConfig = new HashMap<>();

    StoreBuilder<KeyValueStore<String, Invoice>> stateStore = Stores.keyValueStoreBuilder(
                Stores.persistentKeyValueStore(streamConfig.storeName()),
            Serdes.String(),
            invoiceSerde
            )
            .withLoggingEnabled(changeLogConfig)
            .withCachingEnabled();

    final Topology topology = builder.build();
    topology.addSource(streamConfig.sourceName(), Serdes.String().deserializer(), invoiceSerde.deserializer(), streamConfig.inTopic());
    topology.addProcessor(streamConfig.processorName(), DepoInvoiceProcessor::new, streamConfig.sourceName());
    topology.addStateStore(stateStore, streamConfig.processorName());
    log.info("Stream started");
    return topology;
}
```

Invoice splitter

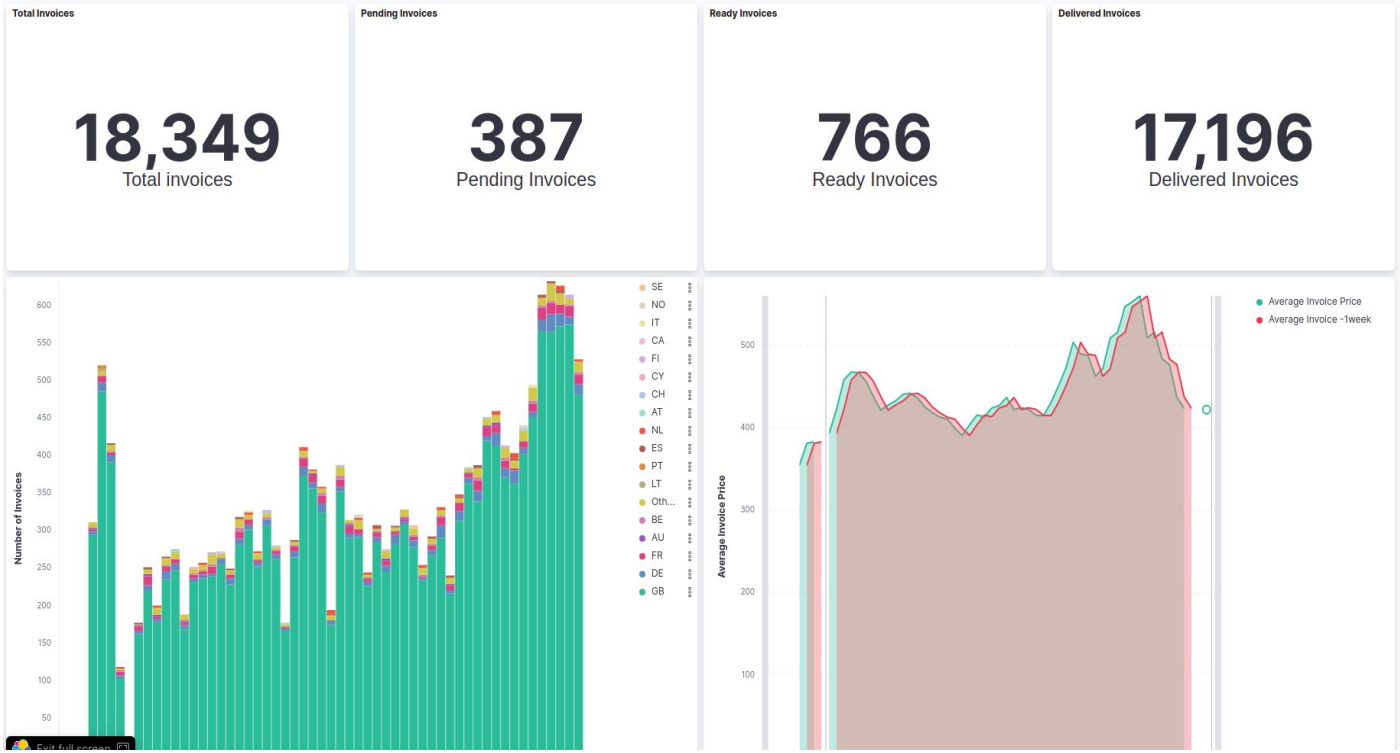
```java
@Produces
public Topology buildStream()
{
    StreamsBuilder builder = new StreamsBuilder();
    ObjectMapperSerde<Invoice> invoiceSerde = new ObjectMapperSerde<>(Invoice.class);
    ObjectMapperSerde<Segment> segmentSerde = new ObjectMapperSerde<>(Segment.class);

    builder.stream(streamConfig.inTopic(), Consumed.with(Serdes.String(), invoiceSerde)) KStream<String, Invoice>
            .filter((k, v) -> v.getStatus() == InvoiceStatus.PENDING)
            .flatMapValues(v -> {
                List<Segment> segments = new ArrayList<>();
                for (var i : v.getItems())
                {
                    Segment segment = new Segment();
                    segment.setInvoiceId(v.getInvoiceNo());
                    segment.setItemId(i.getStockCode());
                    segment.setCustomerId(v.getCustomerID());
                    segments.add(segment);
                }
                return segments;
            }) KStream<String, Segment>
            .to(streamConfig.outTopic(), Produced.with(Serdes.String(), segmentSerde));
    return builder.build();
}
```

Primer dokumenta na elasticsearchu

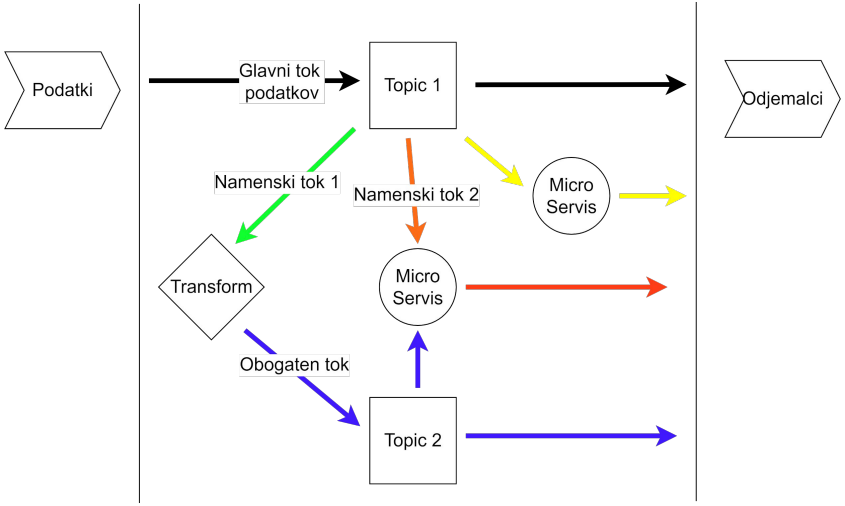
Actions	Field	Value
▫▫▫	_id	581586
▫▫▫	_index	invoices
▫▫▫	# _score	-
▫▫▫	@timestamp	May 30, 2022 @ 07:55:03.716
▫▫▫	k @version	1
▫▫▫	k country	GB
▫▫▫	k customerID	13113
▫▫▫	t event.original	> {"invoiceNo":"581586","customerID":"13113","country":"GB","invoiceDat e":"2021-12-11T12:49:00","items":[{"stockCode":"22061","description":"LA RGE CAKE STAND HANGING STRAWBERY","quantity":8,"unitPrice":2.95},{"stoc kCode":"23275","description":"SET OF 3 HANGING OWLS OLLIE BEAK","quantit y":24,"unitPrice":1.25},{"stockCode":"21217","description":"RED RETROSPO T ROUND CAKE TINS","quantity":24,"unitPrice":8.95},{"stockCode":"2068
▫▫▫	invoiceDate	Dec 11, 2021 @ 13:49:00.000
▫▫▫	k invoiceNo	581586
▫▫▫	# invoicePrice	339.2
▫▫▫	k items.description	LARGE CAKE STAND HANGING STRAWBERY, SET OF 3 HANGING OWLS OLLIE BEAK, R ED RETROSPOT ROUND CAKE TINS, DOORMAT RED RETROSPOT
▫▫▫	# items.quantity	8, 24, 24, 10
▫▫▫	k items.stockCode	22061, 23275, 21217, 20685
▫▫▫	# items.unitPrice	2.95, 1.25, 8.95, 7.08
▫▫▫	k status	DELIVERED

Primer kibana dashboarda



KODA:
https://github.com/dsenica/medius-makeit-2022

Kontekstno odvisni tokovi podatkov - Zaključek



Hvala za pozornost!

Vprašanja?