# Projekat

Tema: *Analiza poljoprivrednih podataka*

Predmet: *Arhitekture sistema velikih skupova podataka*

**Student**
*Stefan Aleksić E2-42-2022*

Novi Sad, januar 2023.

# Sadržaj

# Specifikacija projekta

## Uvod

Poljoprivreda ili agrokultura je praksa uzgoja biljaka i stoke. Poljoprivreda je bila ključni razvoj u usponu sedelačke ljudske civilizacije, pri čemu je poljoprivreda domaćih vrsta stvarala viškove hrane koja je omogućavala ljudima da žive u gradovima. Istorija poljoprivrede počela je pre više hiljada godina. Nakon što su sakupljali divlje žitarice pre najmanje 105.000 godina, poljoprivrednici u nastajanju počeli su da ih sade pre oko 11.500 godina. Ovce, koze, svinje i goveda pripitomljeni su pre više od 10.000 godina. Biljke su samostalno uzgajane u najmanje 11 regiona sveta. Industrijska poljoprivreda zasnovana na monokulturi velikih razmera u dvadesetom veku počela je da dominira poljoprivrednom proizvodnjom, iako je oko 2 milijarde ljudi i dalje zavisilo od poljoprivrede za samostalne potrebe.

Glavni poljoprivredni proizvodi mogu se široko grupisati u hranu, vlakna, goriva i sirovine (kao što je guma). Klase hrane uključuju žitarice (zrna), povrće, voće, ulja za kuvanje, meso, mleko, jaja i gljive. Preko jedne trećine svetskih radnika zaposleno je u poljoprivredi, odmah iza uslužnog sektora, iako se poslednjih decenija nastavlja globalni trend smanjenja broja poljoprivrednih radnika, posebno u zemljama u razvoju, gde male posede preuzima industrijska poljoprivreda. i mehanizacija koja donosi enormno povećanje prinosa.

Savremena agronomija, oplemenjivanje biljaka, agrohemikalije kao što su pesticidi i đubriva, i tehnološki razvoj naglo su povećali prinose useva, ali uzrokuju ekološku i ekološku štetu. Selektivni uzgoj i moderne prakse u stočarstvu su na sličan način povećale proizvodnju mesa, ali su izazvale zabrinutost za dobrobit životinja i štetu po životnu sredinu. Pitanja životne sredine uključuju doprinose globalnom zagrevanju, iscrpljivanju vodonosnih slojeva, krčenju šuma, otpornosti na antibiotike i drugom poljoprivrednom zagađenju. Poljoprivreda je i uzrok i osetljiva na degradaciju životne sredine, kao što je gubitak biodiverziteta, dezertifikacija, degradacija zemljišta i globalno zagrevanje, što sve može da izazove smanjenje prinosa useva. Genetski modifikovani organizmi se široko koriste, iako su neki zabranjeni u određenim zemljama.

## Domen

Domen projekta u širem kontekstu predstavlja poljoprivreda odnosno agrokultura. Poljoprivreda je jedna od 5 grana privrede i predstavlja praksu uzgoja biljaka i stoke. Podaci koji će biti obrađeni su skinuti sa zvaničnog sajta statističkih podataka o hrani i argokulturi Ujedinjenih Nacija (*eng. Food and Argiculture Organization of United Nations – FAO*) [FAOSTAT].

Drugi aspekt ovog projekta jeste analiza podataka o cenama žitarica na tržištu. Ovi podaci su dostupni od strane departmana za poljoprivredu Sjedinjenih Američkih Država (*eng. United States Departman of Agriculture*). Skinuti podaci u bliskoj prošlosti su skinuti, međutim, na raspolaganju je

i javno dostupan API na adresama [Internal Agriculture Transport USDA - Grain Prices] i [Internal Agriculture Transport USDA - Grain Basis].

# Motivacija

S obzirom da je poljoprivreda jedna od glavnih grana privrede, kao i činjenica da izuzetno utiče na ljudski opstanak, smatram da je neophodno u što većoj meri podsticati njen razvitak. Analizom udela zemalja u globalno skladište prinosa može da se proširi slika o tome kako se poljoprivreda skalirala kroz godine, da li je napredak postoji i da li je kontinualan. Ovo dalje otvara vrata da se dublje zađe i razrade pitanje o tome da ukoliko je bilo oscilacija, zašto je do njih došlo. Eventualno da se identifikuju okidači promena i u skladu sa ishodom nađe njihova prevencija ili stimulacija.

# Ciljevi

Cilj ovog projekta jeste da se izvrši analizira istorijskih podataka vezanih za statistiku godišnje proizvodnje različitih vrsti biljaka u različitim zemljama sveta za vremenski period od 1961. do 2021. godine. Takođe je u planu analiza podataka kroz vizuelizaciju i kombinovanje sa podacima o cenama žitarica na tržištu.

# Pitanja i odgovori

Lista konkretnih upita na koja se  očekuje odgovor od strane paketne obrade:

1. Prosečan prinos svake zemlje za svaku kategoriju proizvoda u odgovarajućem periodu?
2. Prosečan globalni prinos svake kategorije proizvoda u određenom periodu?
3. Zemlje koje su imale maksimalan prinos za svaku kategoriju proizvoda u odgovarajućem periodu?
4. Zemlje koje su imale minimalni prinos za svaku kategoriju proizvoda u odgovarajućem periodu?
5. Godine proizvodnje sa maksimalnim i minimalnim prinosom za svaku od zemalja?
6. Rangiranje zemalja na osnovu prinosa odgovarajuce kategorije
7. Rangiranje zemalja na osnovu globalnog prinosa
8. Rangiranje kontinenata na osnovu prinosa odgovarajuce kategorije?
9. Rangiranje kontitnenata na osnovu globalnog prinosa?
10. Globalni prinos za odredjenu kategoriju u odredjenom vremenskom periodu

Lista konkretnih upita na koje se očekuje odgovor od strane obrade podataka u realnom vremenu:

1. Prosečna cena određene žitarice
2. Najmanja i najviša cena određene žitarice na tržištu
3. Rangiranje ponuda za određenu žitaricu
4. Tržišta sa najnižim i najvišim cenama žitarica
5. Varijacija u cenama žitarica

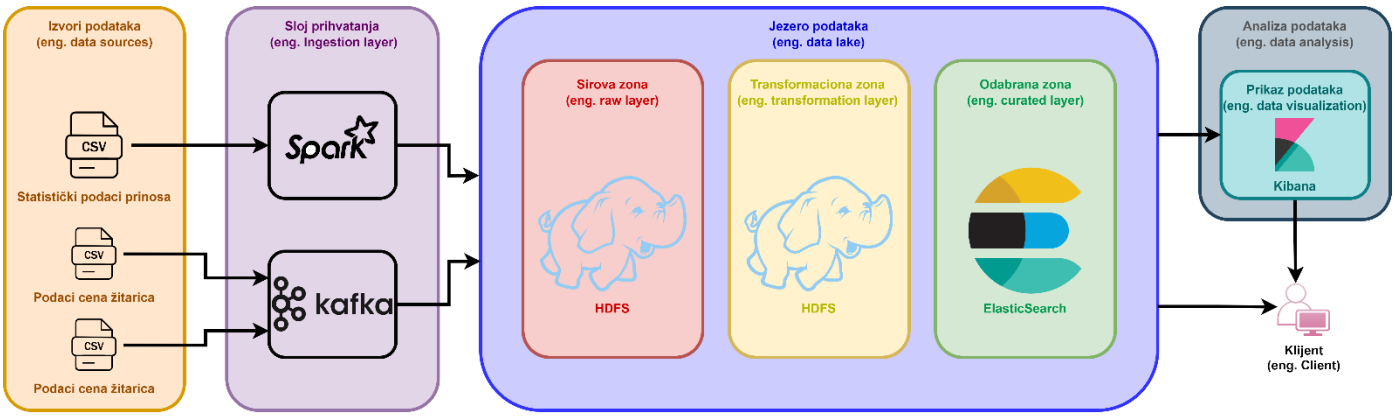
# Arhitektura sistema

## Definicija jezera podataka

## Specifikacija modula

Na slici 1 se nalazi arhitektura sistema za obradu podataka. Sistem se sastoji iz:

- izvora podataka, što predstavlja dataset-ove skinute sa gore navedenih linkova/API-ja,
- sloja prihvatanja (*eng. ingestion layer*) koji se sastoji iz:
  - modula za paketnu obradu – *Spark* i
  - modula za obradu podataka u realnom vremenu – *Kafka* i *Spark Streaming*,
- jezera podataka (*eng. data lake*) koje se sastoji iz tri zona:
  - Sirova zona (*eng. raw layer*) u okviru koje se podaci skladište pomoću *HDFS* modula,
  - Transformaciona zona (*eng. transformation zone*) u okviru koje se podaci skladište pomoću *HDFS* modula i
  - Odabrane zone (*eng. curated zone*) u okviru koje se podaci skladište pomoću *ElasticSearch* modula.
- modula za analizu podataka u okviru kog se nalazi:
  - modul za vizuelizaciju podataka – Kibana.




*Slika 1 – Arhitektura sistema*

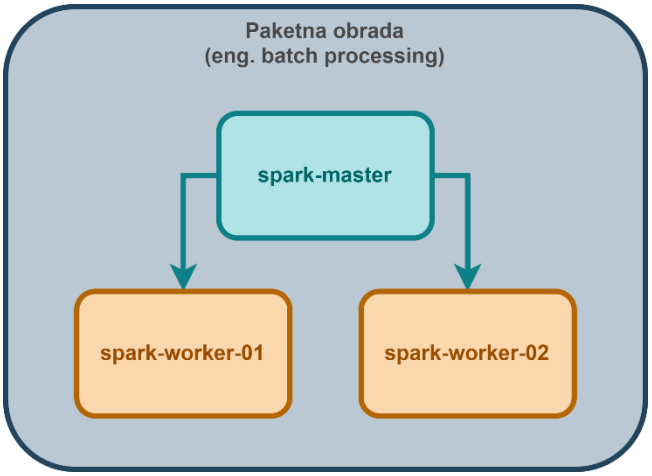
## Sloj prihvatanja podataka

Sloj prihvatanja podataka (*eng. ingestion layer*) je ostvaren uz pomoć dva modula:

- Modul za paketnu obradu podataka (*eng. batch processing*) – *Spark* i
- Modul za obradu podataka u realnom vremenu (*eng. realtime processing*) – *Kafka* + *Spark Streaming.*

### Paketna obrada podataka

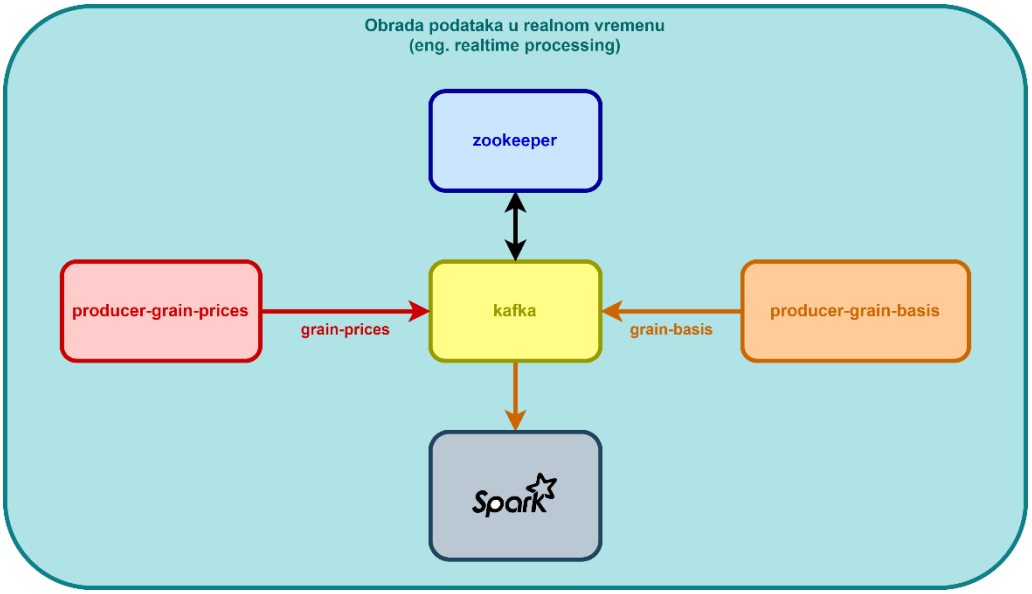
Na slici 2 je prikazan modul za paketnu obradu podataka. Za paketnu obradu je odabran Spark, koji je realizovan jednim master čvorom i sa dva čvorova radnika. Master čvor je pokrenut u okviru kontejnera spark-master, dok su čvorovi radnici pokrenuti u okviru čvorova spark-worker-01 i spark-worker-02.




*Slika 2 – Paketna obrada podakata*

### Obrada podataka u realnom vremenu

Dijagram modula za obradu podataka u realnom vremenu nalazi se na slici 3. Ovaj modul se oslanja na *SparkStreaming* za samu obradu podataka, koji je zapravo pokrenut u okviru modula za paketnu obradu podataka. Samo očitavanje podataka je realizovana uz pomoć *Kafke*. Na odgovarajuće topike se vrši objavljivanje podata. Sam *Kafka* čvor je pokrenut u okviru kontejnera ***kafka***, oslanja se na kontejner ***zookeeper***, dok podatke generišu čvorovi ***producer-grain-prices*** i ***producer-grain-basis***, objavljivanjem podataka iz odgovarajućih *dataset*-ova na topike ***grain-prices*** i ***grain-basis*** respektivno.

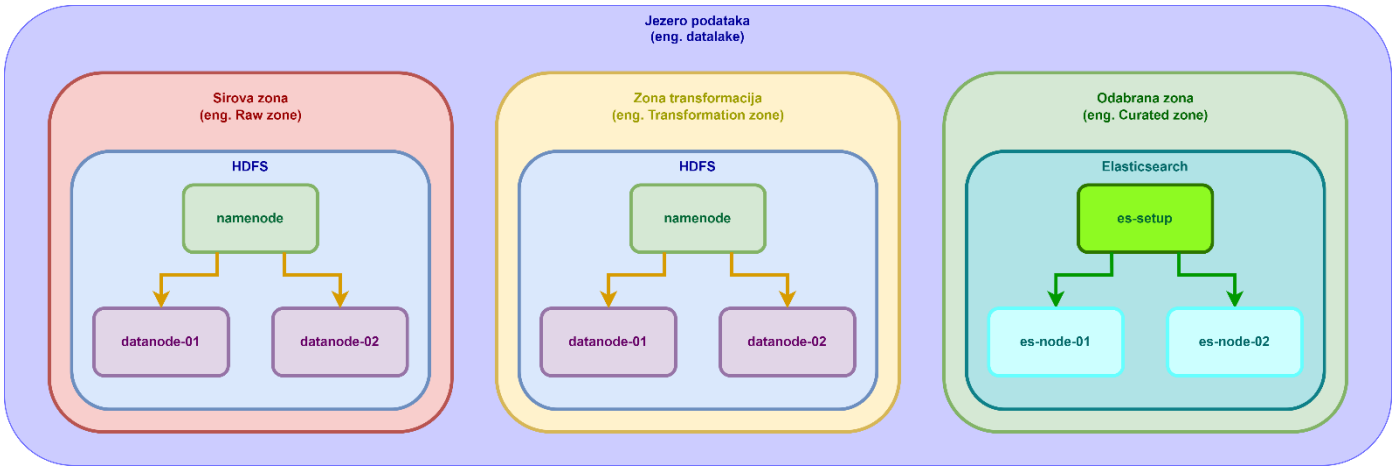



*Slika 3 – Dijagram modula za obradu podataka u realnom vremenu*

## Jezero podataka

Na slici 4 se nalazi detaljniji dijagram arhitekture jezera podataka. Već nabrojani slojevi jezera su:

- Sirova zona (*eng. raw layer*) – podaci se ovde skladište koristeći HDFS koji je sačinjen od jednog imenskog čvora (*eng. namenode*) i dve instance čvora podataka (*eng. datanode*). Imenski čvor je pokrenut u okviru kontejnera ***namenode***, dok su *čvorovi podataka* pokrenuti u okviru kontejnera ***datanode-01*** i ***datanode-02***.
- Zona transformacije (*eng. transformation zone*) – podaci se ovde skladište koristeći *HDFS* koji je sačinjen od jednog imenskog čvora (*eng. namenode*) i dve instance čvora podataka (*eng. datanode*). Imenski čvor je pokrenut u okviru kontejnera ***namenode***, dok su *čvorovi podataka* pokrenuti u okviru kontejnera ***datanode-01*** i ***datanode-02***.
- Zona odabranih podataka (*eng. curated zone*) – podaci se ovde skladište koristeći *ElasticSearch*. Klaster je pokrenut sa dve instance *ElasticSearch* u okviru kontejnera ***es-node-01*** i *es-node-02* za čiju inicijalizaciju je odgovoran kontejner ***es-setup***.




*Slika 4 – Arhitektura jezera podataka*

# Obrada podataka

Paketna obrada

Obrada tokova podataka

# Reference

**There are no sources in the current document.**