

Uvod

Arhitekture sistema velikih skupova podataka

Pravila ocenjivanja - praktični deo - vežbe

- 55 bodova
 - U toku nastave
 - 1 projekat (P1)
 - razvoj sistema velikih skupova podataka
 - 3 projektna zadatka – složeni oblici vežbi (Z1-Z3)
 - specifikacija projekta - prezentacija skupa podataka i željene obrade (5 bodova)
 - inicijalno postavljanje arhitekture - dijagram i kontejnerizovani moduli (5 bodova)
 - završna odbrana projekta (45 bodova)
 - paketna obrada (25 bodova)
 - obrada podataka u realnom vremenu (15 bodova)
 - sistem za razmenu poruka (5 bodova)
 - biće objavljena Specifikacija projekta sa naznačenim bitnim detaljima

Pravila ocenjivanja - praktični deo - vežbe

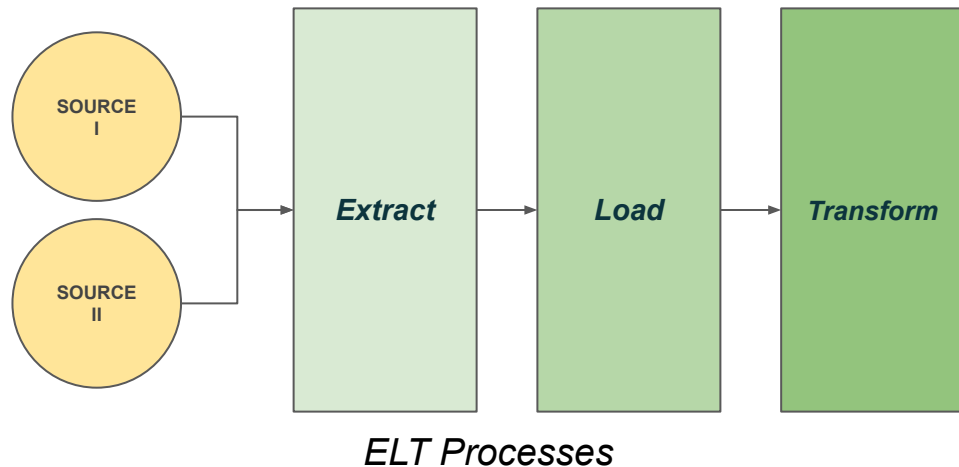
- Pravila realizacije obaveza
 - projektni zadaci – složeni oblici vežbi (Z1 - Z3)
 - odbrana: na nastavi, u toku semestra, na času vežbi, koji je za to unapred planiran, pred asistentom
 - realizuje se na nastavi i u samostalnom radu
 - student odabira temu za projekat i samostalno ga implementira po specifikaciji
 - student dobija na vežbama primere i zadatke koji predstavljaju pripremu za njihovu realizaciju

Kontekst - Skupovi podataka

- Skup podataka I
 - sadrži istorijske podatke na neku temu
 - biće korišten za paketnu obradu
 - postojeći skup ili samostalno kreiran
 - veličine >300MB
- Skup podataka II
 - ima karakteristike toka podataka (eng. *stream*)
 - biće korišten za obradu u realnom vremenu
 - logički povezan sa Skupom podataka I
 - može nastati:
 - korišćenjem javnih API-ja
 - periodičnim dovlačenjem podataka
 - generisanjem toka podataka
 - novi skupa sa istorijskim podacima
 - sadrži vremensku dimenziju
- Skupovi moraju biti različiti

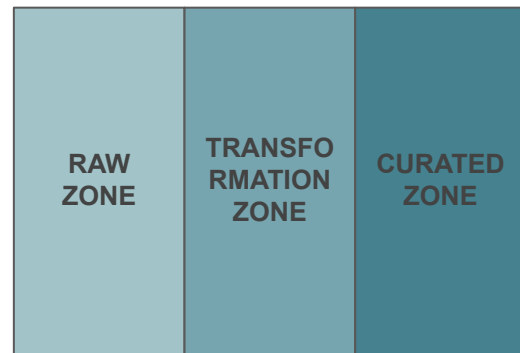
Kontekst - Ekstrakcija, učitavanje i transformacija podataka

- Eng. *Extract, Load, Transform* (*ELT*) procesi
 - Podaci se
 - ekstraktuju sa izvora,
 - učitavaju u jezero podataka (eng. *Data Lake*)
 - i transformišu da bi bili pogodni za upotrebu



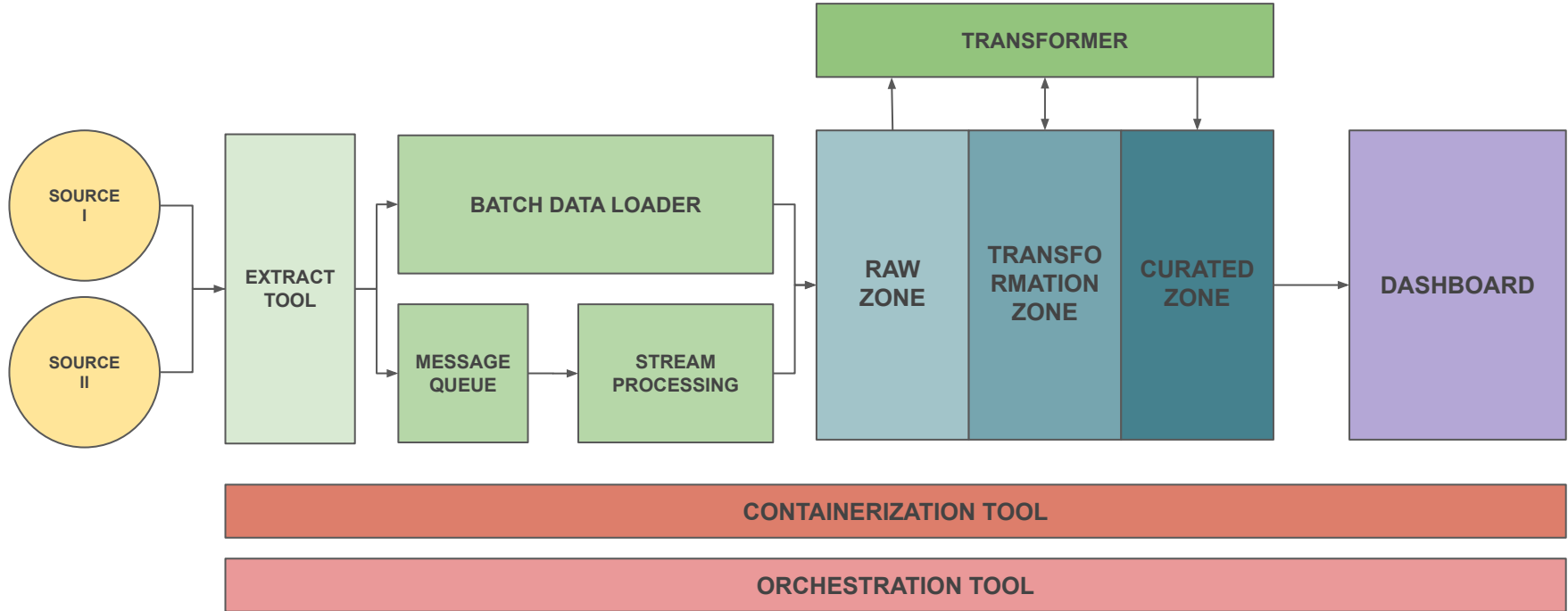
Kontekst - Jezero podataka

- Podaci koji se sa izvora ekstraktuju i učitavaju u jezero podataka
- Podaci se, shodno nameni, čuvaju u tri zone jezera podataka
 - Zona sirovih podataka
 - Zona transformacija
 - *Curated* zona

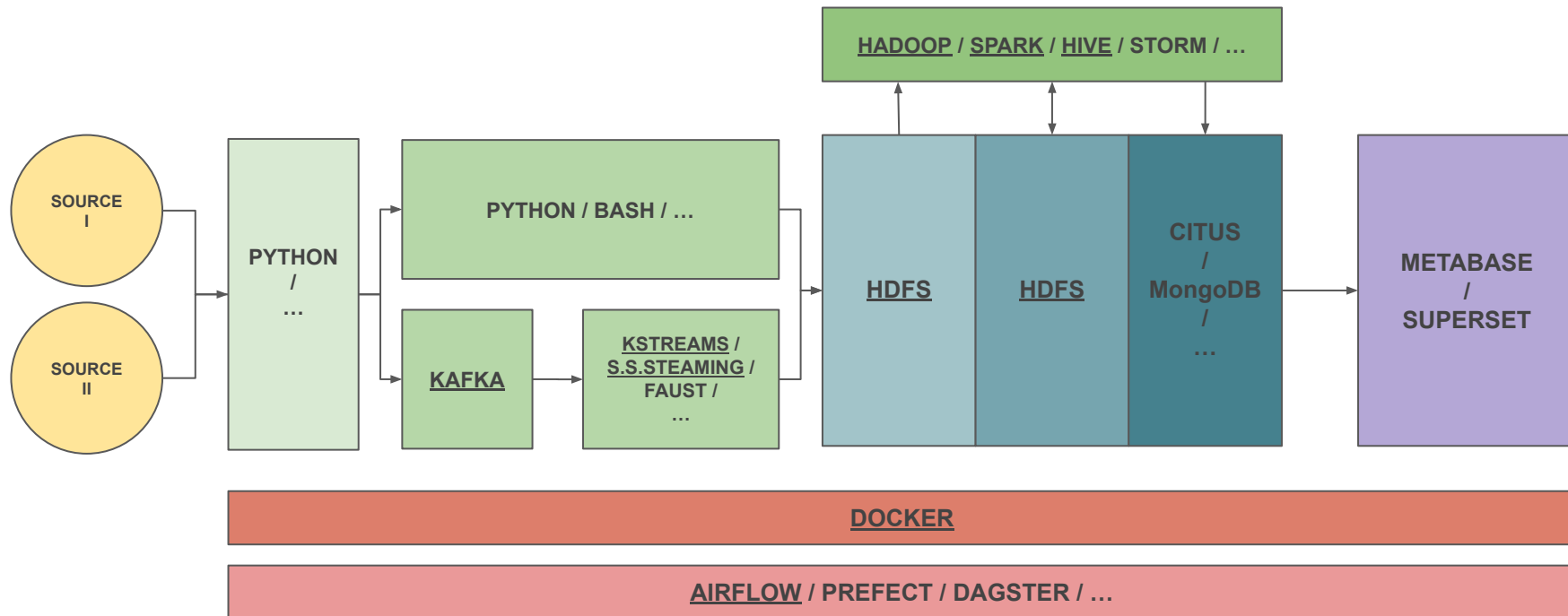


Data Lake

Arhitektura sistema



Implementazioni detalji



Primeri na vežbama - projekat

- Pripremljen projekat
 - Sa svim alatima koje će studenti upoznati u toku nastave
 - Link: <https://github.com/tmiroslav97/asvsp-architecture>
 - Svi alati su dockerizovani
 - Za rad u učionici:
 - Mogu biti neophodne manje korekcije zbog verzije dokera
 - Radi pripreme okruženja:
 - Pogasiti sve nepotrebne kontejnere
 - Odraditi: `docker system prune`
 - Skinuti [VSCode portable](#)
 - Otpakovati i u folderu gde vidite code fajl pokrenuti terminal
 - Pokrenuti `./code`
 - Ručno otvoriti folder iz otvorenog *VSCode*
 - Docker slike verovatno se neće menjati u toku semestra
 - Svuci ih na uvodnom terminu kako bi realizacija vežbi bila olakšana

Primeri na vežbama - projekat

- Pripremljen projekat
 - Pokretanje željenih alata:
 - `./scripts/cluster_up.sh alat1 alat2 ... alatn`
 - Gašenje (svih) alata:
 - `./scripts/cluster_down.sh`

Primeri na vežbama - skup podataka

- Predstavlja kolekciju zapisa mrežne komunikacije
 - Generisan pomoću *Data-Generator* alata
 - [Link](#) ka zip arhivi (~50MB)
 - Unutar archive - *jsonlines* dokument (~160MB)
 - Može doživeti izmene u toku semestra
 - Pravovremena najava od strane asistenata
 - Za osluškivanje (eng. sniff) mrežne komunikacije korišćena [scapy](#) python biblioteka
 - Dokumentacija [sniff](#) metode
 - Svaki zapis u kolekciji predstavlja paket, opisan kroz hijerarhiju OSI slojeva
 - Relevantni linkovi za razumevanje polja zapisa:
 - [Scapy in 15 minutes](#)
 - [IPv4 Packet Header](#)
 - Na istom sajtu se mogu naći pojašnjenja za polja iz zaglavlja ostalih slojeva