

Tehnička specifikacija projekta

Arhitekture sistema velikih skupova podataka

Cilj

Projekat iz ovog predmeta ima za cilj osmišljavanje i realizaciju arhitekture sistema za obradu velikih skupova podataka kao i demonstraciju upotrebe takvog sistema kroz različite primere transformacije, analize i prezentacije podataka.

Opis projekta

- Skupovi podataka:
 - Potrebno je obezbediti bar dva skupa podataka iz različitih izvora.
 - Jedan od ta dva skupa se smatra primarnim i treba da sadrži istorijske podatke na zadatu/izabranu temu.
 - Primarni skup podataka treba biti veći od 300 MB, može biti preuzet sa javnih repozitorijuma podataka (data.gov, data.worldbank.org, data.gov.rs, kaggle.com, datasetsearch.research.google.com) ili može biti sakupljen npr. tehnikom *web scraping-a*.
 - Drugi skup treba da ima karakteristike toka podataka koji je na neki način logički povezan sa temom primarnog skupa podataka.
 - Drugi skup (tok) podataka može nastati korišćenjem javnih API-ja (npr. putem *WebSocket-a*), periodičnim dovlačenjem podataka ili generisanjem toka podataka od postojećeg (istorijskog) skupa podataka koji ima vremensku dimenziju.
 - Bitno je da su ova dva skupa podataka nastala iz različitih izvora, nije dozvoljeno isti početni skup koristiti za kreiranje primarnog skupa i toka podataka.
- Jezero podataka (eng. *data lake*):
 - Potrebno je projektovati i implementirati jezero podataka sa minimalno 3 zone (sloja):
 - sirova (eng. *raw*) zona,
 - zona transformacija,
 - *curated* zona.
 - Potrebno je automatizovati učitavanje odabranog skupa podataka u jezero podataka.
- Obrada podataka i prezentacija rezultata obrade:
 - Potrebno je osmisлити svrsishodnu analizu podataka, kojom bi trebalo da se dobiju korisna saznanja iz odabranog skupa podataka.
 - U ovu svrhu, osmisлити jednu ili dve persone koje predstavljaju zainteresovanu stranu u procesu analize podataka; kroz prizmu ovih persona posmatrati relevantnost i svrsishodnost definisane analize podataka.
 - Rezultate obrade neophodno je vizualizovati krajnjem korisniku.

Zadaci

- Specifikacija projekta - prezentacija skupa podataka i željene obrade:
 - opisati domen, motivaciju, ciljeve, kao i
 - navesti konkretna pitanja na koja bi analiza podataka trebala da da odgovor
 - makar 10 pitanja za paketnu obradu podataka i
 - makar 5 pitanja za obradu podataka u realnom vremenu.
- Inicijalno postavljanje arhitekture - dijagram i kontejnerizovani moduli:
 - definicija jezera podataka i
 - specifikacija modula koji će biti korišćeni za željenu obradu podataka
 - dati dijagramsku predstavu celokupne arhitekture sistema i
 - pripremiti komponente za korišćenje u kontejnerizovanom obliku
 - potrebno kontejnerizovati sve komponente arhitekture,
 - iskoristiti pripremljeni sistem za implementaciju odgovora na makar jedno pitanje postavljeno za paketnu obradu podataka
- Obrada podataka
 - Paketna obrada:
 - potrebno je implementirati makar 10 različitih kompleksnih upita/transformacija na podacima iz jezera podataka
 - u ovu svrhu, za pripremu podataka iskoristiti neki od alata koji omogućavaju paralelnu obradu velike količine podataka,
 - potrebno je koristiti analitičke window funkcije i
 - najmanje 3 rezultata upita/transformacija potrebno je prezentovati koristeći vizualizacionu tehnologiju po želji.
 - Obrada u realnom vremenu / obrada tokova podataka:
 - potrebno je implementirati do 5 kompleksnih transformacija tokova podataka (eng. *stream processors*),
 - potrebno je koristiti spajanje tokova ili spajanje toka sa podacima paketnog tipa, kao i agregaciju sa upotrebom *Windowing*-a i
 - rezultat obrade tokova podataka smestiti u skladište/bazu podataka po želji (npr. *Kudu*, *Druid*, *ElasticSearch*, *Citus*).
 - Orkestracija obrade podataka:
 - potrebno je obezbediti mehanizme za automatizovano pokretanje procesa za obradu podataka.
 - Javni git repozitorijum projekta sa README
 - vezu ka repozitorijumu proslediti nadležnom asistentu PRE odbrane projekta.