

**OBRADA VELIKIH SKUPOVA PODATAKA KORIŠĆENJEM CLOUD TEHNOLOGIJA**  
**PROCESSING LARGE DATA SETS USING CLOUD TECHNOLOGIES***Bosiljka Todić, Fakultet tehničkih nauka, Novi Sad***Oblast – ELEKTROTEHNIKA I RAČUNARSTVO**

**Kratak sadržaj** – U ovom radu analizirani su dnevni valutni kursevi korišćenjem Google BigQuery platforme, koja omogućava obradu velikih skupova podataka i predikciju budućih trendova. Analiza podataka uključuje identifikaciju dugoročnih trendova i volatilnosti valuta, uz korišćenje grafičkih prikaza i prediktivnih modela. Fokus je na upotrebi BigQuery-a kao efikasnog cloud alata, koji nudi skalabilnost i brzinu obrade kompleksnih upita. Cilj rada je pokazati potencijale cloud tehnologija i Big Data alata u analizi i donošenju strateških odluka.

**Ključne reči:** Cloud computing, Big Data, Google Cloud BigQuery, Kaggle, Dnevni valutni kursevi

**Abstract** – This paper analyzes daily currency exchange rates using the Google BigQuery platform, which enables the processing of large datasets and the prediction of future trends. The data analysis focuses on identifying long-term trends and currency volatility, complemented by graphical representations and predictive models. The emphasis is on BigQuery as a robust cloud tool that offers scalability and speed for executing complex queries. The goal is to demonstrate the potential of cloud technologies and Big Data tools in analysis and strategic decision-making.

**Keywords:** Cloud computing, Big Data, Google Cloud BigQuery, Kaggle, Daily exchange rates

**1. UVOD**

Razvoj savremenih tehnologija i sve veća količina dostupnih podataka stvorili su potrebu za efikasnim alatima i platformama za obradu velikih količina informacija. Ova potreba dovela je do širenja koncepta Big Data i cloud računarskih platformi. Danas se mnoge industrije, poput finansija, maloprodaje, zdravstva, trgovine valutama, oslanjaju na brze i precizne informacije kako bi donele strateške odluke. Mogućnost analize velikih skupova podataka u realnom vremenu postala je ključna za uspeh u ovim sektorima.

U ovom radu fokus će biti na obradi i analizi velikih skupova podataka korišćenjem cloud servisa. Konkretno, obrađivaće se skup podataka sa platforme Kaggle, koja služi za deljenje i analizu podataka.

**NAPOMENA:**

**Ovaj rad proistekao je iz master rada čiji mentor je bio dr Srđan Vukmirović, red. prof.**

Skup podataka sadrži dnevne promene valutnih kurseva u odnosu na baznu valutu, što pruža priliku za analizu dugoročnih trendova i promenljivosti valuta. Podaci će biti obrađeni, analizirani i grafički prikazani. Koristiće se Google-ova platforma BigQuery koja je serverless skladište podataka sa ugrađenim mogućnostima mašinskog učenja. Platforma ima mnoštvo funkcija koje pomažu u analitici različitih veličina i tipova podataka. Takođe, platforma nudi nekoliko načina za vizualizaciju, kao i kreiranje predikcija pomoću BigQuery ML ugrađenih funkcija. Cilj ovog rada je da prikaže kako cloud tehnologije i Big Data alati mogu biti korisni u analizi velikih skupova podataka, kroz konkretan primer valutnih kurseva.

**2. OSNOVNI POJMOVI**

U ovom poglavlju će biti detaljno objašnjeni ključni pojmovi koji čine osnovu za dalju razradu teme i analizu podataka u ovom radu. Fokus će biti na definisanju i razumevanju osnovnih principa cloud computinga i big data tehnologija, kao i na njihovoj međusobnoj povezanosti i značaju u savremenim IT sistemima.

**2.1. Cloud computing**

Cloud computing predstavlja isporuku IT resursa na zahtev putem interneta uz plaćanje prema korišćenju. Umesto kupovine, posedovanja i održavanja fizičkih data centara i servera, moguće je pristupiti tehnološkim uslugama, kao što su računarska snaga, skladištenje i baze podataka, po potrebi, od pružalaca cloud usluga [1].

Tipovi modela implementacije cloud computing-a:

- Javni cloud – u vlasništvu i pod upravom trećih strana, poput pružalaca cloud usluga kao što je Google Cloud. Ova vrsta clouda omogućava kompanijama pristup računarskim, skladišnim i mrežnim resursima putem interneta. Resursi su dostupni na zahtev, što firmama pruža fleksibilnost da ih koriste prema svojim specifičnim potrebama i poslovnim ciljevima, bez potrebe za ulaganjima u sopstvenu IT infrastrukturu.
- Privatni cloud – infrastruktura koju gradi, poseduje i upravlja jedna organizacija, pri čemu se resursi hostuju unutar njenih sopstvenih data centara. Ovaj pristup, poznat i kao „on-premises” rešenje, pruža kompanijama veću kontrolu, sigurnost i upravljanje podacima.

- Hibridni cloud – kombinuju modele javnog i privatnog cloud-a, omogućavajući kompanijama da koriste usluge javnog cloud-a, a istovremeno zadrže mogućnosti koje su uobičajene u arhitekturama privatnog cloud-a.

Postoje tri glavna tipa modela cloud computing usluga koje se mogu izabrati u zavisnosti od nivoa kontrole, fleksibilnosti i upravljanja koje korisnik zahteva:

- Infrastructure as a service (IaaS) – obezbeđuje osnovne resurse, kao što su serveri, skladištenje i mogućnosti umrežavanja, upotrebom virtuelnih mašina, kao što su Amazon EC2, Google Compute Engine i Microsoft Azure.
- Platform as a service (PaaS) – obezbeđuje platformu za razvoj, testiranje i implementaciju softverskih aplikacija, kao što su AWS Elastic Beanstalk, Google App Engine i Heroku.
- Software as a service (SaaS) – model usluge u cloud-u koji omogućava korisnicima pristup softverskim aplikacijama putem interneta, bez potrebe za instalacijom i održavanjem tih aplikacija na svojim lokalnim uređajima ili serverima, kao što su Amazon WorkSpaces, Gmail, Salesforce, Dropbox, Google Apps for Work i Microsoft Office 365.

## 2.2. Big Data

Ogromne količine podataka prikupljenih tokom vremena koje je teško analizirati i obrađivati korišćenjem uobičajenih alata za upravljanje bazama podataka. Ovi podaci se analiziraju radi identifikacije tržišnih trendova u poslovanju, kao i u oblastima proizvodnje, medicine i nauke. Tipovi podataka uključuju poslovne transakcije, e-poruke, fotografije, snimke nadzornih kamera, evidenciju aktivnosti i nestrukturirani tekstovi sa blogova i društvenih mreža, kao i ogromne količine podataka koje mogu prikupljati senzori različitih vrsta [2].

Big Data se definiše kroz pet ključnih karakteristika poznatih kao 5V. Prva karakteristika je obim (Volume), koja se odnosi na količinu podataka koji se generišu iz različitih izvora i u različitim formatima, često toliko velika da tradicionalne metode skladištenja i obrade nisu dovoljno efikasne. Zatim, tu je vrednost (Value), koja označava korisne uvide koje se mogu izvući iz podataka, a koji dodaju vrednost poslovnim procesima. Brzina (Velocity) se odnosi na brzinu generisanja, prikupljanja i analize podataka, što omogućava donošenje odluka u realnom vremenu. Raznovidnost (Variety) označava različite tipove podataka, uključujući strukturirane, polustrukturirane i nestrukturirane podatke, prikupljene iz brojnih izvora. Na kraju, tu je validnost/verodostojnost (Validity/Veracity), koja se odnosi na kvalitet i pouzdanost podataka, koji su ključni za preciznu analizu i donošenje tačnih odluka.

## 2.3. Povezanost Cloud computing i Big Data koncepta

Cloud computing i Big Data su međusobno povezane tehnologije koje su promenile način na koji organizacije upravljaju i analiziraju podatke. Big Data, generisana iz raznovrsnih izvora, često dolazi u različitim formatima.

Pružaoči cloud usluga nude alate poput veštačke inteligencije za standardizaciju i analizu ovih podataka.

Cloud platforme, poput Amazon S3 ili Google Cloud Storage, omogućavaju centralizovano i skalabilno skladištenje velikih količina podataka. U kombinaciji sa alatima poput Hadoop-a i Spark-a, podaci se mogu brzo obrađivati i transformisati u korisne uvide. Dodatno, alati za analitiku i vizualizaciju, kao što su Power BI i Tableau, pomažu organizacijama da identifikuju trendove i donesu bolje odluke.

Jedna od najvećih prednosti cloud-a je skalabilnost, zbog toga što organizacije mogu povećavati ili smanjivati resurse prema potrebama, plaćajući samo za ono što koriste. Ova fleksibilnost omogućava preduzećima svih veličina da analiziraju podatke na efikasan i isplativ način. Kombinacija cloud tehnologija i Big Data tako otvara vrata ka inovacijama, smanjenju troškova i optimizaciji poslovanja.

## 3. CLOUD SERVISI ZA OBRADU BIG DATA

Serijska obrada (batch processing) i strim obrada (stream processing) su dva osnovna načina obrade velikih skupova podataka. Ove metode prilagođene su specifičnim potrebama i slučajevima upotrebe. Serijska obrada omogućava analizu podataka iz prošlosti u unapred definisanim vremenskim intervalima, dok je strim obrada optimalna za aplikacije kojima su potrebni trenutni rezultati, kao što su detekcija prevara u realnom vremenu ili praćenje finansijskih transakcija.

Servisi za obradu velikih skupova podataka u cloud-u pružaju skalabilnost, pouzdanost i visoke performanse, omogućavajući kompanijama efikasno upravljanje i analizu podataka. Ključne cloud platforme i njihovi servisi uključuju:

- Amazon Web Services (AWS) – alati poput Amazon S3 za skladištenje, EMR za obradu (Hadoop/Spark), Redshift za analitiku i Kinesis za strimovanje podataka u realnom vremenu.
- Google Cloud Platform (GCP) – BigQuery za analitiku, Dataflow za paralelnu obradu, Dataproc za Hadoop/Spark i Pub/Sub za razmenu poruka.
- Microsoft Azure – HDInsight za Hadoop/Spark, Synapse Analytics za skladištenje i obradu, Data Lake Storage za velike skupove podataka i Stream Analytics za real-time analitiku.
- IBM Cloud – Watson Studio za AI, Cloud Object Storage za skladištenje i Streams za analitiku u realnom vremenu.
- Oracle Cloud – Autonomous Data Warehouse za AI analitiku i Big Data Service za Hadoop bazirane procese.
- Alibaba Cloud – MaxCompute za analizu podataka, DataWorks za ETL i EMR za Hadoop/Spark klastere.
- Cloudera – Platforma za hibridna rešenja koja kombinuje upravljanje podacima, analitiku i AI.
- Apache Hadoop: Osnova za mnoge Big Data servise sa HDFS za skladištenje i MapReduce za obradu.

## 4. OBRADA KAGGLE DATASETA POMOĆU GCP BIGQUERY SERVISA

Google Cloud BigQuery je platforma za podatke koja pomaže pri upravljanju i analiziranju podataka sa ugrađenim funkcijama kao što su mašinsko učenje, pretraga, analiza i poslovna inteligencija. Serverless arhitektura BigQuery-ja omogućava korišćenje programskog jezika poput SQL-a i Pythona za odgovaranje na najveća pitanja, bez potrebe za upravljanjem infrastrukturom [3]. Podaci su automatski replicirani na više lokacija radi visoke dostupnosti i otpornosti.

### 4.1. Opis skupa podataka sa Kaggle platforme

Skup podataka koji je korišćen pri obradi je dostupan na sledećem linku na Kaggle platformi:

<https://www.kaggle.com/datasets/asaniczka/forex-exchange-rate-since-2004-updated-daily/data>

Skup podataka o dnevnim valutnim kursovima pruža istorijske i aktuelne kurseve za preko 160 valuta. Ovaj skup podataka se ažurira svakodnevno i sadrži podatke o dnevnim valutnim kursovima od 2004. godine do danas. Uključuje pet kolona: valuta (currency) je kod strane valute, npr. USD za američki dolar, EUR za evro. Tip podataka je string. Sledeća kolona je osnovna valuta (base\_currency) koja predstavlja kod osnovne valute, u odnosu na koju se kurs izračunava, u ovom slučaju EUR. Tip podataka je string. Sledi kolona za puno ime valute (currency\_name), kao što su „Američki dolar” ili „Evro”. Tip podataka je string. Četvrta kolona je vrednost valutnog kursa (exchange\_rate), tj. odnos između strane valute i osnovne valute. Ovo je broj koji pokazuje koliko osnovne valute vredi jedna jedinica strane valute. Tip podataka je float. Poslednja kolona je datum (date) kada je valutni kurs zabeležen. Tip podataka je datetime.

Skup podataka sadrži skoro 400000 redova podataka o dnevnim valutnim kursovima i trenutna verzija zauzima 17.01 MB, što ga čini dovoljno obimnim za detaljne analize kretanja valutnih kurseva i tržišnih trendova na globalnom nivou.

### 4.2. Koraci za uvoz skupa podataka i SQL analiza

Koraci:

1. Priprema podataka – preuzeti dataset u CSV format i pripremiti za uvoz.
2. Kreiranje projekta – prijava na Google Cloud Console i kreiranje novog projekta.
3. Uvoz podataka – kliknuti na Create Table i izaberite CSV fajl sa lokalnog diska ili Google Cloud Storage-a. Definirati šemu kolona.
4. Opcije uvoza – Append to table ili Replace table
5. Analiza pomoću SQL upita za filtriranje i analizu podataka, kao npr. u listingu 1.

```
SELECT *
FROM `masterrad-438309.dailyForexRates.DailyForexRates`
WHERE date >= '2023-01-01' AND currency = 'USD'
ORDER BY exchange_rate ASC;
```

Listing 1. SQL upit za dobavljanje liste svih kurseva za USD od početka 2023. godine

## 5. ANALIZA SQL UPITA, KREIRANJE MODELA I PREDIKCIJA POMOĆU BIGQUERY ML

### 5.1. Analiza SQL upita i prikaz rezultata

Prvi SQL upit kroz koji su podaci analizirani se odnosi na broj jedinstvenih valuta. SQL upit je korišćen za prebrojavanje različitih valuta u tabeli. Identifikovane su sve jedinstvene vrednosti u koloni valute, što pomaže u analizi pokrivenosti skupa podataka.

Drugi upit se odnosi na prosečni mesečni kursevi između RSD i EUR. Grupisanjem po godini i mesecu, dobijeni su prosečni kursevi, zaokruženi na dve decimale, sa ciljem praćenja trendova na mesečnom nivou. BigQuery nudi nekoliko različitih mogućnosti prikaza rezultata, među kojima je i tabelarni prikaz na slici 1.

Trećim upitom izvučeni su podaci o dnevnim kursovima RSD, hronološki poredani. Rezultati omogućavaju vremensku analizu promenljivosti kursa i vizualizaciju trendova.

| Row | year | month | avg_exchange_rate |
|-----|------|-------|-------------------|
| 111 | 2024 | 1     | 117.21            |
| 112 | 2024 | 2     | 117.21            |
| 113 | 2024 | 3     | 117.2             |
| 114 | 2024 | 4     | 117.13            |
| 115 | 2024 | 5     | 117.14            |
| 116 | 2024 | 6     | 117.07            |
| 117 | 2024 | 7     | 117.05            |
| 118 | 2024 | 8     | 117.03            |
| 119 | 2024 | 9     | 117.06            |
| 120 | 2024 | 10    | 117.04            |

Slika 1. Tabelarni prikaz rezultata SQL upita

### 5.2. Kreiranje prediktivnog modela pomoću BigQuery ML

BigQuery ML je alat koji omogućava kreiranje i treniranje modela mašinskog učenja direktno u BigQuery-u koristeći SQL upite. Ovo pojednostavljuje proces primene mašinskog učenja, jer korisnici mogu raditi sa SQL-om, koji je već poznat mnogim analitičarima i programerima [4]. Prednost BigQuery ML-a leži u mogućnosti direktne primene na postojeće podatke u BigQuery-u, čime se izbegava potreba za prenosom podataka u drugo okruženje. Kreiran je model predikcije za kurs američkog dolara (USD) koristeći vremenske serije ARIMA modela. Model predviđa kurs USD za narednih 15 dana i evaluira se na osnovu podataka u BigQuery bazi podataka.

ARIMA model (Autoregressive Integrated Moving Average) koristi se za analizu i predikciju vremenskih serija, posebno onih koje nisu stacionarne [5]. Cilj ARIMA modela je da pronađe što precizniji matematički opis podataka, što omogućava bolje razumevanje serije i tačnije predviđanje budućih vrednosti.

U ovom procesu je kreiran ARIMA model za predikciju kursa američkog dolara na osnovu podataka o dnevnim kursovima od 1. januara 2016. do 1. septembra 2024. Prvo je postavljen model u BigQuery koristeći podatke iz tabele koja sadrži dnevne kurseve. Parametri modela su automatski podešeni kako bi optimizovali predikciju.

Zatim je upotrebljena funkcija ML.FORECAST za generisanje predikcija kursa na sledećih 15 dana sa 80% nivoa poverenja.

Nakon što su predikcije izvršene, model je evaluiran koristeći stvarne podatke od 1. do 15. septembra 2024. kako bi se proverila tačnost predikcija. Rezultati evaluacije su pokazali različite metrike greške, kao što su MAE, MSE, RMSE, MAPE i SMAPE, što je omogućilo analizu preciznosti modela. Iako su vrednosti MAE i RMSE bile relativno niske, visoki procenti greške (MAPE i SMAPE) ukazali su na to da model može imati poteškoće sa sezonskim promenama ili neredovnim podacima.

### 5.3. Identifikacija vezanih valuta

U sljedećem delu analize, identifikovane su valute vezane za evro iz posmatranog skupa podataka. Valutni kursevi mogu biti fiksni ili promenljivi. Za identifikaciju vezanih valuta posmatrani su podaci sa istorijom valutnih kurseva pre 2015. godine i izračunat je koeficijent varijacije (CV), koji pokazuje stabilnost kursa, pomoću formule (1):

$$CV = \frac{\sigma}{\mu} \quad (1)$$

Početna analiza se bazira na filtriranju valuta koje su ušle u skup podataka pre 2015. godine, jer se pretpostavlja da su ove valute stabilnije, imaju dužu istoriju i mogu biti pogodnije za dalju analizu. SQL upit je korišćen da bi se identifikovale valute koje su prvi put prikazane u skupu podataka pre 1. januara 2015. Na osnovu ovog upita, dobijeno je 134 valuta koje zadovoljavaju ovaj kriterijum.

Za svaku od identifikovanih valuta, izračunata je prosečna vrednost valutnog kursa, kao i standardna devijacija koja pokazuje promenljivost kursa tokom vremena. Ovo je ključno za razumevanje koliko se kurs jedne valute menja u odnosu na evro. Prosečna vrednost kursa daje osnovnu vrednost za analizu, dok standardna devijacija pomaže u oceni stabilnosti kursa.

Nakon što su izračunati srednji kurs i standardna devijacija, koristi se koeficijent varijacije (CV) kao kriterijum za identifikaciju vezanih valuta. Valute sa niskim CV (manje od 0.03) smatraju se stabilnim, što znači da njihovi kursevi imaju malu varijaciju oko prosečne vrednosti, što može ukazivati na to da su vezane za evro. Na osnovu ovog kriterijuma, identifikovane su valute sa malim CV, koje su potencijalno vezane za evro.

Nakon identifikacije vezanih valuta pomoću SQL upita i analize CV, rezultati su upoređeni sa informacijama sa Wikipedije o međunarodnom statusu i upotrebi evra, koje su dostupne na sljedećem linku:

[https://en.wikipedia.org/wiki/International\\_status\\_and\\_age\\_of\\_the\\_euro#Pegged\\_currencies](https://en.wikipedia.org/wiki/International_status_and_age_of_the_euro#Pegged_currencies)

Valute koje su tačno identifikovane kao vezane za evro su BAM, BGN, DKK, KMF, MAD, MKD i XAF. Takođe su identifikovane valute koje nisu bile prisutne u analizi, kao što su CVE, STN, XOF i XPF, pri čemu su valute poput STN-a bile izostavljene zbog nedostatka podataka, dok su ostale dodane u poslednjem ažuriranju podataka krajem 2023. godine. Postoje i dodatne valute koje se nalaze u listi

identifikovanih vezanih valuta, ali nisu prisutne na Wikipediji: RSD, HRK i XDR.

## 6. ZAKLJUČAK

Ovaj rad je uspešno pokazao povezanost Cloud computing i Big Data koncepta, kao i primenu cloud tehnologija za analizu i predikciju tržišta valuta. Korišćenjem Google Cloud BigQuery platforme, analizirani su dnevni kursevi valuta, identifikovani dugoročni trendovi i razvijen model za predviđanje vrednosti kurseva. SQL upiti su omogućili analize, dok je ARIMA model delimično uspešno predvideo vrednosti kurseva u određenim periodima što ukazuje na potencijal, ali i na prostor za unapređenje ovog pristupa.

Google BigQuery se ističe u izvođenju kompleksnih analiza, dok za jednostavnije operacije nije najučinkovitiji. Rad otvara mogućnosti za dalja istraživanja, kao što su korišćenje složenijih mašinskih modela i uvođenje ekonomskih faktora za bolje predikcije. Takođe, širenje analize na veće vremenske intervale i druge valute može doprineti boljem razumevanju globalnih finansijskih tokova. Ovaj rad pruža čvrstu osnovu za buduća istraživanja u oblasti cloud tehnologija i analize podataka.

## 7. LITERATURA

- [1] <https://aws.amazon.com/what-is-cloud-computing/> (pristupljeno u septembru 2024.)
- [2] <https://www.pcmag.com/encyclopedia/term/big-data> (pristupljeno u septembru 2024.)
- [3] <https://cloud.google.com/bigquery/docs/introduction> (pristupljeno u septembru 2024.)
- [4] <https://cloud.google.com/bigquery/docs/create-machine-learning-model#sql> (pristupljeno u oktobru 2024.)
- [5] [https://en.wikipedia.org/wiki/Autoregressive\\_integrated\\_moving\\_average](https://en.wikipedia.org/wiki/Autoregressive_integrated_moving_average) (pristupljeno u oktobru 2024.)

### Kratka biografija:



**Bosiljka Todić** je rođena 2000. godine u Bijeljini, Bosna i Hercegovina. Završila je gimnaziju „Vaso Pelagić” u Brčkom, 2019. godine. Fakultet Tehničkih Nauka u Novom Sadu je upisala 2019. godine. Diplomirala je u septembru 2023. godine na smeru Primenjeno softversko inženjerstvo. Uspešno je ispunila sve akademske obaveze i položila sve ispite predviđene master studijskim programom Primenjeno softversko inženjerstvo.

kontakt: bosiljkatodic00@gmail.com