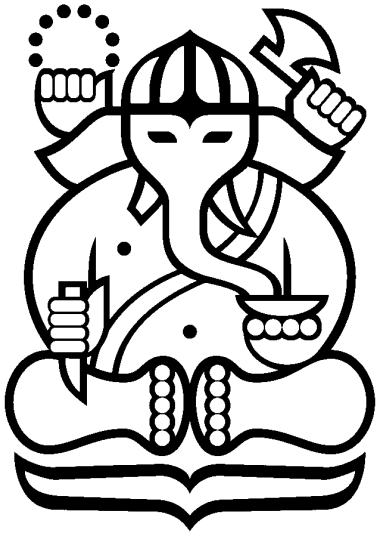
Tugas Proyek Big Data

Laporan Tugas

Dibuat untuk memenuhi nilai
mata kuliah IF4044 Teknologi Big Data



Oleh:

Bryan Bernigen	13520034
Muhammad Alif Putra Yasa	13520135
Ghazian Tsabit Alkamil	13520165

PROGRAM STUDI TEKNIK INFORMATIKA
SEKOLAH TEKNIK ELEKTRO DAN INFORMATIKA
INSTITUT TEKNOLOGI BANDUNG

2022/2023

Latar Belakang

Kripto adalah sebuah mata uang virtual yang keamanannya dijamin dengan kriptografi. Konsep kripto sendiri sudah ada sejak tahun 1983, namun implementasinya secara publik baru dimulai pada tahun 2009 ketika Satoshi Nakamoto meluncurkan bitcoin. Pada awal peluncurannya, bitcoin tidak terlalu mendapat perhatian publik karena barang tersebut belum bisa digunakan oleh publik sebagai alat pembayaran. Pengakuan bitcoin secara masal sendiri baru dimulai pada tahun 2012 ketika beberapa perusahaan ternama mulai mengakui bitcoin sebagai alat pembayaran. Sejak saat itu, bitcoin mulai mendapat perhatian publik.

Semenjak mendapat perhatian publik, bitcoin mulai mengalami pertumbuhan yang pesat hingga harga satu bitcoin mulai menyentuh US$72.000 per koin. Angka tersebut sangat fantastis jika mengingat bahwa pada tahun 2010 harga satu bitcoin hanya sebesar US$0.3 saja. Pertumbuhan tersebut memicu banyak orang untuk melirik kripto sebagai sebuah mata uang yang bernilai. Hal tersebut juga membuat banyak mata uang kripto bermunculan seperti ETH, DOGE, VELO, dan lain-lain.

Adopsi kripto di Indonesia sendiri sudah dimulai pada tahun 2014 ketika perusahaan Indodax mulai berdiri dan mulai memperjualbelikan bitcoin. Saat itu kripto belum terlalu populer sehingga jumlah transaksi harian kripto masih sedikit. Namun dengan berjalannya waktu, semakin banyak orang yang tertarik dengan kripto sehingga pada tahun 2022, Bappebti menyatakan bahwa jumlah investor kripto di Indonesia telah menyentuh angka 16,1 juta jiwa dengan jumlah transaksi melebihi 250 triliun rupiah. Dengan angka tersebut, data kripto menjadi sangat menarik untuk ditambang untuk mendapat insight-insight menarik mengenai keadaan kripto di Indonesia.

Sumber Data

Sumber data yang digunakan pada laporan kali ini adalah transaksi kripto dan bitcoin yang disediakan secara publik oleh Indodax. API tersebut akan mengirimkan data setiap ada transaksi yang dilakukan melalui Indodax. Kami memilih menggunakan sumber data dari Indodax karena Indodax merupakan salah satu platform exchange kripto terbesar di Indonesia sehingga data tersebut dapat merepresentasikan mayoritas transaksi di Indonesia.

Detail mengenai sumber data yang kami gunakan dapat dilihat pada tautan berikut: https://github.com/btcid/indodax-official-api-docs

Detail mengenai API yang kami gunakan untuk mendapatkan data dapat dilihat pada tautan berikut:
https://github.com/btcid/indodax-official-api-docs/blob/master/Marketdata-websocket.md

Data Hasil Pemrosesan

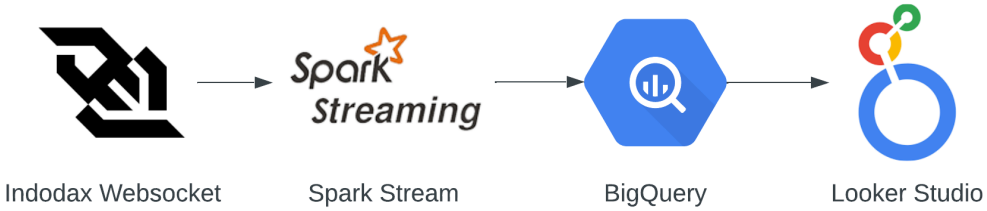
Data source yang digunakan pada proyek ini berupa *websocket* yang tersedia secara publik milik Indodax. *Websocket* yang digunakan melakukan *stream* terhadap *message* yang berisi data kondisi market cryptocurrency dalam rentang waktu 24 jam dan transaksi bitcoin yang terjadi di platform Indodax. Pada proyek ini tidak dilakukan proses transformasi terhadap data yang diperoleh dari sumber data sehingga data yang diterima dari *websocket* dengan data yang disimpan di dalam *data warehouse* akan sama persis. Contoh data yang diperoleh dari *websocket* dapat dilihat pada Gambar 2.1 dan Gambar 2.2.

Tujuan Mining

Terdapat dua tujuan dalam mining data kali ini, yakni mengenai penerapan kripto secara umum di Indonesia dan penerapan bitcoin di Indonesia. Pada mining penerapan kripto secara umum di Indonesia, kami ingin tahu berapa banyak transaksi kripto di Indonesia beserta koin kripto apa yang paling populer. Kami juga ingin tahu mengenai tren transaksi kripto selama 7 hari terakhir.

Pada mining penerapan bitcoin di Indonesia, kami ingin tahu berapa banyak transaksi bitcoin yang dilakukan di Indonesia beserta volume yang ditransaksikan. Kami juga ingin tahu apakah lebih banyak orang menjual atau membeli bitcoin selama 7 hari terakhir. Lalu kami juga ingin mencari berapa harga rata-rata orang menjual dan membeli bitcoin selama 7 hari terakhir. Kami juga ingin melihat tren harga kripto selama 7 hari terakhir.

Arsitektur Umum



Indodax Websocket

Data source yang digunakan pada proyek ini berupa *websocket* yang tersedia secara publik milik Indodax. *Websocket* yang digunakan melakukan *stream* terhadap *message* yang berisi data kondisi market cryptocurrency dalam rentang waktu 24 jam dan transaksi bitcoin yang terjadi di platform Indodax. Salah satu contoh *message* yang didapatkan dari websocket publik Indodax dapat dilihat pada Gambar 2.1 dan 2.2.

```
{
    "result": {
        "channel": "market:summary-24h",
        "data": {
            "data": [
                [
                    "dogeidr",                  // pair
                    1635134109,                 // epoch timestamp in second
                    3810,                       // last price
                    3480,                       // lowest price in the last 24h
                    3980,                       // highest price in the last 24h
                    3523,                       // price at T-24h
                    "112745093944.00000000",    // IDR volume in the last 24h (DOGE/IDR)
                    "30241791.15270789"         // DOGE volume in the last 24h (DOGE/IDR)
                ],
                [
                    "usdtidr",
                    1635134410,
                    14124,
                    14076,
                    14130,
                    14123,
                    "194798674207.00000000",    // IDR volume in the last 24h (USDT/IDR)
                    "13798116.12995762"         // USDT volume in the last 24h (USDT/IDR)
                ]
            ],
            "offset": 2444948
        }
    }
}
```

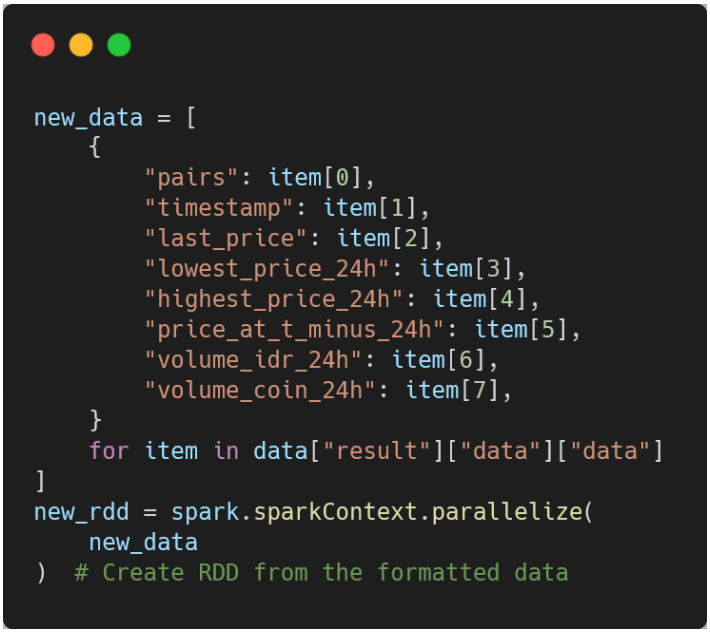
Gambar 2.1 Message Market Summary Cryptocurrency

Gambar 2.2 Message Bitcoin Transaction

Message yang berhasil diperoleh dari websocket akan diproses menggunakan *spark stream* dan kemudian dimasukkan ke dalam Google BigQuery setiap lima menit.

Spark Stream

Spark Stream digunakan untuk memproses *messages* yang diterima dari *websocket*. Setiap kali menerima *messages*, *messages* akan di-format sesuai format field BigQuery dari kedua data.



Gambar 2.3 Message di-format dan diubah menjadi RDD

Setelah di-format, *messages* tersebut ditambahkan ke sebuah RDD. Setiap 5 menit, *messages* yang telah disimpan sebagai sebuah RDD di-insert ke BigQuery.

```python
def insert_rdd_to_bigquery():
    global rdd
    print("Inserting RDD to BigQuery")
    if not rdd.isEmpty():
        rows = rdd.collect()  # Collect RDD data into a list of rows
        insert_into_bigquery(rows)  # Insert all collected rows at once
        rdd = spark.sparkContext.emptyRDD()  # Reset RDD after insertion
```

Gambar 2.4 Memasukkan data di RDD ke BigQuery

Google BigQuery

Pada proyek ini, Google BigQuery digunakan sebagai tempat penyimpanan data tujuan atau *data warehouse*. Pada proyek ini dibuat sebuah data set dengan nama *crypto_transaction_indodax* yang memiliki dua tabel, tabel *btc_transaction* dan tabel *crypto-transaction*. Data yang disimpan pada kedua tabel tersebut sama persis dengan *message* yang didapatkan dari *websocket* yang digunakan, yang berarti pada proyek ini tidak dilakukan proses transformasi data. Deskripsi tabel *btc_transaction* dan *crypto-transaction* dapat dilihat pada Gambar 2.5 dan 2.6.

Field name	Type	Mode
pairs	STRING	NULLABLE
timestamp	TIMESTAMP	NULLABLE
seq_number	INTEGER	NULLABLE
action	STRING	NULLABLE
price	FLOAT	NULLABLE
volume_idr	FLOAT	NULLABLE
volume_btc	FLOAT	NULLABLE

Gambar 2.5 Deskripsi Tabel *btc_transaction*

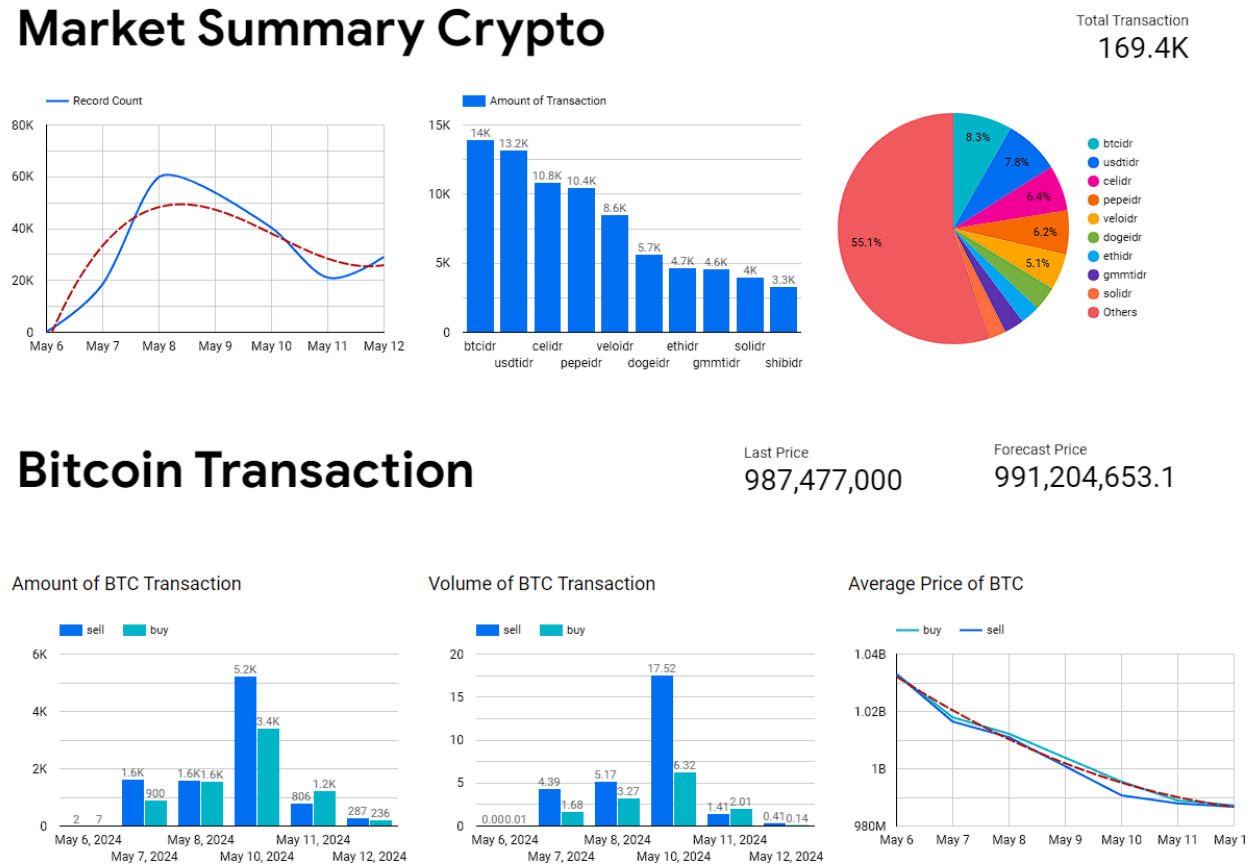
Field name	Type	Mode
pairs	STRING	NULLABLE
timestamp	TIMESTAMP	NULLABLE
last_price	FLOAT	NULLABLE
lowest_price_24h	FLOAT	NULLABLE
highest_price_24h	FLOAT	NULLABLE
price_at_t_minus_24h	FLOAT	NULLABLE
volume_idr_24h	FLOAT	NULLABLE
volume_coin_24h	FLOAT	NULLABLE

Gambar 2.6 Deskripsi Tabel *crypto-transaction*

Data yang berhasil disimpan di dalam Google BigQuery kemudian akan divisualisasikan menggunakan *Looker Studio*.

Looker Studio

Looker studio digunakan untuk memvisualisasi data yang telah dimasukkan ke dalam Google BigQuery. Kami memilih untuk menggunakan Looker studio karena looker studio sudah terkoneksi secara native ke Google BigQuery. Alasan lain kami memilih looker studio adalah karena looker studio dan Google BigQuery sama-sama merupakan produk dari GCP sehingga koneksi, setup, pembayaran, dan masalah-masalah lainnya akan lebih mudah untuk diselesaikan karena keduanya terhubung dalam satu platform dengan satu akun yang sama. Dashboard yang kami visualisasikan dapat dilihat pada gambar 2.7



Gambar 2.7 Dashboard Visualisasi Kripto

Dashboard pada gambar 2.7 digunakan untuk memvisualisasikan dua hal, yakni market kripto secara keseluruhan dan market bitcoin saja. Pada market keseluruhan, terdapat tiga visualisasi, yakni jumlah transaksi per hari untuk segala macam koin, jumlah transaksi untuk 10 koin paling populer, dan market share dari keseluruhan transaksi. Terdapat juga tren transaksi per hari yang direpresentasikan dengan garis merah putus-putus pada visualisasi jumlah transaksi per hari.

Pada visualisasi market bitcoin, terdapat beberapa hal yang divisualisasikan, yakni jumlah transaksi, volume yang ditransaksikan, harga per hari, harga terakhir, dan harga yang diprediksi. Visualisasi jumlah transaksi, volume transaksi, dan harga per hari diterapkan untuk pembelian dan penjualan. Harga terakhir dan harga prediksi hanya ditampilkan untuk pembelian.

Machine Learning and Forecasting

Pada proyek ini juga dilakukan pembuatan model machine learning untuk melakukan forecasting. Model yang dibuat ditujukan untuk memprediksi harga beli bitcoin pada jangka waktu tertentu. Model tersebut dibuat menggunakan bigquery dengan metode ARIMA. Berikut syntax yang digunakan untuk mentrain model.

```
CREATE OR REPLACE MODEL`crypto_transaction_indodax.btc_buy_forecasting`
OPTIONS
  (model_type = 'ARIMA_PLUS',
   time_series_timestamp_col = 'timekey',
   time_series_data_col = 'price',
   auto_arima = TRUE,
   data_frequency = 'AUTO_FREQUENCY',
   decompose_time_series = TRUE
  ) AS
SELECT
  TIMESTAMP_SECONDS(15*60 * DIV(UNIX_SECONDS(timestamp), 15*60)) AS timekey,
  AVG(price) AS price
FROM `tubes-big-data-422412.crypto_transaction_indodax.btc_transaction` AS d
WHERE action='buy'
GROUP BY timekey
```

Gambar 2.8 Syntax untuk melakukan training model

Model tersebut lalu digunakan untuk memprediksi harga bitcoin untuk beberapa jam kedepan. Hasil tersebut forecast tersebut ditampilkan pada dashboard.

Lessons Learned

Berdasarkan pengerjaan tugas besar teknologi big data ini, terdapat beberapa poin *lessons learned* sebagai berikut:

1. Digunakannya sumber data *websocket* yang sudah melakukan stream data akan mempermudah pengerjaan tugas besar karena tidak diperlukannya lagi setup untuk kafka.
2. Diperlukan proses autentikasi secara berkala agar dapat terhubung ke websocket Indodax yang dijadikan sumber data.
3. Bigquery sebagai solusi data warehouse yang ditawarkan oleh Google memiliki satu kelebihan utama dibandingkan dengan data warehouse lainnya, yakni kemampuannya untuk terintegrasi secara native dengan layanan-layananan Google lainnya seperti looker studio untuk memvisualisasikan data ataupun Bigquery ML untuk melakukan training model.
4. Looker studio sebagai tools visualisasi memiliki beberapa kelebihan dibandingkan tools visualisasi lainnya seperti grafana ataupun metabase. Kelebihan tersebut terdapat pada user interface looker studio yang dapat digunakan oleh user non teknis karena visualisasi dilakukan melalui GUI. Hal tersebut berbanding terbalik dengan beberapa tools visualisasi lainnya yang menggunakan query sehingga perlu pemahaman mengenai SQL untuk menggunakan tools tersebut. Looker studio juga terkoneksi dengan BigQuery secara native sehingga tidak perlu tambahan library ataupun connector lainnya.
5. Bigquery ML menawarkan layanan training model secara mudah, namun model yang dihasilkan kurang lengkap dibandingkan training sendiri. Model yang dihasilkan oleh Bigquery ML tidak memiliki informasi mengenai precision ataupun recall yang umumnya digunakan sebagai parameter keberhasilan suatu model. Bigquery ML juga tidak memberikan fleksibilitas mengenai hyperparameter tuning ataupun tuning model lainnya.

Lampiran

Github: https://github.com/ZianTsabit/tubes-big-data

Looker Studio: https://lookerstudio.google.com/u/0/reporting/27424802-cd57-4261-8d45-d06d63a93ebf/page/tEnnC