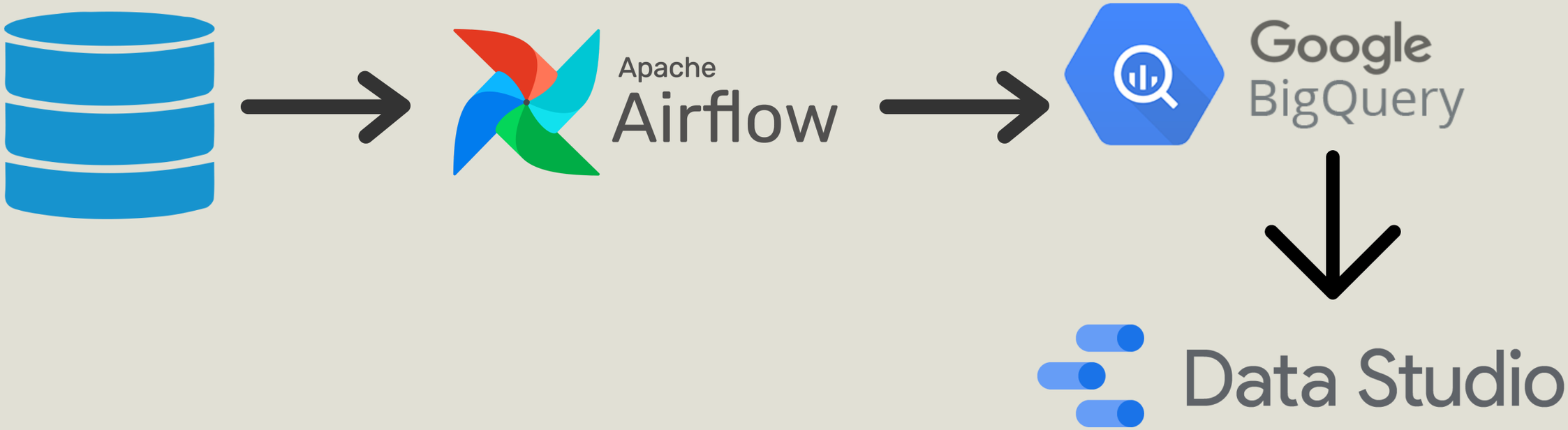
Final Project

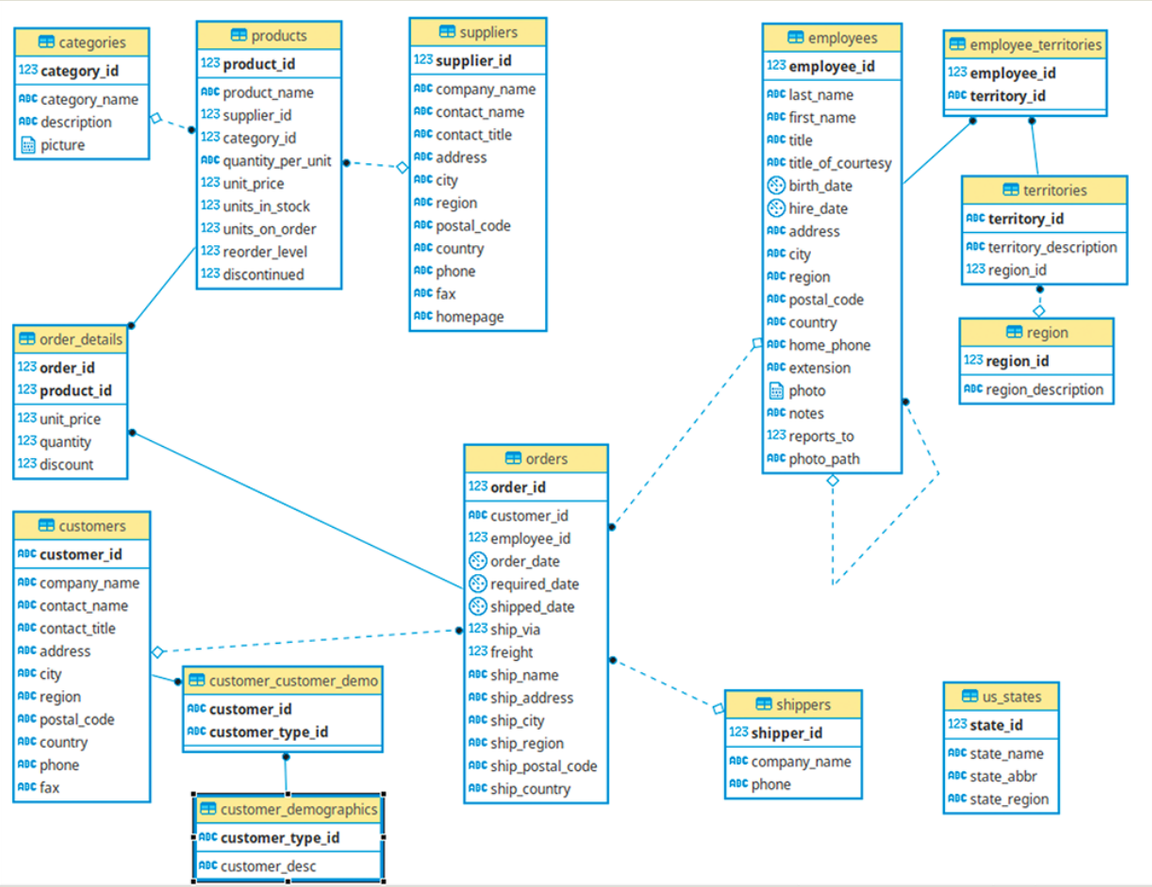
Hilda Meiranita Prastika Dewi

Arsitektur end to end dari source (postgresql) hingga ke datawarehouse(BigQuery)



Arsitektur end to end dari source (postgresql) hingga ke datawarehouse(BigQuery)

Sumber data: northwind database



Setting github repository

- Fork repository: https://github.com/saungkertas/digitalskola_dataops

Membuat file Ingest

Ingest_orders

```
sql = """select order_id, customer_id, employee_id, order_date, required_date, shipped_date,
ship_country from orders o where cast(order_date as date) = '"""+sys.argv[1]+"""'"""
csv_file_path = '/root/output/hilda/orders/orders_'+sys.argv[1]+'.csv'
```

Ingest_order_details

```
sql = """select order_id, product_id, unit_price, quantity, discount from order_details where order_id in (
    select order_id from orders where cast(order_date as date) = '"""+sys.argv[1]+"""')"""
csv_file_path = '/root/output/hilda/order_details/order_details_'+sys.argv[1]+'.csv'
```

Ingest_categories

```
sql = """select category_id, category_name, description, picture from categories"""
csv_file_path = '/root/output/hilda/categories/categories_'+sys.argv[1]+'.csv'
```

Ingest_customers

```
sql = """select customer_id, company_name, contact_name, contact_title, address, city,
region, postal_code, country, phone, fax from customers"""
csv_file_path = '/root/output/hilda/customers/customers_'+sys.argv[1]+'.csv'
```

DigitalSkola

Arsitektur end to end dari source (postgresql) hingga ke datawarehouse(BigQuery)

Membuat file Ingest

Ingest_products

```
sql = """select p.product_id,
    p.product_name,
    s.supplier_id,
    c.category_id,
    p.quantity_per_unit,
    p.unit_price,
    p.units_in_stock,
    p.units_on_order,
    p.reorder_level,
    p.discontinued
    from products p
    join suppliers s on p.supplier_id = s.supplier_id
    join categories c on p.category_id = c.category_id
    where product_id in (select product_id from order_details od where od.order_id in
    (select order_id from orders o where cast(o.order_date as date) = '"""+sys.argv[1]+"""'))"""
csv_file_path = '/root/output/hilda/products/products_'+sys.argv[1]+'.csv'
```

Ingest_suppliers

```
sql = """select supplier_id, company_name, contact_name, contact_title, address, city,
region, postal_code, country, phone, fax, homepage from suppliers"""
csv_file_path = '/root/output/hilda/suppliers/suppliers_'+sys.argv[1]+'.csv'
```

Membuat file Init

```
with DAG('init_hilda',
    schedule_interval="@once",
    start_date=datetime(2022, 7, 6)
) as dag:

    start = DummyOperator(
        task_id='start'
    )
    #orders
    ingest_orders = BashOperator(
        task_id='ingest_orders',
        bash_command="""python3 /root/airflow/dags/ingest/hilda/ingest_orders.py {{ execution_date.format('YYYY-MM-DD') }}"""
    )
    to_datalake_orders = BashOperator(
        task_id='to_datalake_orders',
        bash_command="""gsutil cp /root/output/hilda/orders/orders_{{ execution_date.format('YYYY-MM-DD') }}.csv gs://digitalskola-de-batch7/hilda/staging/orders/"""
    )
    data_definition_orders = BashOperator(
        task_id='data_definition_orders',
        bash_command="""bq mkdef --autodetect --source_format=CSV gs://digitalskola-de-batch7/hilda/staging/orders/* > /root/table_def/hilda/orders.def"""
    )
    to_dwh_orders = BashOperator(
        task_id='to_dwh_orders',
        bash_command="""bq mk --external_table_definition=/root/table_def/hilda/orders.def de_7.hilda_orders"""
    )
```

```
start >> ingest_orders >> to_datalake_orders >> data_definition_orders >> to_dwh_orders
start >> ingest_order_details >> to_datalake_order_details >> data_definition_order_details >> to_dwh_order_details
start >> ingest_products >> to_datalake_products >> data_definition_products >> to_dwh_products
start >> ingest_customers >> to_datalake_customers >> data_definition_customers >> to_dwh_customers
start >> ingest_suppliers >> to_datalake_suppliers >> data_definition_suppliers >> to_dwh_suppliers
start >> ingest_categories >> to_datalake_categories >> data_definition_categories >> to_dwh_categories
```

DigitalSkola

Arsitektur end to end dari source (postgresql) hingga ke datawarehouse(BigQuery)

Scheduler menggunakan airflow untuk ingest data tiap hari ke datalake

Membuat file daily

```
with DAG('daily_hilda',
    schedule_interval='0 0 * * *',
    start_date=datetime(2022, 7, 1)
) as dag:

    start = DummyOperator(
        task_id='start'
    )

    #orders
    ingest_orders = BashOperator(
        task_id='ingest_orders',
        bash_command="""python3 /root/airflow/dags/ingest/hilda/ingest_orders.py {{ execution_date.format('YYYY-MM-DD') }}"""
    )

    to_datalake_orders = BashOperator(
        task_id='to_datalake_orders',
        bash_command="""gsutil cp /root/output/hilda/orders/orders_{{ execution_date.format('YYYY-MM-DD') }}.csv gs://digitalskola-de-batch7/hilda/staging/orders/"""
    )

    #order_details
    ingest_order_details = BashOperator(
        task_id='ingest_order_details',
        bash_command="""python3 /root/airflow/dags/ingest/hilda/ingest_order_details.py {{ execution_date.format('YYYY-MM-DD') }}"""
    )
    to_datalake_order_details = BashOperator(
        task_id='to_datalake_order_details',
        bash_command="""gsutil cp /root/output/hilda/order_details/order_details_{{ execution_date.format('YYYY-MM-DD') }}.csv gs://digitalskola-de-batch7/hilda/staging/order_details/"""
    )

    start >> ingest_orders >> to_datalake_orders
    start >> ingest_order_details >> to_datalake_order_details
```

Membuat directory untuk output

```
root@de7-final-project:~/output/hilda# ls
categories  customers  order_details  orders  products  suppliers
```

Membuat directory untuk table_def

```
root@de7-final-project:~/table_def/hilda# ls
categories.def  customers.def  order_details.def  orders.def  products.def  suppliers.def
```

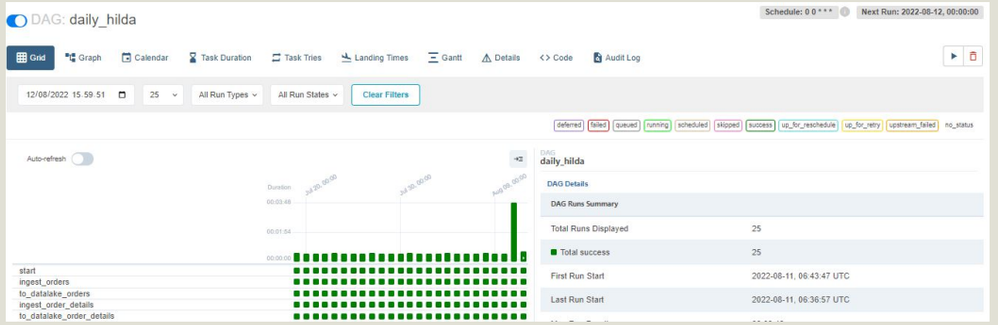
Mengupload file ingest, daily, dan init pada github dan commit untuk pull request

Menginput file ke airflow

```
root@de7-final-project:~/airflow/dags/ingest/hilda# ls
ingest_categories.py  ingest_order_details.py  ingest_products.py
ingest_customers.py   ingest_orders.py         ingest_suppliers.py
```

File dag daily dan init pada folder airflow/dags/

Daily pada airflow



Arsitektur end to end dari source (postgresql) hingga ke datawarehouse(BigQuery)

Membuat datamart view menggunakan BigQuery

Gross revenue total per hari

```
SELECT EXTRACT(MONTH FROM order_date) as month,
      o.order_date,
      SUM((1-ood.discount) * ood.unit_price * ood.quantity) AS gross_revenue
FROM `data-engineering-2-329815.de_7.hilda_orders` AS o
JOIN `data-engineering-2-329815.de_7.hilda_order_order_details` AS ood
ON o.order_id = ood.order_id
GROUP BY o.order_date
ORDER BY o.order_date
```

Gross revenue per product per bulan

```
SELECT
      EXTRACT(MONTH FROM order_date) as month,
      p.product_name,
      SUM((1-ood.discount) * ood.unit_price * ood.quantity) AS gross_revenue
FROM `data-engineering-2-329815.de_7.hilda_orders` AS o
JOIN `data-engineering-2-329815.de_7.hilda_order_order_details` AS ood
ON o.order_id = ood.order_id
JOIN `data-engineering-2-329815.de_7.hilda_products` AS p
ON ood.product_id = p.product_id
GROUP BY month, p.product_name
ORDER BY month, gross_revenue
```

Jumlah total pembelian per product per bulan

```
SELECT
      EXTRACT(MONTH FROM order_date) as month,
      p.product_name,
      SUM(ood.quantity) AS total_purchase
FROM `data-engineering-2-329815.de_7.hilda_orders` AS o
JOIN `data-engineering-2-329815.de_7.hilda_order_order_details` AS ood
ON o.order_id = ood.order_id
JOIN `data-engineering-2-329815.de_7.hilda_products` AS p
ON ood.product_id = p.product_id
GROUP BY month, p.product_name
ORDER BY  month, total_purchase DESC
```

Jumlah total pembelian per kategori product per bulan

```
SELECT
      EXTRACT(MONTH FROM order_date) as month,
      c.category_name,
      SUM(ood.quantity) AS total_purchase
FROM `data-engineering-2-329815.de_7.hilda_orders` AS o
JOIN `data-engineering-2-329815.de_7.hilda_order_order_details` AS ood
ON o.order_id = ood.order_id
JOIN `data-engineering-2-329815.de_7.hilda_products` AS p
ON ood.product_id = p.product_id
JOIN `data-engineering-2-329815.de_7.hilda_categories` AS c
ON p.category_id = c.category_id
GROUP BY month, c.category_name
ORDER BY total_purchase DESC
```

Jumlah total pembelian per negara per bulan

```
SELECT
      EXTRACT(MONTH FROM order_date) as month,
      o.ship_country AS country,
      SUM(ood.quantity) AS total_purchase
FROM `data-engineering-2-329815.de_7.hilda_orders` AS o
JOIN `data-engineering-2-329815.de_7.hilda_order_order_details` AS ood
ON o.order_id = ood.order_id
GROUP BY month, country
ORDER BY month,total_purchase
```

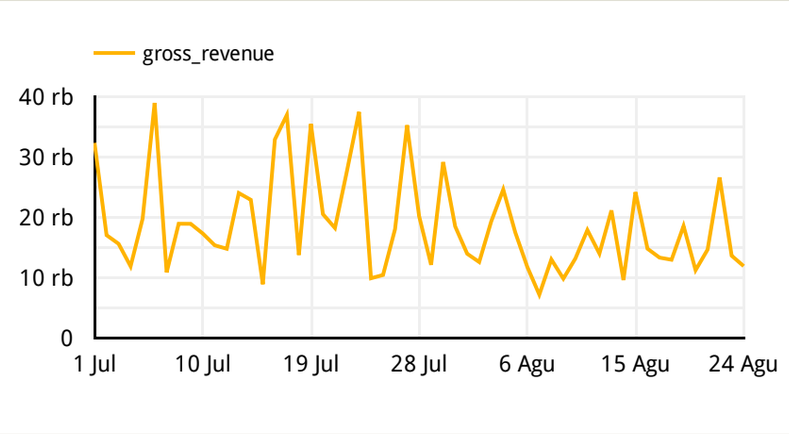
DigitalSkola

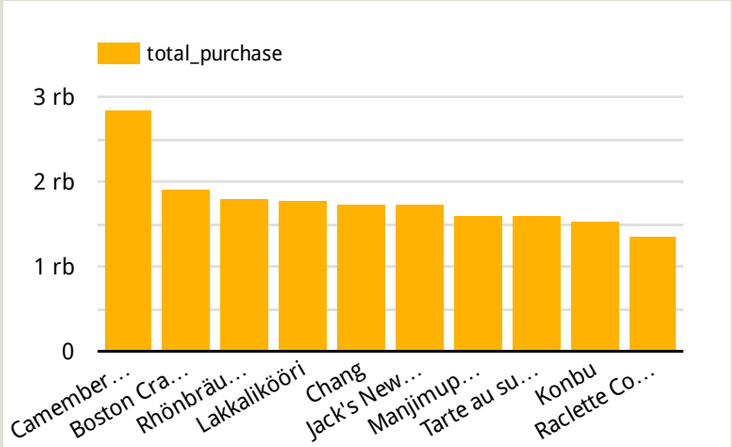
SALES DASHBOARD

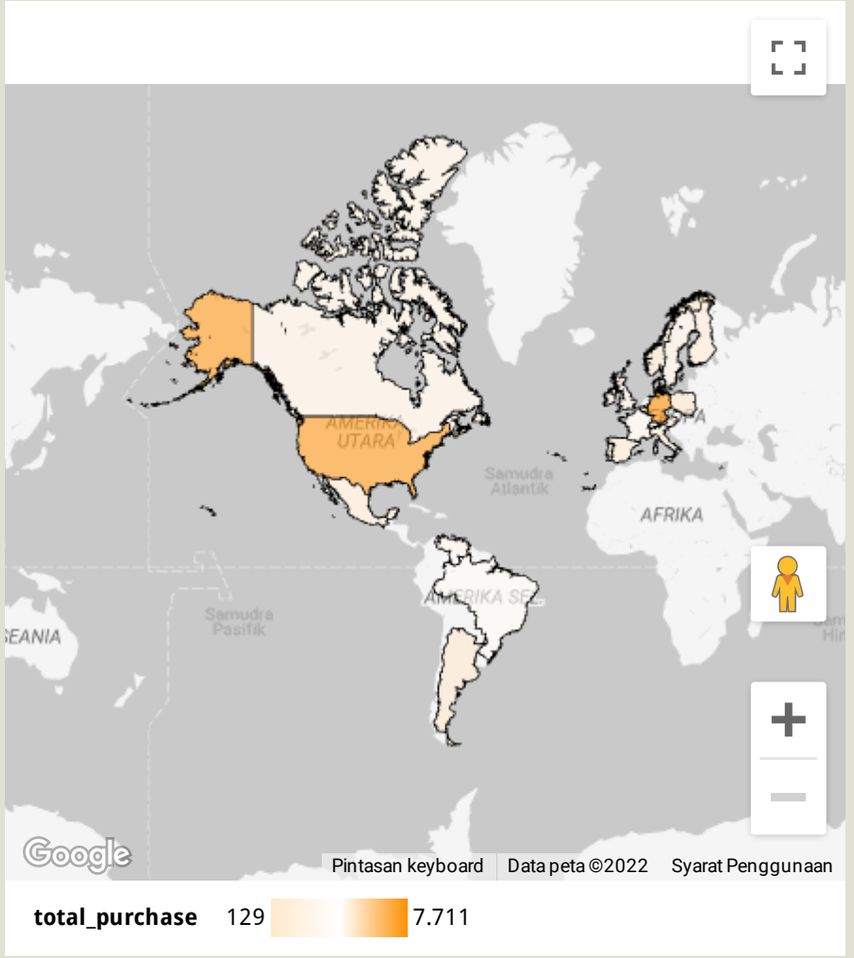
gross_revenue
1,0 jt

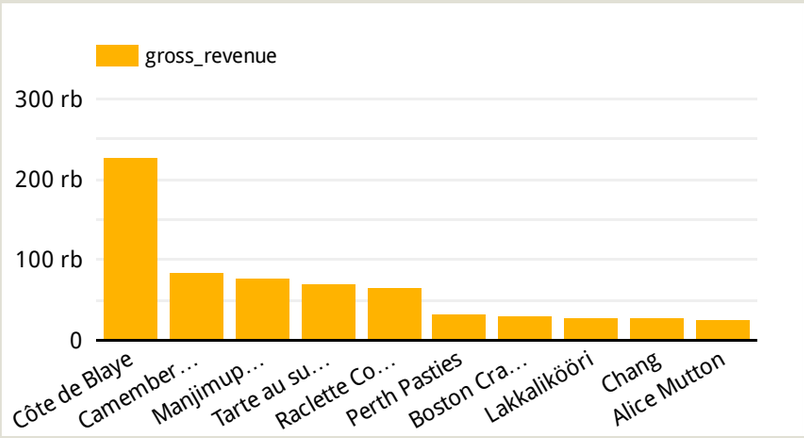
total_purchase
45,7 rb

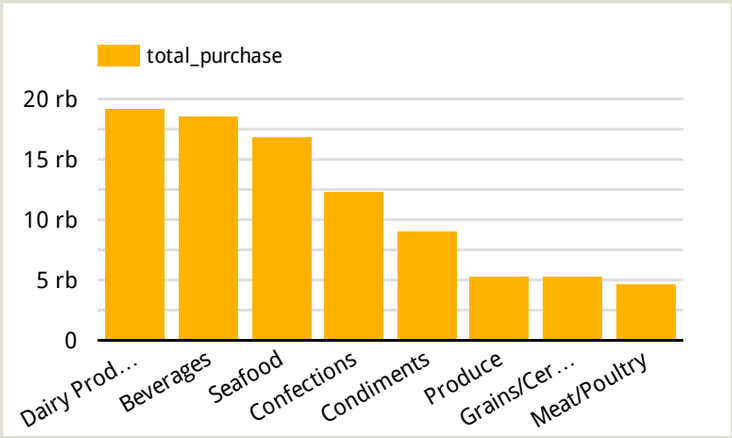
month











DigitalSkola

Terima kasih