In [1]:

```
import pandas as pd
from sqlalchemy import create_engine
import mysql.connector
from mysql.connector import Error

# Connect to Database
def connect_to_database():
    # Connect to the MySQL database
    try:
        db_connection = mysql.connector.connect(
            host="127.0.0.1",
            user="root",
            password="0000",  # Add your password here
            database="e-commerce-dwh"
        )
    except mysql.connector.Error as err:
        print("Error connecting to database:", err)
        exit()  # Terminate script if connection fails
    return db_connection

# Fetch data from rawdata table
def fetch_data(db_connection,query):
    cursor = db_connection.cursor(dictionary=True)
    cursor.execute(query)
    result = cursor.fetchall()
    return pd.DataFrame(result)

connection = connect_to_database()
```

Business Questions

When is the peak season of our ecommerce ?

In [10]:

```
query = '''
SELECT
    d.month_name AS peak_month,
    d.year,
    COUNT(o.order_key) AS total_orders,
    SUM(o.price) AS total_revenue
FROM
    fact_order o
JOIN
    dim_date d ON o.order_date_id = d.date_id
GROUP BY
    d.year,d.month_name
ORDER BY
    total_orders desc, total_revenue desc
limit 5;
'''
fetch_data(connection,query)
```

Out[10]:

	peak_month	year	total_orders	total_revenue
0	November	2017	8665	1010271370.00
1	March	2018	8217	983213440.00
2	January	2018	8208	950030360.00
3	April	2018	7975	996647750.00
4	May	2018	7925	996517680.00

Our Peak is November due to black friday which is highest in revenue and ordes numbers Then March and January for 2018 Year

What time users are most likely make an order or using the ecommerce app?

In [ ]:

```
query_time = '''
SELECT
    d.hour,
    COUNT(o.order_key) AS total_orders
FROM
    fact_order o
JOIN
    dim_date d ON o.order_date_id = d.date_id
```

```
GROUP BY
    d.hour
ORDER BY
    total_orders DESC
Limit 5;
'''
fetch_data(connection, query_time)
```

Out[ ]:

	hour	total_orders
0	16	7653
1	14	7565
2	11	7432
3	13	7403
4	15	7370

Seems Like 4 in the afternoon is our Peak time also all afternoon times is the are most times to make orders

What is the preferred way to pay in the ecommerce?

In [25]:

```
query_payment = '''
with payment_method as (
    SELECT
        p.payment_type,
        COUNT(o.order_key) AS total_orders
    FROM
        fact_order o
    JOIN
        dim_payment p ON o.payment_id = p.payment_id
    GROUP BY
        p.payment_type
    )

SELECT payment_type, total_orders,round(total_orders/(SELECT SUM(total_orders) FROM payment_method) * 100,1) as perce
From payment_method
Order by total_orders desc
```

```
'''
fetch_data(connection, query_payment)
```

Out[25]:

	payment_type	total_orders	percentage
0	credit_card	85030	75.5
1	blipay	22867	20.3
2	voucher	3060	2.7
3	debit_card	1690	1.5

Seems Like People prefer credit card 75% of people prefer this payment method

How many installment is usually done when paying in the ecommerce?

In [35]:

```
query_installments = '''
with payment_installments_counnt as (
SELECT
    payment_installments,
    COUNT(o.order_key) AS total_orders
FROM
    fact_order o
JOIN
    dim_payment p ON o.payment_id = p.payment_id
GROUP BY
    payment_installments)

SELECT payment_installments, total_orders, total_orders/sum(total_orders) over() as percentage
FROM payment_installments_counnt
ORDER BY total_orders DESC
LIMIT 5;
'''
fetch_data(connection, query_installments)
```

Out[35]:

	payment_installments	total_orders	percentage
0	1	54357	0.4825
1	2	13548	0.1203
2	3	11631	0.1033
3	4	7896	0.0701
4	10	6763	0.0600

Seems like most people pay all in first time for 50 of purchases and almost 1 percentage of people tend to long term installements

What is the average spending time for user for our ecommerce?

In [50]:

```
query_avg_spending_time = '''
SELECT
    AVG(
        CASE
            WHEN d2.hour > d1.hour THEN (d2.hour - d1.hour) * 60 + (d2.minute - d1.minute)
            ELSE d2.minute - d1.minute
        END
    ) AS avg_decision_time_minutes
FROM
    fact_order o
JOIN
    dim_date d1 ON o.order_date_id = d1.date_id
JOIN
    dim_date d2 ON o.order_approved_date_id = d2.date_id;
'''
fetch_data(connection, query_avg_spending_time)
```

Out[50]:

	avg_decision_time_minutes
0	31.3784

average time only was to see the customer time between the approval date and order date

What is the frequency of purchase on each state?

In [ ]:

```
query_state_frequency = '''
    SELECT
        customer_state,
        Round(AVG(total_orders)) AS avg_orders_per_month
    FROM (
        SELECT
            u.customer_state,
            COUNT(o.order_key) AS total_orders
        FROM
            fact_order o
        JOIN
            dim_user u ON o.user_id = u.user_id
        JOIN
            dim_date d ON o.order_date_id = d.date_id
        GROUP BY
            u.customer_state, d.month_name
    ) AS state_orders
    GROUP BY
        customer_state
    ORDER BY
        avg_orders_per_month DESC
    LIMIT 5;
    '''
fetch_data(connection, query_state_frequency)
```

Out[ ]:

	customer_state	avg_orders_per_month
0	BANTEN	2017
1	JAWA BARAT	1200
2	DKI JAKARTA	1196
3	JAWA TENGAH	806
4	JAWA TIMUR	791

Banten state got the highest month order with 2017 order per month then jawa barat with 800 less

Which logistic route that have heavy traffic in our ecommerce?

In [79]:

```
query_rush_hour = '''
WITH RankedOrders AS (
    SELECT
        s.seller_city AS pickup_city,
        s.seller_state AS pickup_state,
        u.customer_city AS delivery_city,
        u.customer_state AS delivery_state,
        d.hour,
        COUNT(o.order_key) AS order_count,
        ROW_NUMBER() OVER (
            PARTITION BY
                s.seller_city, s.seller_state, u.customer_city, u.customer_state
            ORDER BY
                COUNT(o.order_key) DESC
        ) AS rank_
    FROM
        fact_order o
    JOIN
        dim_seller s ON o.seller_id = s.seller_id
    JOIN
        dim_user u ON o.user_id = u.user_id
    JOIN
        dim_date d ON o.delivered_date_id = d.date_id
    GROUP BY
        s.seller_city, s.seller_state, u.customer_city, u.customer_state, d.hour
)
SELECT
    pickup_city,
    pickup_state,
    delivery_city,
    delivery_state,
    hour AS max_order_hour,
    order_count
FROM
    RankedOrders
WHERE
    rank_ = 1
```

```
ORDER BY
    order_count DESC
LIMIT 10;
'''
fetch_data(connection, query_rush_hour)
```

Out[79]:

	pickup_city	pickup_state	delivery_city	delivery_state	max_order_hour	order_count
0	KOTA TANGERANG	BANTEN	KOTA TANGERANG	BANTEN	18	184
1	KABUPATEN BERAU	KALIMANTAN TIMUR	KOTA TANGERANG	BANTEN	18	57
2	KOTA TANGERANG	BANTEN	KOTA JAKARTA BARAT	DKI JAKARTA	17	52
3	KOTA JAKARTA SELATAN	DKI JAKARTA	KOTA TANGERANG	BANTEN	18	48
4	KOTA JAKARTA TIMUR	DKI JAKARTA	KOTA TANGERANG	BANTEN	20	47
5	KOTA TANGERANG	BANTEN	KOTA JAKARTA TIMUR	DKI JAKARTA	16	44
6	KOTA TANGERANG	BANTEN	KABUPATEN PURBALINGGA	JAWA TENGAH	16	40
7	KOTA TANGERANG	BANTEN	KOTA PROBOLINGGO	JAWA TIMUR	17	39
8	KOTA TANGERANG	BANTEN	KOTA BONTANG	KALIMANTAN TIMUR	16	33
9	KOTA JAKARTA BARAT	DKI JAKARTA	KOTA TANGERANG	BANTEN	18	33

This query shows the hour where most orders are sent which will be more traffic

How many late delivered order in our ecommerce? Are late order affecting the customer satisfaction?

In [ ]:

```
query_late_orders = '''
WITH late_orders AS (
    SELECT
        o.order_key,
        CASE
            WHEN date(d2.full_timestamp) > o.estimated_time_delivery THEN 1
            ELSE 0
        END AS is_late
    FROM
        fact_order o
    JOIN
        dim_date d2 ON o.delivered_date_id = d2.date_id
)
SELECT
    is_late,
    COUNT(o.order_key) AS total_orders,
    AVG(f.feedback_score) AS avg_satisfaction
FROM
    late_orders lo
JOIN
    fact_order o ON lo.order_key = o.order_key
JOIN
    dim_feedback f ON o.feedback_id = f.feedback_key
GROUP BY
    is_late;
'''
fetch_data(connection, query_late_orders)
```

Out[ ]:

	is_late	total_orders	avg_satisfaction
0	0	102931	4.1948
1	1	7265	2.2434

Late arrival got a big infelunce in the satisfication rate for customers

How long are the delay for delivery / shipping process in each state?

In [ ]:

```
query_delay_per_state = '''
WITH delay AS (
    SELECT
        u.customer_state,
        DATEDIFF(d2.full_timestamp, d1.full_timestamp) AS delay_days
    FROM
        fact_order o
    JOIN
        dim_user u ON o.user_id = u.user_id
    JOIN
        dim_date d2 ON o.delivered_date_id = d2.date_id
    JOIN
        dim_date d1 ON o.pickup_date_id = d1.date_id
    WHERE
        DATEDIFF(d2.full_timestamp, o.estimated_time_delivery) > 0
)
SELECT
    customer_state,
    AVG(delay_days) AS avg_delay_days
FROM
    delay
GROUP BY
    customer_state
ORDER BY
    avg_delay_days
    limit 5;
'''
fetch_data(connection, query_delay_per_state)
```

Out[ ]:

	customer_state	avg_delay_days
0	BANTEN	18.5260
1	KALIMANTAN TIMUR	26.0625
2	JAWA TENGAH	26.1481
3	DI YOGYAKARTA	26.6345
4	SULAWESI TENGAH	26.7945

The states with lowest delay days between pick up and delivery are higher in number of orders count

How long are the difference between estimated delivery time and actual delivery time in each state?

In [ ]:

```
query_delivery_time_diff = '''
WITH delivery_diff AS (
    SELECT
```

```
        u.customer_state,
        DATEDIFF(o.estimated_time_delivery,d2.full_timestamp) AS delivery_time_diff
    FROM
        fact_order o
    JOIN
        dim_user u ON o.user_id = u.user_id
    JOIN
        dim_date d2 ON o.delivered_date_id = d2.date_id
)
SELECT
    customer_state,
    AVG(delivery_time_diff) AS avg_delivery_time_diff_from_estimation
FROM
    delivery_diff
GROUP BY
    customer_state
ORDER BY
    avg_delivery_time_diff_from_estimation DESC
LIMIT 5;
'''
fetch_data(connection, query_delivery_time_diff)
```

Out[ ]:

	customer_state	avg_delivery_time_diff_from_estimation
0	MALUKU	14.1241
1	ACEH	13.8581
2	BENGKULU	13.8255
3	NUSA TENGGARA TIMUR	13.3982
4	PAPUA BARAT	13.3108