Assignment – 1

SQL MASTERY - The E-Commerce Analytics Challenge

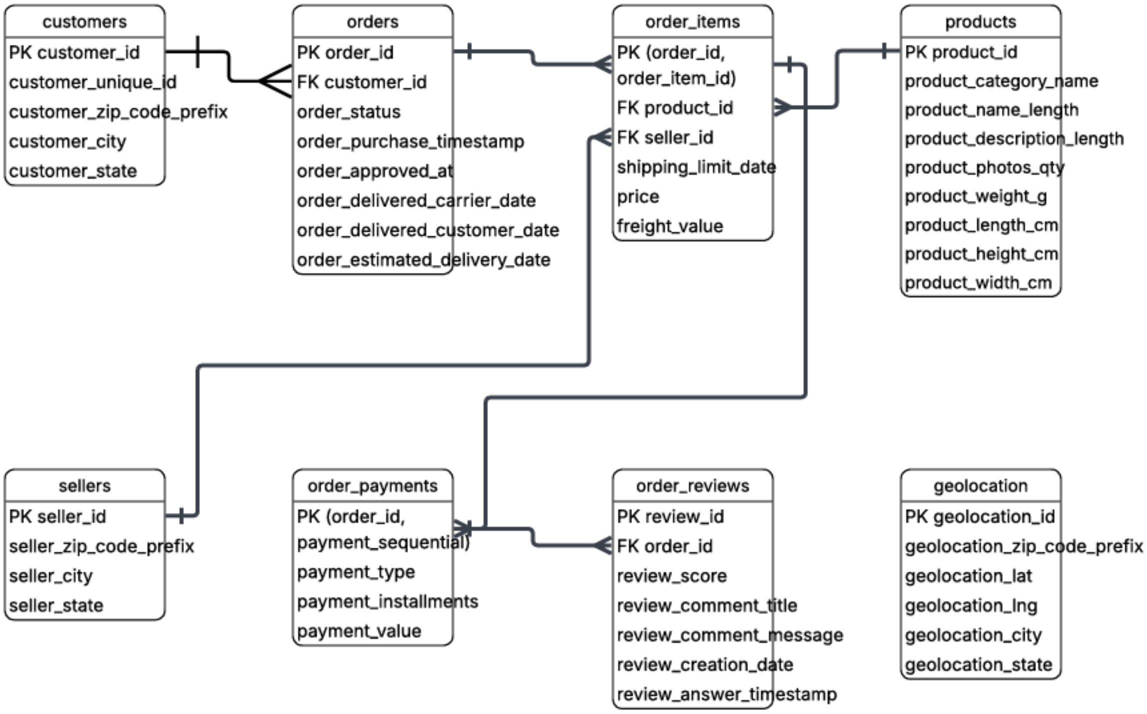
Part A: Database Design & Data Quality

Question -1. Design a normalized database schema (3NF) with ERD showing all relationships between 8 tables.

Solution –

Database schema is designed following third normal form(3nf) principles: -

- Each tables represents as single entity
- All non-key attributes depends only on the primary key
- No transitive dependencies exist
- Relationships are enforced using foreign keys



Question 2. Identify and document 10+ data quality issues in the raw CSV files (nulls, duplicates, format inconsistencies, orphan records).

Solution – 10+ data quality issues in the raw CSV files are:-

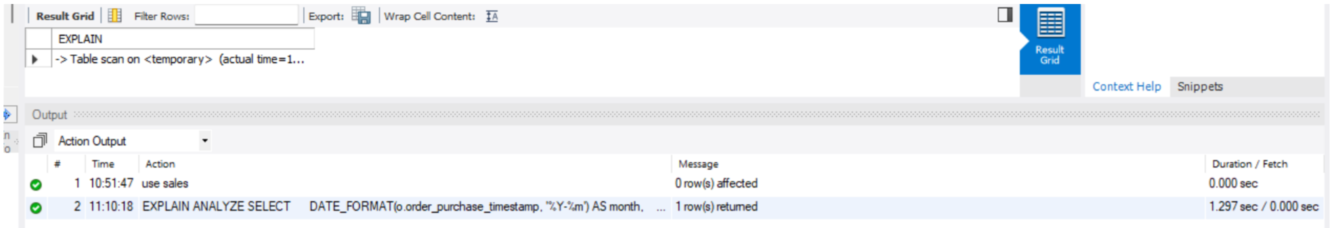
1. **Missing values** - several columns like order_delivered_customer_date, review_comment_message etc. contain missing values.
2. **Duplicate customer record** – multiple records exist for same customer but with different ids.
3. **Orphan records** – some columns refereeing to other columns where values are missing.
   Eg. order_items reference missing product_id
4. **Inconsistent datetime format** – Date columns are stored as string and may contain null or invalid values.
5. **Duplicate order items entries** – Duplicate combinations of multiple columns exist.
   Eg. geolocation_zip_code_prefix, geolocation_lat and geolocation_lng have duplicate combination values
6. **Invalid numerical values** – some record contain invalid numeric values like zero, negative for payment_value , price , etc
7. **Payment and order value mismatch** – Sum of order_items dono match payment_values.
8. **Missing reviews for delivered orders** – reviews are missing.
9. **Geolocation duplication** – The geolocation table contains multiple rows for the same zip code with different latitude and longitude values.
10. **Text formatting issues** – Customer and city names contain inconsistent letters and casing.

Question 30. Identify 3 slowest queries using EXPLAIN ANALYZE, optimize with appropriate indexes (B tree, Hash), show before/after execution time.

Solution –

1) Explain before –

```
EXPLAIN ANALYZE
SELECT
  DATE_FORMAT(o.order_purchase_timestamp, '%Y-%m') AS month,
  SUM(p.payment_value) AS revenue
FROM orders o
JOIN order_payments p
  ON o.order_id = p.order_id
GROUP BY month;
```



```
EXPLAIN
-> Table scan on <temporary>  (actual time=1...
```

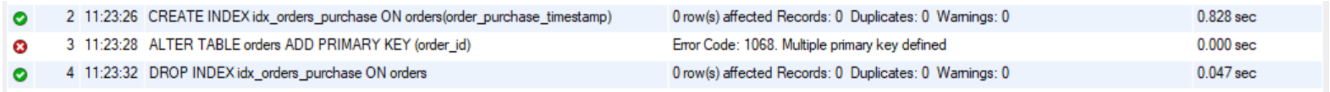
```
#   Time      Action                                                                      Message              Duration / Fetch
1   10:51:47  use sales                                                                   0 row(s) affected    0.000 sec
2   11:10:18  EXPLAIN ANALYZE SELECT    DATE_FORMAT(o.order_purchase_timestamp, '%Y-%m') AS month,  ...  1 row(s) returned    1.297 sec / 0.000 sec
```

2) Create Indexes –

```
-- Index for join
CREATE INDEX idx_payments_order
ON order_payments(order_id);

-- Index for grouping/filtering
CREATE INDEX idx_orders_purchase
ON orders(order_purchase_timestamp);

-- Always ensure PK exists
ALTER TABLE orders
ADD PRIMARY KEY (order_id);
```

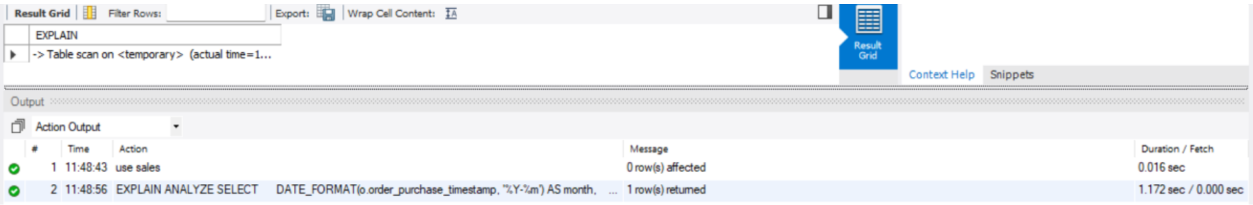


```
2  11:23:26  CREATE INDEX idx_orders_purchase ON orders(order_purchase_timestamp)   0 row(s) affected Records: 0  Duplicates: 0  Warnings: 0   0.828 sec
3  11:23:28  ALTER TABLE orders ADD PRIMARY KEY (order_id)                         Error Code: 1068. Multiple primary key defined             0.000 sec
4  11:23:32  DROP INDEX idx_orders_purchase ON orders                              0 row(s) affected Records: 0  Duplicates: 0  Warnings: 0   0.047 sec
```

3) Explain after

```
-- Query 1: AFTER Optimization

EXPLAIN ANALYZE
SELECT
  DATE_FORMAT(o.order_purchase_timestamp, '%Y-%m') AS month,
  SUM(p.payment_value) AS revenue
FROM orders o
JOIN order_payments p
  ON o.order_id = p.order_id
GROUP BY month;
```



EXPLAIN
-> Table scan on <temporary> (actual time=1...

	#	Time	Action	Message	Duration / Fetch
✓	1	11:48:43	use sales	0 row(s) affected	0.016 sec
✓	2	11:48:56	EXPLAIN ANALYZE SELECT DATE_FORMAT(o.order_purchase_timestamp, '%Y-%m') AS month, ...	1 row(s) returned	1.172 sec / 0.000 sec