STAGE 1

SQL Process Notes

Clean and Transform

1. Load & Restructure

1st SCHEMA

* 1. Created ‘raw’ schema
     1. Allows rollback to original data

2nd SCHEMA

* 1. Created ‘clean’ schema
     1. Enabled pipeline to stages of cleaning
     2. Kept structure of clean load simple (datatypes only)
        1. Matched raw file table/column structure
  2. Removed columns irrelevant to project context
     1. Phone, email, zip\_code enable to analyse
     2. Dropped before cleaning to reduce redundancy (ie clean then remove)

1. Profiling & Assessment
   1. Row count matches raw data
   2. Checked for key-based duplicates
      1. None found
   3. Ran full-row duplicate checks
      1. Nine found
   4. Ran NULL and distinct checks on all columns
      1. NULL values in shipping\_date, manager\_id
   5. Ran summary stats on key metrics (numerical columns)
      1. List\_price shows expected range
      2. Discount tiers adhere to expected norm
      3. Model\_year within bounds
      4. Stock level as expected
2. Handle Missing Data
   1. NULL values in shipping\_date
      1. Expected and left as NULL
      2. Poss use flag column (delivered/not delivered) at later stage
   2. NULL values in manager\_id
      1. Likely CEO, left as is
3. Remove/Resolve Duplicates
   1. n/a
4. Standardise
   1. Trimmed whitespace on all string columns
   2. Removed unnecessary syntax on:
      1. products\_name column
         1. Info held elsewhere – year, model etc
      2. category\_name column (as above - info elsewhere)
5. Datatype Conversion
   1. Reformatted string date columns to correct Date data type
      1. Order\_date, required\_date, shipped\_date
         1. Shipped\_date carried NULL’s
6. Normalisation/Transformation
   1. Created date table
      1. Added Boolean flag (yes/no) for comparison analysis
      2. Created sequence and added continuous dates
         1. Used static dates for ease of analysis
      3. Populated other descriptive columns
      4. Added FK’s to original date source table
         1. Used new columns for integrity (will match original cols exactly)
      5. Populated FK’s columns from new date table key column
   2. Created geography dimension table
      1. Created table stage
      2. Populated new table with values from source tables (customers, stores)
      3. Added ID columns to source tables (customers, stores) and dropped original source columns (state, city)
   3. Created order status dimension table
      1. Created staging
      2. Inserted likely string values to match numerical ranking
         1. Values unknown, derived from logically likeliness
      3. Altered source col (order\_status) in source table (orders) to ID col
7. Model

3rd SCHEMA

* 1. Created ‘finalBK’ schema to enable multi-stage save and load (raw, clean,final)
  2. Combined orders and order\_items in one fact table: sales
     1. Increased efficiency in ‘star/snowflake schema’ model
     2. Stocks stays as 2nd fact table
  3. Created all other dimension tables
     1. Changed table names:
        1. date\_dim (dates); order\_status\_dim(order status); staffs(staff)
  4. Added all constraints
     1. Nulls, primary keys, foreign keys
  5. Added indexes
  6. Inserted data from ‘clean’ schema into ‘finalBK’ schema

1. Transformations
   1. Added calculated column
      1. Sale\_price from list\_price and discount
   2. Added concatenated name column
      1. Full\_name from first\_name and last\_name
2. Validation
   1. Ran structural consistency table check
      1. No errors
   2. Checked for missing rows in ‘parent’ to ‘child’ table relationship
      1. No NULL’s returned
   3. Ran multiple joins across multiple tables
      1. All data looked correct and no NULL’s returned
   4. Checked for missing dates
      1. None found
   5. Ran aggregation query
      1. Results as expected