Case Study Exam: Banggood Product Data Pipeline & Analysis

Scenario

You are working for a startup that wants to analyze product trends on Banggood.com. Your task is to scrape, clean, analyze, store, and aggregate data for five selected categories.

Part 1: Data Extraction (Web Scraping)

1. Select any 5 categories from Banggood.

2. Scrape product data including product name, price, rating, reviews, and URL.

3. Use Python libraries such as requests, BeautifulSoup, Selenium.

4. Implement pagination to fetch multiple pages.

Part 2: Data Cleaning & Transformation

1. Load scraped data into pandas DataFrames.

2. Clean price, rating, review counts, and handle missing values.

3. Create at least two additional derived features.

Part 3: Python Exploratory Analysis (Minimum 5 Analyses)

Examples:

- Price distribution per category

- Rating vs Price correlation

- Top reviewed products

- Best value metric per category

- Stock availability analysis

Part 4: Load Data into SQL Server

1. Create database schema (one table per category or unified table).

2. Use pyodbc/pymysql to connect and insert data.

3. Validate inserts by querying row counts.

Part 5: SQL Aggregated Analysis (Minimum 5 Queries)

Examples:

- Average price per category

- Average rating per category

- Product count per category

- Top 5 reviewed items per category

- Stock availability percentage

Part 6: Final Report

Submit a report including:

- Architecture Diagram

- GitHub repo with proper README

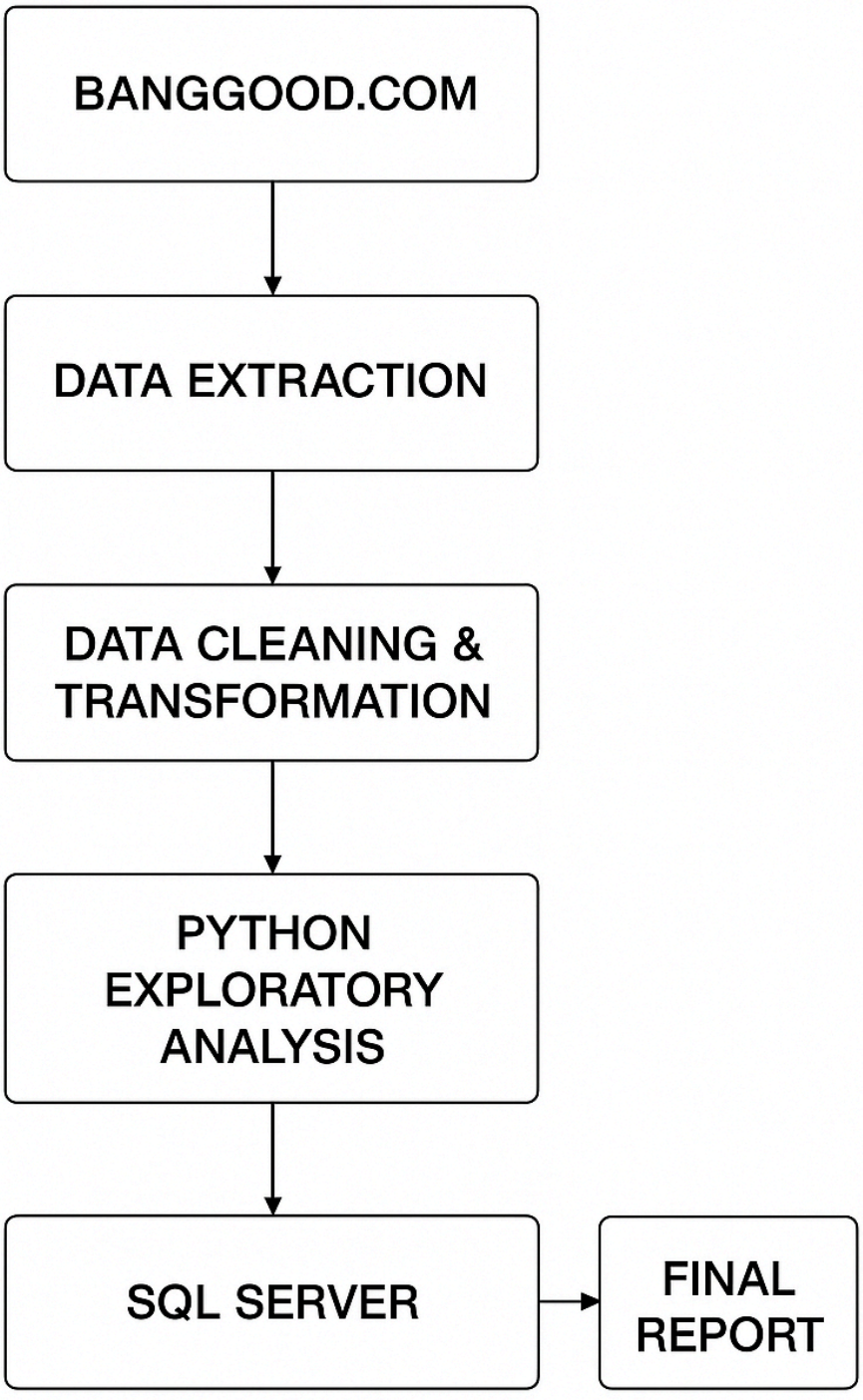
- Scraping methodology

- Cleaning/transformation steps

- Python analyses (with graphs)

- SQL aggregated insights

- Final conclusions and recommendations


BANGGOOD.COM
DATA EXTRACTION
DATA CLEANING & TRANSFORMATION
PYTHON EXPLORATORY ANALYSIS
SQL SERVER
FINAL REPORT