Stock Analytics Pipeline — Business Requirements

Context:

The investment analytics team needs a robust data processing pipeline to evaluate stock performance across sectors. The goal is to transform raw monthly stock data into actionable insights for portfolio decisions.

Input Data:

Stock	Sector	Month	PriceStart	PriceEnd
AAPL	Technology	2025-01	120.5	175.2
MSFT	Technology	2025-01	210.1	280.5
WMT	Consumer Staples	2025-01	140.0	150.0
TSLA	Automotive	2025-01	600.0	880.0
JPM	Financials	2025-01	95.0	105.0
BABA	Consumer Discretionary	2025-01	180.0	160.0
RELIANCE	Energy	2025-01	2200.0	2400.0
TCS	Technology	2025-01	3100.0	3500.0
HDFC	Financials	2025-01	1500.0	1650.0
BAJAJ-AUTO	Automotive	2025-01	4800.0	5200.0

Requirements

---

1. Track Price Movement

Calculate how much each stock's price changed during the month. Add a column PriceChange to show the difference between closing and opening prices. This helps identify volatility.

---

2. Determine Market Trend

Classify each stock as **Gain** or **Loss** based on whether its closing price is higher or lower than its opening price. This trend indicator will be used for quick performance checks.

---

3. Compute Return on Investment

For every stock, calculate the percentage return using the formula:

Return(%) = ((PriceEnd - PriceStart) / PriceStart) * 100

Round to two decimal places. This metric is critical for investment decisions.

4. Rank Stocks by Return

Sort all stocks in descending order of Return(%) so analysts can quickly identify top performers and prioritize investment opportunities.

---

5. Detect Outliers in Return

Use statistical logic (e.g., returns greater than mean + 2*std deviation) to flag outlier stocks. These could indicate unusual market behavior.

---

6. Ensure Data Quality

Replace any invalid price values (≤ 0) with NULL to maintain data integrity and prevent incorrect calculations.

Check if any stock violates business rules like:

- PriceStart should always be > 0
- Return should not exceed 200% (flag unrealistic data)

---

7. Export Processed Data

Save the transformed dataset in a structured format (CSV or Parquet) for downstream analytics and reporting. Log invalid records separately for audit and compliance.

Folder Structure & Data Management Standards

To ensure consistency and maintainability in the data processing pipeline, the following directory structure and naming conventions must be implemented:

Project Folder Structure

<processing_script>.py

/input/

/output/

/error/

Input Directory

- Create an **input/** folder to store all incoming data files.
- Only process files that have been added within the last 24 hours to ensure timely and relevant data ingestion.

Output Directory

- Create an **output/** folder to store all successfully transformed datasets.
- **Naming Convention:**
  Transformed_Data_<timestamp>.<csv/parquet>
  This ensures traceability and version control for processed data.

Error Directory

- Create an **error/** folder to capture and store all records that fail validation or processing due to bad data.
- **Naming Convention:**
  Bad_Data_<timestamp>.csv
  This allows for easy identification and review of problematic records.