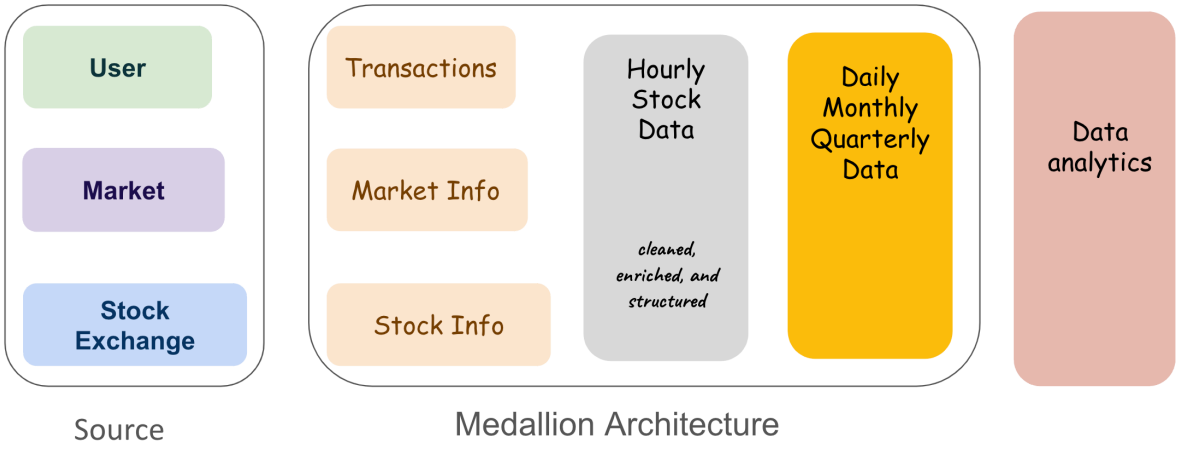
# Medallion architecture

## Objective

Design and implement **Medallion Architecture** on stock data, from source to data analytics.

## Stock data introduction




Stock-related data* is generated every second from multiple sources, including users, the market, and stock exchanges. The **Bronze Layer** captures raw data, consisting of user transactions, market information, and stock and company registration details from the stock exchange.

After aggregation and cleaning, the **Silver Layer** contains structured and detailed data at a granular level, making it suitable for further processing.

In the **Gold Layer**, data is aggregated into daily, monthly, and yearly summaries, optimizing it for reporting, dashboards, and business intelligence applications.

**The data used in this project is artificially generated and does not represent real financial data. However, it is designed to closely mimic real-world scenarios, ensuring that stock prices fluctuate within a reasonable range over time.*

1. **Load `data/stock_market.sql` into MySQL database**

2. **Load `data/transactions.json` into Mongodb (5 points)**

3. **Extract data from mysql and mongodb database(10 points)**
   - Load data into spark sessions (5 points)
4. **Aggregate data together for further data analysis (15 points)**
   **Note: only pyspark functions are allowed for data aggregation, other python packages such as pandas may not be used**
   a. The final table should contain all information shown below, stock price fluctuates and the table requires average stock prices used for transactions, the volume of the transactions (sum of buy and sell), and the market index. Column sequence and name should follow the example given below.
   b. For further data analysis, aggregate the data into different granular level

      i. **Hourly data (4)**

```
Hourly data preview (5 rows):
+-------------------+------+--------------------+---------+------+------------+
|datetime           |ticker|company_name        |avg_price|volume|market_index|
+-------------------+------+--------------------+---------+------+------------+
|2024-10-01 09:00:00|AAPL  |Apple Inc.          |175.74   |560   |2290.82     |
|2024-10-01 09:00:00|AMZN  |Amazon.com Inc.     |145.35   |350   |2290.82     |
|2024-10-01 09:00:00|JPM   |JPMorgan Chase & Co.|147.94   |380   |2290.82     |
|2024-10-01 09:00:00|META  |Meta Platforms Inc. |299.91   |910   |2290.82     |
|2024-10-01 09:00:00|NVDA  |NVIDIA Corporation  |485.81   |180   |2290.82     |
+-------------------+------+--------------------+---------+------+------------+
only showing top 5 rows
```

ii. **Daily data (4)**

```
Daily data preview (5 rows):
+----------+------+--------------------+---------+------+------------+
|date      |ticker|company_name        |avg_price|volume|market_index|
+----------+------+--------------------+---------+------+------------+
|2024-10-01|AAPL  |Apple Inc.          |175.42   |9940  |2291.07     |
|2024-10-01|AMZN  |Amazon.com Inc.     |145.56   |7480  |2291.36     |
|2024-10-01|GOOGL |Alphabet Inc.       |140.69   |9300  |2291.11     |
|2024-10-01|JPM   |JPMorgan Chase & Co.|148.14   |8860  |2291.36     |
|2024-10-01|META  |Meta Platforms Inc. |300.18   |11460 |2291.36     |
+----------+------+--------------------+---------+------+------------+
only showing top 5 rows
```

iii. **Monthly data (4)**

```
Monthly data preview (5 rows):
+-------+------+--------------------+---------+------+------------+
|month  |ticker|company_name        |avg_price|volume|market_index|
+-------+------+--------------------+---------+------+------------+
|2024-10|AAPL  |Apple Inc.          |177.59   |197580|2277.62     |
|2024-10|AMZN  |Amazon.com Inc.     |146.31   |199880|2277.59     |
|2024-10|GOOGL |Alphabet Inc.       |141.69   |188410|2277.82     |
|2024-10|JPM   |JPMorgan Chase & Co.|149.7    |194580|2277.55     |
|2024-10|META  |Meta Platforms Inc. |300.38   |203920|2277.80     |
+-------+------+--------------------+---------+------+------------+
only showing top 5 rows
```

iv. Quarterly Data (Oct - Dec Summary)

```
Quarterly data preview (5 rows):
+-------+------+--------------------+---------+------+------------+
|quarter|ticker|company_name        |avg_price|volume|market_index|
+-------+------+--------------------+---------+------+------------+
|2024 Q4|AAPL  |Apple Inc.          |173.82   |571770|2252.27     |
|2024 Q4|AMZN  |Amazon.com Inc.     |146.55   |579280|2252.79     |
|2024 Q4|GOOGL |Alphabet Inc.       |139.76   |567280|2253.10     |
|2024 Q4|JPM   |JPMorgan Chase & Co.|150.84   |531950|2253.37     |
|2024 Q4|META  |Meta Platforms Inc. |300.68   |583100|2252.80     |
+-------+------+--------------------+---------+------+------------+
only showing top 5 rows
```