**Setup and instructions on how to run the code:**

1. **Setup**

1. Install Python 3.12.3

2. Install dependencies:

pip install -r requirements.txt

3. Install pyspark==3.5.1

2. **Running the Script**

To run the script, use the following command:

python batch\_aggregation.py <input\_csv\_file> <output\_directory> <time\_bucket\_duration>

3. **Command Line Arguments**

- `<input\_csv\_file>`: Path to the input CSV file.

- `<output\_directory>`: Directory where the output files will be written.

- `<time\_bucket\_duration>`: Time bucket duration for aggregation (e.g., "1 day", "1 hour", "30 minutes"). Please use quotes for this argument as it has spaces in it.

4. **Input File Format**

The input CSV file should have a header with the following columns:

- `metric`: The metric identifier

- `timestamp`: The timestamp of the measurement

- `value`: The measurement value

5. **Output**

The script will output aggregated data (average, minimum and maximum values) into the specified output directory, grouped by the provided time bucket duration.