Batch Aggregation Coding Challenge

The analytics teams handle a lot of time series data that has to be aggregated. For this challenge, we want you to write a batch job that can perform such an aggregation. The expected input dataset will contain a metric column with the name of the metric, a value column with the value of the metric, and a timestamp column. The batch job is supposed to aggregate the data by metric and time bucket. For each time bucket and each metric in the bucket, we want you to calculate the average value of the metric.

Example input data

Metric	Value	Timestamp
temperature	88	2022-06-04T12:01:00.000Z
temperature	89	2022-06-04T12:01:30.000Z
precipitation	0.5	2022-06-04T14:23:32.000Z
...	...	...

Constraints

- You may use Java and Spark (or MapReduce) or Python and PySpark for this project.
- The input dataset and the generated output can be in any format that you find useful for this challenge.
- You may share the code via public repository such as GitHub. If you do this, please use a random name and don't reference Genesys. You can also share your code via email.
- You should spend no more than 3 to 4 hours on this challenge. Consider this a proof of concept. The implementation does not have to be perfect.

Notes

- You may assume a fixed time bucket duration, such as 24 hours. If you can support flexible time buckets that's great, too.
- You may assume that all the columns are always present; there is no missing data.
- Timestamps can be in any format you like (epoch timestamps, date strings, ...).
- You should add comments to your code where you find them useful but we don't expect full documentation.
- Bonus: Can you extend this to also report minimum and maximum value.
- If you have questions or clarifications, please contact your recruiter.
- We will discuss the code you submitted during the technical interview.
- Have fun!