

△ > StarTree ThirdEye > Getting Started > ThirdEye data requirements

# ThirdEye data requirements

StarTree ThirdEye searches through your data looking for data points that are unexpected. For anomaly detection to be effective, your data must be configured in a way that makes sense to the algorithm examining it. This page provides the data schema requirements and configuration to help you use ThirdEye effectively.

**i** Time series data may have different scales, ranges, and units. You must normalize your data for it to be balanced in how it affects anomaly detection algorithms.

There are three types of columns that are required for ThirdEye to work; a timestamp, a metric, and a dimension. Further considerations are listed after a presentation of details for each of these followed by examples.

## Timestamp

This field is vital for detection accuracy. ThirdEye works with seasonal or trend-based patterns, but does not perform well with random or cyclic data. It also requires at least 30 days worth of data for granularity in the hourly/daily range.

ThirdEye works with timestamps that have:

- One column, with type of 'DateTime'
- When this column isn't set, the timestamp is automatically set as the start time of each ingestion period
- A format that Pinot can use, see [DateTime strings in Pinot](#) for more.

## Metric

A metric is a quantifiable value. With ThirdEye, metrics can be:

- One or more columns containing numeric or categorical values in the data feed. For each data feed, you can specify multiple metrics, but you must select at least one column as `metrics`.
- Directly configured during alert configuration and need not be a column in the dataset, such as events streams on which you can use count metric.

**i** If you want to use `SUM` to aggregate your data, make sure your metrics are additive in each dimension. Fraction-based metrics are non-additive. This includes ratio, percentage, and so on.

You can use pre-aggregated metrics or configure aggregation methods in ThirdEye and run it during anomaly detection at run time. If the Pinot data has correct indexes and data volume is not too high then you will get sub-seconds results for aggregated queries.

## Dimension

A dimension is one or more categorical values. For ThirdEye:

- A combination of different values identifies a particular single-dimension time series, for example: country, language, tenant.
- You can select zero or more columns as dimensions.
- The dimension columns can be of any data type.

**i** Avoid using large volumes of columns and values to prevent excessive numbers of dimensions from being processed, which is very slow.

## Further considerations

If you are doing real-time anomaly detection, verify the logical steps done correctly. These are:

- Get event

- Write event
- Query historic events (requires at least 30 days' worth of data)
- Run anomaly detector

You want to ensure missing data is handled well or use filters in anomaly detection configuration do avoid that becoming a problem.

Find the right granularity of your data.

- When data granularity is very high, you have more data, but it will become noisy
- As your data granularity goes down, it becomes tough to detect anomalies
- Pay attention to consistency, use the same granularity across data so you don't have one source being recorded with a granularity of 1 minute and another with a granularity of 5 minutes.

## Examples

---

### Revenue metrics

Timestamp	Category	Market	Revenue
2020-6-1	Food	US	1000
2020-6-1	Apparel	US	2000
2020-6-2	Food	UK	800

In this table, the:

- Metric is Revenue
- Dimensions are Category and Market
- Timeseries pattern is daily

### Application errors

Timestamp	Application component	Region	Error count
2020-6-1	Employee database	WEST EU	9000
2020-6-1	Message queue	EAST US	1000
2020-6-2	Message queue	EAST US	8000

In this table, the:

- Metric is Error count
- Dimensions are Application component and Region
- Timeseries pattern is Daily

Last updated on May 21, 2024