

Lectura 2: Prometheus & Grafana

Diego Granados Retana 2022158363

Bases de datos II

22 de agosto, 2023

Responda las siguientes preguntas:

¿En qué consisten los datos timeseries?

Son datos que se registran a lo largo del tiempo (Grafana Labs, s.f.). Usualmente pertenecen a la misma métrica y al mismo conjunto de dimensiones. Se pueden desplegar en tablas, pero lo más útil es usar un gráfico. Usualmente, las mediciones no se actualizan, solo se añaden nuevas. Permiten entender el pasado al mostrar información de cualquier punto en el tiempo. De acuerdo con IBM (2021), hay diferentes tipos de series:

- Dependiente: Para esta se desea hacer previsiones.
- Predictora: Ayuda a explicar un objetivo.
- Evento: Es una serie predictora que se utiliza para tomar en cuenta eventos recurrentes y predecibles.
- Intervención: Es una serie predictora que se usa para tomar en cuenta incidentes específicos que han ocurrido en el anteriormente.

¿Qué son métricas?

Son mediciones numéricas (Grafana Labs, s.f.). Algunos ejemplos pueden ser la cantidad de usuarios que inician sesión, la cantidad de lecturas que se realizan, la cantidad de memoria que se está usando, entre otros. Dicen cuánto existe de algo.

Explique en que consiste la Observabilidad.

La observabilidad es poder ver el estado interno del sistema con base en los datos que se producen, hacerlo más transparente (Grafana Labs, s.f.). Existen herramientas que nos ayudan a recolectar métricas y mantener una bitácora de todo lo que ocurre en el sistema. Con estos, podemos medir el rendimiento y disponibilidad de nuestro sistema, encontrar errores y ver si las soluciones que implementamos sí estén funcionando. Hay tres pilares fundamentales que son registradas a lo largo del tiempo:

- Logs: registros de eventos que ocurrieron.
- Métricas: representaciones numéricas de datos.
- Rastros o traces: Una cadena de eventos que ocurren en el sistema. Existe una diferencia entre la observabilidad y el monitoreo de rendimiento. El monitoreo se enfoca más en recolectar y analizar datos del rendimiento del sistema y su conducta y así revisar si el sistema está funcionando correctamente. La observabilidad se enfoca más en entender cómo los componentes de sistemas complejos interactúan entre sí.

Explique el concepto de dimensiones en datos timeseries.

Las dimensiones son las características o el contexto en el cual se están midiendo los datos (Grafana Labs, s.f.). En la lectura se menciona un ejemplo con las dimensiones job y device. Esto significa que cada time series fue obtenido en una tarea y dispositivo específicos. Es posible que, si se cambiara el dispositivo, las métricas darían resultados diferentes. Las dimensiones son importantes para estudiar cómo se comporta nuestro sistema con diferentes variables.

¿Por qué los tags en métricas permiten generar mejores gráficos en Grafana?

Los tags o labels se utilizan para definir dimensiones en Grafana (Grafana Labs, s.f.). Esto permite agregar el contexto de dónde se obtuvo la información. Además, permiten tener datos de diferentes dimensiones en el mismo gráfico, generando una forma de comparar directamente cómo se diferencian las dimensiones. Los labels se almacenan en pares de llaves y valores.

Suponiendo que se están recolectando datos IoT (Internet of Things) de miles de dispositivos, los mismos generan una métrica cada 15 segundos con el consumo de energía y temperatura, explique:

- **¿Por qué una base de datos relacional no es una buena opción para almacenar esta información?**

Las bases de datos relacionales no están optimizadas para almacenar time series. Es posible que los datos de time series puedan ser almacenados estructuralmente, ya que usualmente tienen una característica identificadora, como el tiempo en que fueron tomadas y al almacenar normalmente datos numéricos, sí es posible construir un esquema. No obstante, en la operación, hay varias desventajas (Grafana Labs, s.f.). Los datos de time series se actualizan o eliminan muy poco, entonces esto permite optimizar el almacenamiento. Puede ser que se almacene un tiempo exacto y a partir de ahí, solo se almacene la diferencia que hay en los siguientes registros a ese tiempo. Como los datos se almacenan en intervalos regulares y cortos, la diferencia va a ser menor, entonces se pueden almacenar las diferencias en menos espacio ya que son números pequeños. Si se almacenara todo el número, habría un desperdicio de espacio porque habría información redundante. Otra razón es que una base de datos relacional no puede soportar la velocidad en la que se insertan datos (Watchel, 2023). Además, algunos datos tienen dimensiones y campos diferentes, entonces no se podrían añadir a la misma tabla rápidamente ya que habría que cambiar el esquema. Por estas razones, existen bases de datos especializadas para time series. Estas aplican estas optimizaciones para mejorar el rendimiento y uso de recursos.

- **Dada la naturaleza de datos timeseries, ¿de qué forma la localidad puede ayudarnos a ahorrar dinero?**

La localidad es muy útil porque implica que todos los datos de las mismas dimensiones se almacenan en ubicaciones cercanas. Así, puedo agarrar toda la información que necesito para un gráfico de una vez en lugar de buscar por todo el disco y recolectarla. Eso me ayuda a generar el gráfico más rápido y le evita a la base de datos tener que subir tantas páginas o bloques de información a la memoria, ya que todo estaría en el mismo.

Bibliografía

IBM. (2021, 17 agosto). Datos de series temporales. IBM.com. Recuperado 5 de agosto de 2023, de <https://www.ibm.com/docs/es/spss-modeler/saas?topic=models-time-series-data>

What is Prometheus? | Grafana Documentation. (s. f.). Grafana Labs. <https://grafana.com/docs/grafana/latest/fundamentals/intro-to-prometheus/>

Wachtel, J. (2023, 24 abril). Don't let time series data break your relational database. The New Stack.

<https://thenewstack.io/dont-let-time-series-data-break-your-relational-database/#:~:text=Relational%20databases%20can%27t%20handle,reach%20a%20scaling%20inflection%20point.>