Pedro Lima

Master in Informatics and Computing Engineering
Management of large volumes of data collected by wind monitoring sensors
Pedro Filipe Agrela Faria

—

Abstract

Nowadays, virtually everything around us generates large amounts of data. With this comes the need to store such large volumes of data in a scalable and efficient way. Conventional databases, which follow relational models, are no longer scalable when applied to large amounts of data since it follows the properties: Atomicity, Consistency, Isolation, Durability — ACID. In these situations, losing the ACID properties, NoSQL databases are preferred due to easy scalability. There are also time series databases that are designed to store pairs of timestamp and value attributes.

In the case of the renewable energy, in particular, the wind energy area, the decision making is often — if not always — sustained by large data sets collected by different instruments and sensors. The constant fluctuation of the meteorological variables, such as wind speed, direction, air temperature, etc. increases the importance of accurate measurements of such variables. This can be done during prospect phase, where local wind conditions are assessed to ensure wind turbines suitability, or post construction, to track the performance of the wind farms, monitor the conditions of critical components and optimize the wind turbines control. These data can be recorded at different frequencies, typically with sampling rates of 1 Hz and integration times of 10 minutes where several statistics are generated. In certain circumstances, the sampling frequencies can reach higher values and the integration time may not even be applied, meaning all data needs to be stored for later processing, which creates a large amount of data.

The main objective of this work is to create an information system capable of storing and managing large volumes of data, processing, cleaning invalid records, and subsequent reporting of that data. This challenge was proposed by INEGI, who has an implemented system currently, although with some limitations, namely in what concerns scalability and flexibility. The created platform aims to solve those problems and is designed to support three database systems: PostgreSQL, MongoDB and Influx. We studied the performance by inserting and querying the raw data, and also the disk space occupied by each database system.

The performed tests not only helped us to choose the best database to do such operations but also allowed us to improve the platform during these tests. The study revealed that a Relational Database Management System — RDBMS — and a Time Series Database — TSBD — can coexist in the same system, using PostgreSQL to store and handle all the meta information of the system and Influx to store the raw data provided by the wind towers.

To develop the platform, we studied the different Actors in the system, refering to the different access levels in the platform by each user. We analyzed the different User Stories that contain the individual interactions in the platform by those users, and we also designed the Model of the Domain to describe the entities and their subsequent relationships present in the system. The platform was implemented in Django, a Python framework, that follows an Modal-View-Controller — MVC — architecture to separate the different layers in the Web system.

Resumo

No decorrer dos dias de hoje, e cada vez mais, estamos rodeados de dados, o que gera uma necessidade de armazená-los de uma forma eficiente e escalável. As base de dados convencionais, que seguem o esquema relacional, deixam de ser escaláveis quando se deparam com grande volumes de dados, pois seguem as propriedades: Atomicidade, Consistência, Isolamento e Durabilidade — ACID. Nestas situações, perdendo as

propriedades ACID, são preferidas as base de dados NoSQL devido à sua fácil escalabilidade. Existem também base de dados especificamente desenhadas para guardar pares de *timestamp* e valor, que são designadas por base de dados *time series*.

No caso da energia eólica, a tomada de decisão é muitas vezes — se não sempre — sustentada por grandes conjuntos de dados recebidos por diferentes instrumentos e sensores. A constante flutuação das variáveis meteorológicas, como a velocidade e direção do vento, temperatura do ar, etc., aumenta a importância de ler essas variáveis de uma forma mais precisa. Isso pode acontecer durante a fase prospetiva, onde as condições locais de vento são avaliadas para garantir o correto funcionamento das turbinas eólicas, ou pós-construção, para monitorizar o desempenho dos parques eólicos, as condições dos componentes críticos e otimizar o controle das turbinas eólicas. Esses dados podem ser registados em frequências diferentes, geralmente com taxas de amostragem de 1 Hz e tempos de integração de 10 minutos, onde são geradas várias estatísticas. Em determinadas circunstâncias, as frequências de amostragem podem atingir valores mais altos e o tempo de integração pode nem ser aplicado, o que significa que todos os dados precisam de ser armazenados para processamento posterior, o que cria uma grande quantidade de dados.

O principal objetivo desta dissertação passa por criar um sistema de informação capaz de armazenar e manipular grandes volumes de dados, processá-los, limpar registos inválidos e posteriormente gerar relatórios desses dados. Este desafio foi proposto pelo INEGI, que atualmente tem um sistema implementado, contudo possui limitações de escalabilidade e de flexibilidade. Esta plataforma visa resolver esses problemas e é projetada para suportar 3 sistemas de base de dados: PostgreSQL, MongoDB e Influx. Para isso, estudou-se o desempenho, inserindo e consultando os dados brutos, e também o espaço em disco ocupado por cada sistema de base de dados.

Os testes realizados, ajudaram a escolher qual a melhor base de dados para realizar essas operações e também permitiram melhorar a plataforma durante os mesmos. O estudo revelou que as base de dados relacionais e as *time series* podem coexistir no mesmo sistema, usando o PostgreSQL para armazenar e manipular a meta-informação do sistema e o Influx para armazenar os dados brutos fornecidos pelas turbinas eólicas.

Para desenvolver a plataforma, estudou-se os diferentes Atores do sistema, no qual permitenos diferenciar os níveis de acesso na plataforma por cada tipo de utilizador. Analisou-se as diferentes *User Stories* que contêm as interações individuais na plataforma por cada utilizador e desenhou-se também o Modelo Conceptual com o objettivo de descrever as entidades e as suas subconsequentes relações no sistema. A plataforma criada foi implementada utilizando o Django, uma framework Python, que segue uma arquitetura *Model-View-Controller* — MVC —, no qual separa as diferentes camadas de um sistema Web.

Jury

- Chair: Prof. Sérgio Sobral Nunes
- External Examiner: Prof. Maria Benedita Campos Neves Malheiro
- Supervisor: Prof. João Correia Lopes
- Date: 11/07/2019