XXIII SIMPÓSIO BRASILEIRO DE RECURSOS HIDRÍCOS

MODELO DIMENSIONAL DE BANCO DE DADOS HIDROLÓGICOS EM POSTGRESQL: ESTUDO DE CASO SSDPCJ

João Rafael Bergamaschi Tercini [1] ; Victor Alberto Romero Gonzalez [2] ; Carla Voltarelli Franco da Silva [3] ; Joaquin Ignacio Bonnecarrère Garcia [4] ; Arisvaldo Vieira Méllo Júnior [5] ; Cristiano de Pádua Milagres Oliveira [6] ; Felipe de Aguiar Pupo Luongo [7] ; Mayara Sakamoto Lopes [8] ; Diogo Bernardo Pedrozo [9] ; Aline Doria de Santi [10] ; Eduardo Cuoco Léo [11]

RESUMO – O modelo dimensional de banco de dados aplicado em recursos hídricos tem grande potencial de uso devido a sua generalização na entrada e saída de dados além de facilitar a agregação dos dados para geração de informações. O presente trabalho tem objetivo de demonstrar a aplicação deste modelo no software PostgreSQL, que é um motor de banco de dados potente e gratuito, ou seja, não gera custos com licenças para sua implementação e mesmo implementado em ambientes com servidores de pequeno porte, o modelo é capaz de realizar consultas complexas em base de dados enormes (bilhões de dados) com tempo de resposta muito rápidos, atingindo velocidades superiores a 300.000 dados processados por segundo. O modelo dimensional foi aplicado ao módulo de monitoramento do Sistema de Suporte a Decisões das Bacias PCJ (SSDPCJ), que tem a função de divulgar dados de estações de monitoramento telemétrico da bacia e demonstrou ser genérico, ter alta performance e baixo custo.

ABSTRACT– The dimensional model of database applied in water resources has great potential of use due to its generalization in the input and output of data besides facilitating the aggregation of the data for information generation. This paper aims to demonstrate the application of this model in PostgreSQL software, which is a powerful and free database engine, that is, it does not generate license costs for its implementation and even implemented in environments with small servers, the model is able to perform complex queries on huge database (billions of data) with very fast response times, reaching speeds in excess of 300,000 processed data per second. The dimensional model was applied to the monitoring module of the PCJ Basin Decisions Support System (SSDPCJ), which has the function of disseminating data from basin monitoring stations and was generic, high performance and low cost.

Palavras-Chave – dados hidrológicos, banco de dados, modelo dimensional

1) Fundação Centro Tecnológico de Hidráulica – FCTH, joao.tercini@fcth.br
2) Fundação Centro Tecnológico de Hidráulica – FCTH, victor.romero@fcth.br
3) Fundação Centro Tecnológico de Hidráulica – FCTH, carla.voltarelli@fcth.br
4) Escola Politécnica da Universidade de São Paulo– EPUSP, joaquinbonne@usp.br
5) Escola Politécnica da Universidade de São Paulo– EPUSP, arisvaldo@usp.br
6) Escola Politécnica da Universidade de São Paulo– EPUSP, cpmoliveira@usp.br
7) Escola Politécnica da Universidade de São Paulo– EPUSP, felipe.luongo@usp.br
8) Fundação Centro Tecnológico de Hidráulica – FCTH, mayara.lopes@fcth.br
9) Fundação Centro Tecnológico de Hidráulica – FCTH, diogo.pedrozo@fcth.br
10) Fundação Centro Tecnológico de Hidráulica – FCTH, aline.santi@fcth.br
11) Agência das Bacias Hidrográficas dos Rios Piracicaba, Capivari e Jundiaí - Agência das Bacias PCJ, eduardo.leo@agencia.baciaspcj.org.br

INTRODUÇÃO

Dados hidrológicos são essenciais para a tomada de decisão em recursos hídricos, tem importância desde o planejamento até o gerenciamento dos projetos, seja para abastecimento público, geração de energia, lançamento de efluentes, enfim os múltiplos usos da água. Dados hidrológicos ajudam a entender a variabilidade espacial e temporal da água e dão suporte aos tomadores de decisão sobre questões de segurança hídrica, permitindo mitigar danos que possam ser provocados por eventos extremos (estiagens e cheias).

Segundo ANA (2019) a coleta e registro de dados pluviométricos iniciou-se em 1846 na estação [338018] Aeroporto (Fortaleza). Para dados fluviométricos o início foi em 1900 na estação [58975000] Campos. Por mais de um século os dados hidrológicos foram armazenados em papel e para consulta-los era preciso acessar a biblioteca dos órgãos competentes.

Com o advento da computação no início da década de 70, o órgão responsável por gerenciar as informações coletadas em toda a rede hidrometeorológica nacional promoveu o desenvolvimento de um sistema de banco de dados, denominado Sistema de Informações Hidrológicas (SIH). O sistema foi evoluindo e hoje chame-se Hidro, que constitui uma aplicação de banco de dados do tipo cliente/servidor onde a base de dados hidrometeorológica é armazenada centralizadamente em um banco de dados relacional (ANA, 2010).

Terakawa (2013) avaliou 6 sistemas de gerenciamento de dados hidrológicos utilizando em diferentes países, todos os sistemas utilizavam softwares de banco de dados prontos para uso, o que o autor considerou benéfico, pois diminui o tempo e o custo do desenvolvimento. Oracle, Sybase e SQL Server são os motores de banco de dados utilizados pelos sistemas analisados, vale ressaltar que estes softwares são pagos e necessitam de licença para seu uso.

O uso de banco de dados no setor de recursos hídricos tornou-se indispensável devido a grande quantidade de dados gerados por estações de monitoramento, radares, sensoriamento remoto ou outros processos. Contudo, um sistema de banco de dados deve fornecer muito mais que armazenamento dos dados, principalmente em ambientes como comitês de bacias que necessitam de informações baseada nestes dados para tomadas de decisões, são necessárias também possibilidades de consultas inteligentes gerando informações mais palpáveis aos tomadores de decisão.

O inciso IV do artigo 44 da Lei das Águas (Lei Nº 9.433, de 8 de janeiro de 1997) diz que compete às Agências de Água, no âmbito de sua área de atuação, gerir o Sistema de Informações sobre Recursos Hídricos. A Agência das Bacias PCJ é responsável pela manutenção de 40 estações de monitoramento e divulga dados de mais de 750 estações, sendo 182 postos manuais, 164 postos telemétricos de chuva e vazão e 406 postos de qualidade da água. A padronização das diferentes fontes de dados que motivou o presente estudo.

Compreender a natureza dos dados hidrológicos, de forma que facilite os processos de geração de informação para tomada de decisão foi o objetivo deste trabalho que propõe e testa a modelagem dimensional de banco de dados. Dentre as premissas adotadas foi o uso de software livre e de código aberto, PostgreSQL. O modelo de banco de dados foi aplicado ao módulo de monitoramento do Sistema de Suporte a Decisões das Bacias PCJ (SSDPCJ), e vem sendo utilizado em caráter de testes pelas partes interessadas.

MODELO DIMENSIONAL APLICADO A DADOS HIDROLÓGICOS

Segundo Kimball e Ross (2013), a modelagem dimensional de banco de dados é amplamente aceita como a técnica para apresentar dados analíticos, pois organiza os dados de maneira compreensiva para analistas e fornece desempenho nas consultas, ou seja, são simples de entender e rápidas. Analisando os dados de monitoramento das bacias PCJ foi detectada a característica dimensional que os dados possuem, ou seja, para uma mesma estação podemos relacionar vários parâmetros, sendo que os parâmetros são medidos de tempos em tempos. A Figura 1 apresenta os três atributos identificados a partir da análise dos dados: Estação, Parâmetro e Data.

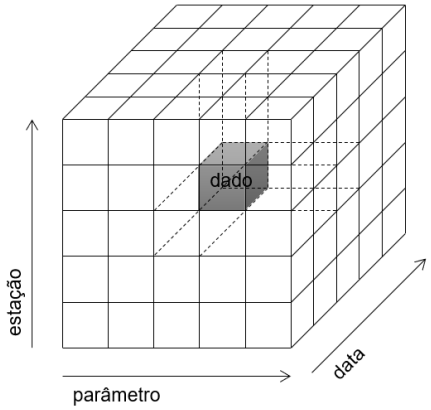

Figura 1 - Estrutura dimensional dos dados hidrológicos

A estrutura de dados em forma de cubo (3 dimensões) apresentada na Figura 1 significa que cada dado hidrológico armazenado sempre estará alinhado a uma estação, a um parâmetro e a data que ocorreu a medição. Tal estrutura proporciona a flexibilidade e generalização no armazenamento dos dados, por exemplo: se uma nova estação for implantada, ou um novo sensor for instalado em uma estação já existente conferindo novo parâmetro monitorado, não seria necessário alterar a estrutura do banco de dados.

Além de generalizar as condições de armazenamento proporciona generalização nas consultas das informações de acordo com as necessidades do tomador de decisão, a Figura 2 apresenta os tipos de consultas que geram informações agregando uma das características dos dados hidrológicos,

podendo ser: (i) no tempo variando as estações e parâmetros, (ii) no espaço variando os parâmetros e datas ou (iii) em indicadores variando as estações e datas.

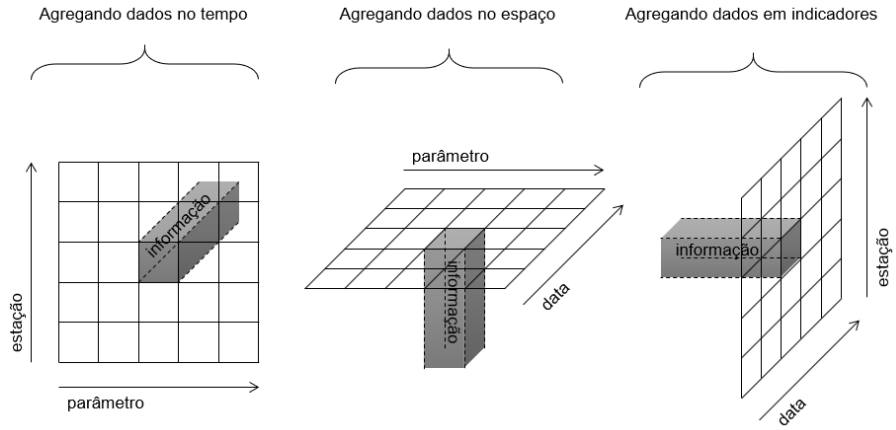

Figura 2 - Tipos de informações agregando uma das características dos dados hidrológicos

Os tipos de informações agregando uma das características dos dados hidrológicos apresentados na Figura 2 consegue cobrir muitos dos relatórios requisitado pelos tomadores de decisão em recursos hídricos, por exemplo: a agregação no tempo responderia qual a chuva média em uma determinada estação, a agregação no espaço responderia qual a série histórica do volume armazenado num sistema de reservatórios, a agregação em indicadores responderia qual a série histórica do IQA (Índice de Qualidade da Água) de uma determinada estação.

Em hidrologia resumir dados em intervalos de tempo padronizados com a finalidade de compor séries históricas diárias, mensais ou anuais dos parâmetros hidrológicos é muito comum e a estrutura dimensional permite realizar tal função com facilidade. A Figura 2 apresenta como se organiza o tipo de consulta de série histórica padronizada na estrutura dimensional de dados.

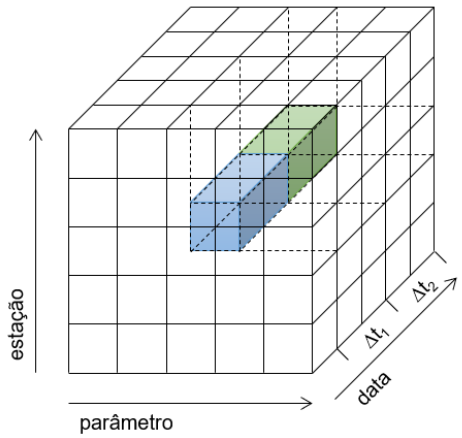

Figura 3 - Série histórica padronizada de dados hidrológicos

A série histórica padronizada de dados hidrológicos apresentada na Figura 2 consegue produzir informações comparáveis no tempo e são dados de entrada para modelos hidrológicos e de otimização, por exemplo: séries de precipitações acumuladas mensais, séries de vazões médias diárias

ou série da mediana do oxigênio dissolvido anual. Esse tipo de consulta pode ser realizado sobre outra consulta e assim poderíamos obter as vazões médias diárias e sobre esses resultados consultar as vazões máximas diárias anuais.

Outros tipos de informação são possíveis de ser geradas com este modelo dimensional, agregando-se duas das características dos dados hidrológicos é possível cobrir outros questionamentos de tomadores de decisão, por exemplo: indicadores no tempo responderia qual o IQA médio em uma determinada estação no tempo e espaço responderia qual a chuva média em uma bacia. Outro tipo de informação seria agregando todas características dos dados.

MÉTODOS

A programação da estrutura de dados dimensional foi executada no PostgreSQL que é um sistema de gerenciamento de banco de dados relacional de objeto, desenvolvido inicialmente no Departamento de Ciência da Computação da Universidade da Califórnia em Berkeley. Suporta uma grande parte do padrão SQL e oferece muitos recursos como: consultas complexas, chaves estrangeiras, gatilhos, visualizações atualizáveis, integridade transacional, controle de concorrência multiversão (PostgresSQL, 2019)

Outra característica do PostgreSQL é que pode ser usado, modificado e distribuído gratuitamente por qualquer pessoa, seja para fins particulares, comerciais ou acadêmicos. A modelagem da proposta dimensional no software PostgreSQL foi desenvolvida no âmbito entidade-relacionamento, a Figura 4 apresenta o modelo entidade-relacionamento dimensional dos dados hidrológicos.

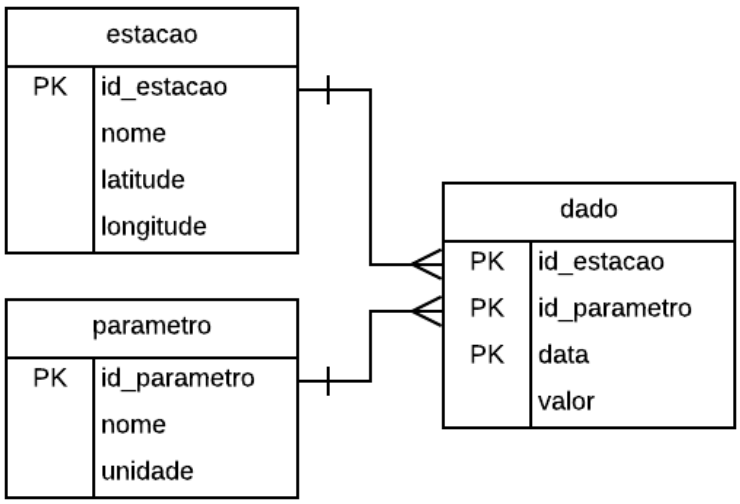

Figura 4 - Modelo entidade-relacionamento dimensional dos dados hidrológicos.

O modelo entidade-relacionamento dimensional dos dados hidrológicos apresentada na Figura 4, é a maneira sistemática de descrever e definir um processo no banco de dados, onde as entidades (tabelas) estão ligadas umas às outras por relacionamentos que expressam as dependências e exigências entre elas. Este modelo leva em consideração a natureza dimensional dos dados de monitoramento pois garante que o dado armazenado sempre estará relacionado com uma estação a um parâmetro e uma data.

A Figura 5 apresenta o código de criação das tabelas e relacionamentos dos dados hidrológicos no banco de dados.

```
CREATE TABLE estacao (
    id_estacao serial NOT NULL,
    nome varchar NOT NULL,
    latitude float  NULL,
    longitude float NULL,
    CONSTRAINT id_estacao_pk PRIMARY KEY (id_estacao)
);

CREATE TABLE parametro (
    id_parametro serial NOT NULL,
    nome varchar NOT NULL,
    unidade varchar NULL,
    CONSTRAINT id_parametro_pk PRIMARY KEY (id_parametro)
);

CREATE TABLE dado (
    id_estacao int NOT NULL,
    id_parametro int NOT NULL,
    data timestamp NOT NULL,
    valor float NOT NULL,
    CONSTRAINT id_dado_pk PRIMARY KEY (id_estacao, id_parametro, data),
    CONSTRAINT dado_estacao_fk FOREIGN KEY (id_estacao)
        REFERENCES estacao(id_estacao) ON UPDATE CASCADE,
    CONSTRAINT dado_parametro_fk FOREIGN KEY (id_parametro)
        REFERENCES parametro(id_parametro) ON UPDATE CASCADE
);
```

Figura 5 - Código de criação das tabelas e relacionamentos dos dados hidrológicos.

O código de criação das tabelas e relacionamentos dos dados hidrológicos apresentado na Figura 5 identificam tabelas que representam as estações, parâmetros e dados, porém os atributos das tabelas são meramente ilustrativos (como latitude, longitude, unidade) podendo ser alterados conforme necessidade. No presente estudo o importante é o relacionamento entre as tabelas. Sendo diferente do armazenamento em uma matriz de 3 dimensões, pois nem todas estações tem dados para todos os parâmetros e nem mesmo para os mesmos períodos de tempo. Portanto essa modelagem garante que serão armazenados somente dados medidos e não haverá dados nulos.

Para testar o modelo dimensional vislumbrou-se um cenário hipotético de banco de dados com 150 estações de monitoramento, 5 parâmetros de medição e 20 anos de dados com intervalo de coleta de 10 minutos cada estação/parâmetro. Os dados foram gerados aleatoriamente e inseridos no banco de dados resultando em aproximadamente 0,8 bilhões de dados e 78 GB. A Tabela 1 apresenta os testes propostos de velocidade das consultas de séries históricas padronizadas e agregações temporais.

Tabela 1- Testes de velocidade de consultas agregadas no banco de dados

Teste	Descrição
Série da média anual	Série da média anual para todos os dados de uma estação e um parâmetro. O teste deve retornar 20 valores, um para cada ano.

	`SELECT date_trunc('year', data)::date, AVG(valor)`
	`FROM dado`
	`WHERE id_estacao = {e_aleatorio} AND id_parametro = {p_aleatorio}`
	`GROUP BY date_trunc('year', data)`
	`ORDER BY date_trunc('year', data);`

Média mensal	Média mensal para todos os dados de uma estação e um parâmetro. O teste deve retornar 12 valores, um para cada mês.

	`SELECT date_part('month', data)::int, AVG(valor)`
	`FROM dado`
	`WHERE id_estacao = {e_aleatorio} AND id_parametro = {p_aleatorio}`
	`GROUP BY date_part('month', data)`
	`ORDER BY date_part('month', data);`

Série da média mensal	Série da média mensal para cinco anos de dados de uma estação e um parâmetro. O teste deve retornar 60 valores, um para cada mês dos cinco anos.

	`SELECT date_trunc('month', data)::date, AVG(valor)`
	`FROM dado`
	`WHERE id_estacao = {e_aleatorio} AND id_parametro = {p_aleatorio}`
	`  AND data >= {dt_aleatoria} AND data < {dt_aleatoria}.AddYears(5)`
	`GROUP BY date_trunc('month', data)`
	`ORDER BY date_trunc('month', data);`

Série da média diária	Série da média diária para um mês de dados de uma estação e um parâmetro. O teste deve retornar 30 valores, um para cada dia do mês.

	`SELECT date_trunc('day', data)::date, AVG(valor)`
	`FROM dado`
	`WHERE id_estacao = {e_aleatorio} AND id_parametro = {p_aleatorio}`
	`  AND data >= {dt_aleatoria} AND data < {dt_aleatoria}.AddMonths(1)`
	`GROUP BY date_trunc('day', data)`
	`ORDER BY date_trunc('day', data);`

Série da média horária	Série da média horária para uma semana de dados de uma estação e um parâmetro. O teste deve retornar 168 valores, um para cada hora dos sete dias.

	`SELECT date_trunc('hour', data)::timestamp, AVG(valor)`
	`FROM dado`
	`WHERE id_estacao = {e_aleatorio} AND id_parametro = {p_aleatorio}`
	`  AND data >= {dt_aleatoria} AND data < {dt_aleatoria}.AddDays(7)`
	`GROUP BY date_trunc('hour', data)`
	`ORDER BY date_trunc('hour', data);`

Todos dados	Todos os dados para um dia de uma estação e um parâmetro. O teste deve retornar 144 valores, um para cada 10 minutos do dia

	`SELECT data, valor`
	`FROM dado`
	`WHERE id_estacao = {e_aleatorio} AND id_parametro = {p_aleatorio}`
	`  AND data >= {dt_aleatoria} AND data < {dt_aleatoria}.AddDays(1)`
	`ORDER BY data;`

Tal modelo de banco de dados proposto por este trabalho foi aplicado ao módulo de monitoramento do SSDPCJ, o qual tem função de divulgar dados de estações de monitoramento telemétrico. Foi desenvolvida interface gráfica web para o módulo utilizando as linguagens HTML,

CSS e JavaScript. A interface com o banco de dados foi realizada por meio de APIs utilizando tecnologia ASP.NET Core.

RESULTADOS

Os testes de velocidade de consultas agregadas no banco de dados apresentados na Tabela 1 foram realizados utilizado PostgreSQL 10 instalado em um servidor Linux, com as seguintes características técnicas: 2 processadores Intel Xeon E5410 a 2.33GHz (8 núcleos de processamento); 9 Gb de memória RAM; 2 discos HD convencional de 7200RPM de 2TB de armazenamento cada (RAID 1); e sistema operacional Debian 9 com configuração padrão. A Tabela 2 apresenta o resultado dos testes.

Tabela 2- Resultado dos testes de velocidade de consultas agregadas no banco de dados

Teste	Tempo t (s)	Dados d	Velocidade d/t (10^3.s^{-1})
Série da média anual	2,976	1051920	353
Média mensal	2,682	1051920	392
Série da média mensal	0,946	262980	278
Série da média diária	0,040	4320	107
Série da média horária	0,032	1008	32
Todos dados	0,028	144	5

Os resultados dos testes de velocidade de consultas agregadas no banco de dados apresentado na Tabela 2 mostra o tempo médio de 1000 consultas consecutivas realizadas para cada teste em segundos, a quantidade de dados processada em cada teste e a velocidade da consulta representada pela quantidade de dados pelo tempo, resultando na unidade mil dados por segundo.

Os resultados mostraram que o modelo é capaz de fazer consultas com agregações complexas em tabelas de proporções enormes (aproximadamente 1 bilhão de dados) com tempo de resposta inferior a um segundo para consultas de até cinco anos de dados. Mesmo utilizando servidor de pequeno porte os resultados demonstram uma performance muito rápida do modelo dimensional aplicado ao banco de dados PostgreSQL, realizando consultas complexas com 20 anos de dados em menos de 3 s com velocidade superior a 300.000 dados processados por segundo. Além de promover consultas de baixa complexidade com tempos de resposta considerados imediatos, inferiores a 0,05 s.

O modelo de banco de dados desenvolvido neste trabalho foi aplicado ao módulo de monitoramento do SSDPCJ, sendo implementada interface gráfica para permitir a fácil geração das consultas agregadas, além da apresentação das informações em tabelas e gráficos de forma prática e rápida, o que acelera o entendimento dos problemas e a tomada de decisões. Foram implementados

dois tipos de resultados: resumo do monitoramento (Figura 6) e série temporal do monitoramento (Figura 7).

Postos		Adicionar	Excluir	Excluir	Excluir
			PLU (mm)	Q (m3/s)	Q (m3/s)
			Hoje	Hoje	Últimos 15 Dias
			Instantâneo	Média	Média
D3-040T/3D-009T	Rio Jaguari em Buenópolis		5,40	2,63	4,38
E3-111T/3E-063T	Rio Atibaia em Atibaia		0,80	8,20	19,01
D3-051T/3D-007T	Rio Atibaia Captação Valinhos		0,00	25,23	48,03

Figura 6 - Tabela de resumo do monitoramento no SSDPCJ

A tabela de resumo do monitoramento no SSDPCJ apresentada na Figura 6, constitui informações agregadas por estação monitoramento nas linhas e parâmetro monitorado, data e tipo de agregação por coluna. Neste exemplo os dados apresentados para três estações de monitoramento para chuva e vazão de hoje e vazão média dos últimos 15 dias. Vale ressaltar que essa ferramenta permite adicionar ou remover linhas e colunas, gerando relatórios genéricos.

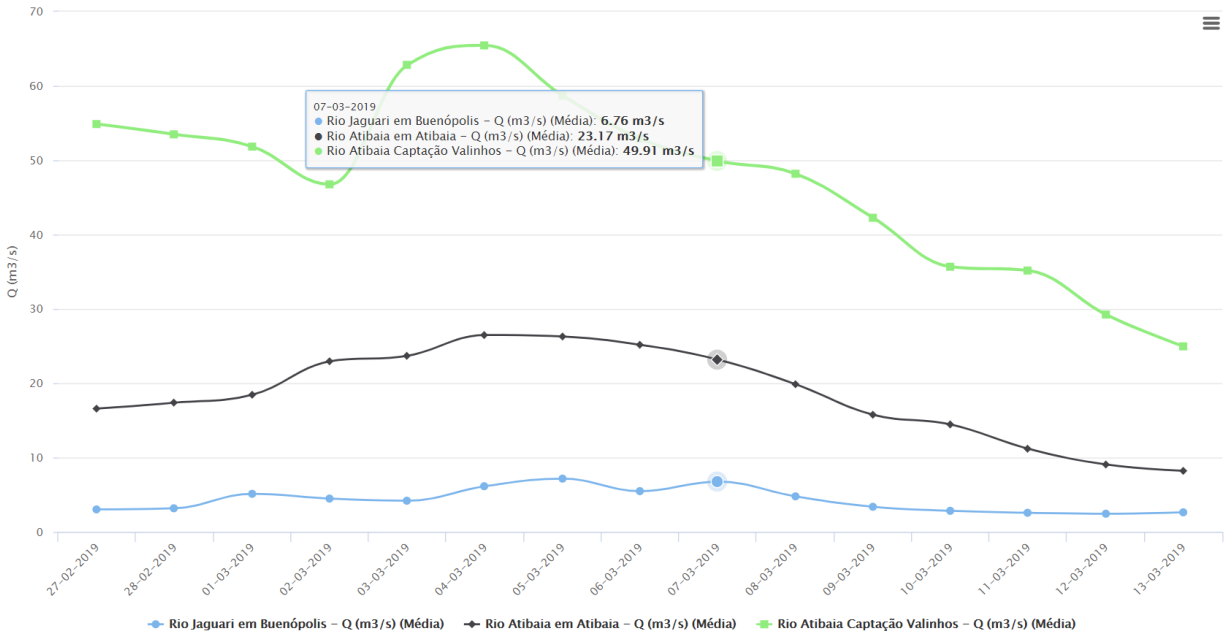

Figura 7 - Gráfico de série temporal do monitoramento no SSDPCJ

O gráfico de série temporal do monitoramento no SSDPCJ apresentado na Figura 7, o eixo y representa a intersecção de uma estação e um parâmetro e o eixo x o tempo, que pode estar na base horária, diária, mensal, anual ou sem discretização. Neste exemplo foram selecionadas três estações com o parâmetro vazão a base de tempo diária, a agregação dos dados pela média num período de 15 dias.

CONCLUSÃO

Conclui-se que o modelo dimensional aplicado em recursos hídricos tem grande potencial de uso devido a sua generalização na entrada de dados e fácil agregação dos dados para geração de informações. Foi demonstrada a aplicação deste modelo no PostgreSQL, que é um motor de banco de dados potente e gratuito, ou seja, não gera custos com licenças para sua implementação.

Mostrou-se viabilidade mesmo em ambientes com servidores de pequeno porte, onde o modelo foi capaz de realizar consultas complexas em base de dados de aproximadamente 1 bilhão de dados com tempo de resposta aceitável a sistema de informação, menos de 3 segundos, com velocidade superior a 300.000 dados processados por segundo. Além de realizar consultas de baixa complexidade com tempos de resposta considerados imediatos, inferiores a 0,05 segundos.

A Agência das Bacias PCJ teve êxito na aplicação do modelo dimensional no SSDPCJ, permitindo a integração dos dados de diversas fontes e gerando resultados de resumo (informações agregadas por estação monitoramento, parâmetro monitorado, data e tipo de agregação) e série temporal (informação agregada em uma base de discretização temporal por estação e parâmetro).

AGRADECIMENTOS

Os autores agradecem ao Laboratório de Sistemas de Suporte a Decisões em Engenharia Ambiental e de Recursos Hídricos, LabSid, ao Departamento de Engenharia Hidráulica e Ambiental da Escola Politécnica da Universidade de São Paulo e à Fundação Centro Tecnológico de Hidráulica, FCTH, pelo ambiente criativo que apoia pesquisa e desenvolvimento. E a Agência das Bacias PCJ, pelo financiamento e oportunidade de aplicar pesquisa e desenvolvimento em problemas reais de engenharia de recursos hídricos.

REFERÊNCIAS

ANA, Agência Nacional de Águas (2010). "*HIDRO: Sistema de Informações Hidrológicas Versão 1.0*".

ANA, Agência Nacional de Águas (2019). "*Inventario de Estacoes Hidrometeorologicas*". Disponível em: <http://www.snirh.gov.br/hidroweb/publico/baixar_documento.jsf>. Acesso em: 26 mar. 2019.

KIMBALL, R. e ROSS, M. (2013). "*The data warehouse toolkit: The definitive guide to dimensional modeling*". John Wiley & Sons.

PostgreSQL (2019). "*PostgreSQL Documentation*". Disponível em: <https://www.postgresql.org/docs>. Acesso em: 26 mar. 2019.

Terakawa, A. (2003). "*Hydrological data management: Present state and trends*". Secretariat of the World Meteorological Organization. WMO.