Ciência Cidadã Baseada em Big Data Aplicada ao Planejamento Urbano

Eliza Gomes[1], M.A.R. Dantas[1], Douglas D. J. de Macedo[2]
Júlio Dias[3], Carlos De Rolt[3], Marcelo Brocardo[3], Luca Foschini[4]

[1]Departamento de Informática e Estatística
Universidade Federal de Santa Catarina (UFSC)
Caixa Postal 476 – 88.040-900 – Florianópolis – SC – Brasil

[2]Departamento de Ciência da Informação
Universidade Federal de Santa Catarina (UFSC)
88040-900 – Florianópolis – SC – Brasil

eliza.gomes@posgrad.ufsc.br, {mario.dantas, douglas.macedo}@ufsc.br

[3]Centro de Ciências da Administração e Socioeconômicas
Universidade do Estado de Santa Catarina (UDESC)
Av. Madre Benvenuta, 2037 – 88.035-001 – Itacorubí – Florianópolis – SC

{julio.dias,rolt}@udesc.br, marcelo.luiz.brocardo@gmail.com

[4]Universidade de Bologna (DISI), Bologna, Itália

luca.foschini@unibo.it

Abstract. *In this paper we presented a model and an architecture for building and structuring environments that need to store, search and analyze, efficiently, of large amount of data. For the implementation of the model we used ParticipACT Brazil project. This project aims to deploy in Florianópolis city the citizen science process to obtain relevant information to help managers to solve urban problems. We can be observed that the implementation of the structure has facilitated the environment and helped in choice of specialized tools in manipulating big data.*

Resumo. *Neste artigo são apresentados um modelo e uma arquitetura para construção e estruturação de ambientes que necessitam realizar, eficientemente, armazenamento, busca e análise de grandes volumes de dados. Para a implementação do modelo foi utilizado o projeto ParticipACT Brasil. Este projeto visa implantar na cidade de Florianópolis o processo de ciência cidadã para obter informações relevantes que auxiliem os gestores a resolverem problemas urbanos. Pode-se observar, então que a implementação do modelo facilitou a estruturação do ambiente, e ajudou na escolha de ferramentas especializadas na manipulação de big data.*

1. Introdução

O constante uso da internet em diferentes equipamentos e lugares (*Internet of Things*) faz com que a implementação de cidades inteligentes tenha um aumento significativo

[Jara et al. 2014]. Uma cidade com sensores pode obter uma variedade de dados e, consequentemente, gerar uma grande quantidade de informações que podem ser utilizadas tanto pelos cidadãos quanto pelos administradores para auxiliar na tomada de decisões [Moreno-Cano et al. 2015]. Além disso, a interação entre dispositivos de tecnologia da informação e comunicação (TIC) e os cidadãos pode contribuir com o desenvolvimento e a manutenção da cidade inteligente. Por exemplo, informações cedidas pelos habitantes sobre a situação do tráfego em uma determinada rua e horário pode ajudar a tornar o sistema de transporte mais inteligente [Bicocchi et al. 2013]. Esta grande quantidade e variedade de dados gerados pelas cidades inteligentes cria um *big data*.

Adicionalmente, dados de entidades prestadoras de serviços públicos podem ser disponibilizados para integrar o *big data*, pois sua utilização pode ajudar na resolução de problemas urbanos. No entanto, por serem fornecidos por diferentes fontes, esses dados podem apresentar uma variedade de formatos. Para resolver isso, são necessárias ferramentas de integração de dados. Em outras palavras, ferramentas que transformem dados diferentes em um único formato a fim de possibilitar o armazenamento em uma única base de dados.

Contudo, o aumento de técnicas e ferramentas para *big data* [Inacio and Dantas 2014] torna o processo de implementação de infraestrutura computacional um desafio, uma vez que é necessário conhecer e escolher as ferramentas mais adequadas para o ambiente, de acordo com a quantidade e o tipo de dado. Uma solução seria desenvolver um modelo de infraestrutura de *big data* que apresente como os dados devem se comportar no ambiente e quais os tipos de ferramentas que devem ser usadas e escolhidas. Entretanto, existem poucos artigos que apresentam um modelo de plataforma para *big data* [Cheng et al. 2015], [Ma and Liang 2015].

Para resolução deste problema, é proposto nesse artigo um modelo de infraestrutura para *big data*. O principal objetivo do modelo é estruturar um ambiente para receber e trabalhar eficientemente com grandes quantidades de dados. O modelo também visa fornecer as tecnologias que atendam aos requisitos para o desenvolvimento do ambiente dando autonomia para o desenvolvedor na escolha de ferramentas.

Para implementar o modelo foi realizado um estudo de caso utilizando como base o projeto ParticipACT Brasil [ParticipACT 2016b]. Este projeto tem como base os conceitos de cidades inteligentes (*smart cities*), bem como o objetivo de projetar, promover e desenvolver um sistema de gerenciamento sócio-técnica para formar gradual e progressivamente um *big data* para analisar problemas de uma área urbana.

Este artigo tem a seguinte estrutura organizacional: na Seção 2 é feita uma descrição do projeto ParticipACT Brasil, detalhando seus objetivos e estrutura; na Seção 3 é detalhado o modelo apresentado nesse artigo; um estudo de caso realizado com o projeto ParticipACT Brasil é apresentado na Seção 4; na Seção 5 são apresentados alguns trabalhos relacionados à proposta deste artigo; por fim, na Seção 6 são apresentadas as conclusões e indicações para trabalhos futuros.

2. Projeto ParticipACT Brasil

ParticipACT Brasil [ParticipACT 2016b] é uma extensão da plataforma *crowdsensing* ParticipACT [ParticipACT 2016a] da Universidade de Bologna. Este projeto está sendo

desenvolvido pela Universidade do Estado de Santa Catarina em parceria com a Universidade Federal de Santa Catarina e recebe apoio financeiro e de disponibilização de dados de entidades públicas e privadas. O objetivo do ParticipACT Brasil é utilizar os dados disponibilizados por entidades que prestam serviços básicos para os habitantes e promover "campanhas"de *crowdsensing*, para que os habitantes da cidade, voluntariamente, enviem e disponibilizem informações que possam auxiliar os gestores na resolução de problemas urbanos. Em outras palavras, com o projeto ParticipACT será possível desenvolver o processo de ciência cidadã, não somente com a ajuda dos habitantes, mas também com a ajuda de instituições públicas e privadas com a disponibilização de dados úteis para o desenvolvimento e aprimoramento dos serviços oferecidos aos cidadãos [Gomes et al. 2016].

O projeto está sendo desenvolvido, inicialmente, na cidade de Florianópolis/SC. O município de Florianópolis foi escolhida por ser a cidade sede das universidades, turística e com problemas de mobilidade e infraestrutura, o que aumenta a necessidade de um estudo concreto para a resolução desses problemas.

ParticipACT Brasil tem como objetivo incial apresentar evidências científicas que determinem, com maior exatidão possível, a quantidade de turistas que visitam a cidade, ou seja, a população flutuante, principalmente, nos meses de Dezembro a Janeiro. Para isso, pretende-se em um primeiro momento obter dados de instituições fornecedoras de serviços básicos como: companhia de coleta de resíduos sólidos, companhia de distribuição de energia elétrica, companhia de água e saneamento, secretarias da prefeitura do município e censo (para determinar o número de habitantes).

A infraestrutura do ParticipACT Brasil, conforme mostra a Fig. 1, é dividida em três partes principais: *Big Data*, *Crowdsensing* e *Website*.

- **Big Data**: é formado pelo *Database Server*, *Server* e *Big Data Server*. *Database Server* é responsável por armazenar as informações capturadas pelos cidadãos através de "campanhas"de *crowdsensing*. *Server* é um servidor fornecido pelo ParticipACT da Bologna/Itália responsável por gerenciar as "campanhas"de *crowdsensing*. Finalmente, o *Big Data Server* integra as bases de dados, Database Server e dados provenientes de diferentes instituições e correlaciona esses dados para gerar informações importantes para tomada de decisão.
- **Crowdsensing**: neste está incluso o componente *App - Phone* e o *Server*. O *App - Phone* coleta dados resultantes das campanhas de participação e cooperação voluntária dos cidadãos. O *crowdsensing* é realizado por meio de smartphones (incialmente para sistemas Android e IOS), com acesso ao aplicativo do ParticipACT Brasil, para coleta de dados. O objetivo de se gerenciar uma campanha de *crowdsensing* é coordenar um grupo de pessoas para coletar um certo tipo, e talvez complexo, de dados. Os participantes do *crowdsensing* podem acessar as informações coletadas nas campanhas através portal.
- **Website**: é o portal disponibilizado para a interação entre o usuário e o sistema ParticipACT. Esse *website* permite que o usuário visualize os resultados das análises dos dados, faça download do aplicativo, entre outros.

Diante disso, o ParticipACT Brasil está diretamente relacionado ao processo de ciência cidadã, uma vez que o projeto incentiva cidadãos, agências governamentais, em-

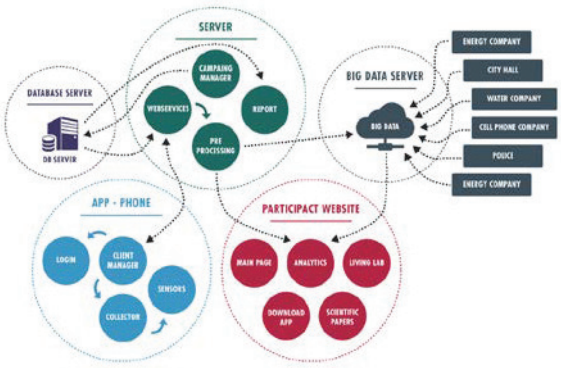

Figura 1. Infraestrutura do projeto ParticipACT Brasil [Gomes et al. 2016]

presas privadas e academia, a participarem de campanhas e assim fornecerem dados úteis para pesquisas e estudos [Conrad and Hilchey 2011].

3. Proposta

Nessa seção são apresentados o modelo e a arquitetura para um ambiente que necessite manipular grandes quantidades de dados.

Big Data é um conjunto de dados caracterizado pelo armazenamento de grandes volumes e variedades de dados, bem como pela velocidade e flexibilidade que devem ser oferecidas para manipulação dos dados como armazenamento, consulta e análise [Kitchin 2013]. Diante disso, um modelo e arquitetura de plataforma para *big data* se torna importante uma vez que o uso correto e estruturado de ferramentas determina positivamente o uso eficiente dos dados.

O principal objetivo do modelo é estruturar um ambiente que atenda aos requisitos exigidos por uma plataforma para *big data*, ou seja, volume e variedade de dados, velocidade de busca e armazenamento, flexibilidade para manipulação dos dados e possibilidade de análise sistemática dos dados.

A Figura 2 apresenta o modelo composto por cinco camadas, bem como a arquitetura desenvolvida com base no modelo proposto.

1. A primeira camada consiste dos dados fornecidos pelas instituições que são armazenados na base de dados. Diferentes instituições fornecem dados de diferentes tipos e formatos. Os tipos mais comuns são: txt, csv, planilha e base de dados relacional.
2. A segunda camada representa a interação dos dados para que sejam armazenados em uma única base de dados. Para converter os dados em um único formato é necessário utilizar uma ferramenta para extrair, transformar e carregar (*Extract, Transformation and Load* - ETL) dados. ETL, por sua vez, é um mecanismo utilizado para migrar dados heterogêneos de uma ou mais fontes de dados para um repositório de dados, *data marts* ou *data warehouse* [Albrecht and Naumann 2008]. A ferramenta utilizada, inicialmente, para o desenvolvimento da arquitetura foi o Pentaho Data Integration (também conhecido como Kettle) [Pentaho 2016].
3. A terceira camada apresenta a base de dados. De acordo com o tipo e quantidade de dados, foi utilizado a base de dados NoSQL Apache Cassandra

[Cassandra 2016]. Cassandra é um sistema de gerenciamento de banco de dados distribuído, orientado a coluna, para gerenciamento de grandes quantidades de dados estruturados distribuídos por muitos servidores.

4. A quarta camada representa a etapa de processamento de dados. Nesta fase os dados são relacionados, manipulados e analisados. A ferramenta utilizada foi o Programa R, um ambiente para análises e gráficos estatísticos. Este programa possui um pacote, RCassandra, que possibilita a conexão entre a base de dados e o programa R [R 2016].
5. A última camada consiste de uma interface amigável que apresenta ao usuário as informações geradas. Essas informações podem ser visualizadas através de gráficos, mapas ou desenhos.

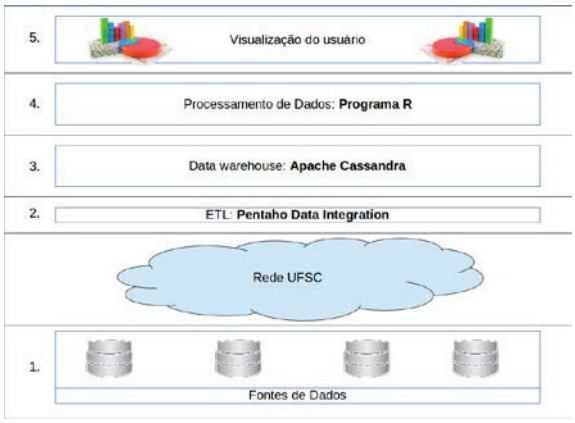

Figura 2. Modelo e Arquitetura de infraestrutura para ambientes de big data

As ferramentas Kettle e Cassandra foram escolhidas para os experimentos iniciais porque o Kettle apresenta uma interface amigável e todas as funcionalidades necessárias para a transformação e integração dos dados. Além disso, possui uma versão para a comunidade *open source*. Por outro lado, o Apache Cassandra foi escolhido, pois além de ser uma base de dados NoSQL colunar, se destaca pela boa escalabilidade, alta disponibilidade, baixa latência e por permitir a replicação dos dados em múltiplos *datacenters* [Deka 2014].

4. Estudo de Caso

Para implementar o modelo e arquitetura apresentados, foi utilizado o projeto ParticipACT Brasil, descrito na seção 2. Um dos objetivos desse projeto é receber dados de companhias prestadoras de serviços públicos e analisá-los para fornecer resultados científicos relevantes para os administradores e para a população de um modo geral. O objetivo específico desse estudo de caso é realizar uma comparação entre dois bairros, com a quantidade de habitantes relativamente parecida, da cidade de Florianópolis (Bairro C e Bairro T). O Bairro C é um bairro que atrai muitos turistas, pois oferece amplo serviço de hotéis e restaurantes, além de estar localizado em uma região cuja praia possui água com temperatura mais agradável e ampla faixa de areia. Já o Bairro T é mais residencial, ou seja, não possui grandes quantidades de hotéis ou restaurantes, por isso não é frequentado por turistas. O objetivo da comparação entre esses bairros é de verificar a quantidade

de resíduos sólidos produzidos durante o ano de 2015, considerando as características de cada bairro.

O funcionamento da arquitetura pode ser visto na Fig. 2. Os dados recebidos do provedor estão, normalmente, em formatos csv ou txt. Por isso, os dados são carregados no Kettle, que os transforma, em um formato compatível com o Cassandra, e os envia para a base de dados para armazenamento. Uma vez armazenados, os dados são capturados pelo Programa R e, a partir disso, podem ser manipulados. Após a manipulação, a informação gerada a partir dos dados é apresentada para o usuário final através de gráficos ou na forma escrita por meio de um *website*.

A Figura 3 apresenta o gráfico comparativo entre os Bairro C e Bairro T, no qual é possível observar a variação da produção de resíduos sólidos no decorrer do ano. Entretanto, está em fase de desenvolvimento um conjunto de algoritmos para avaliação dos resultados (conforme aqueles apresentados na Fig. 3).

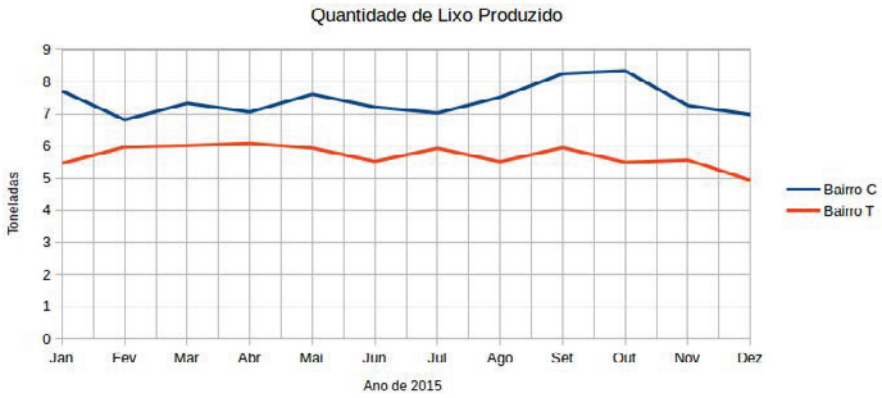

Figura 3. Quantidade de resíduos sólidos produzidos em dois bairros de Florianópolis

Outras análises poderiam ser realizadas para que seja possível explicar o motivo dessa variação, como por exemplo, se o aumento da produção de resíduos sólidos no Bairro T no mês de Julho se deve ao período de férias escolares ou se existe alguma relação entre a primavera, a safra de frutas e verduras e o aumento da produção de resíduos sólidos nos meses de Setembro e Outubro. Vale ressaltar que o objetivo desse estudo de caso é implementar o modelo e a arquitetura propostos e verificar o seu funcionamento em uma situação real. Em outras palavras, o foco não é proporcionar uma análise estatística ampla e completa.

4.1. Experiências

Para o desenvolvimento do ambiente de *big data* foram encontradas algumas dificuldades no que diz respeito às versões e instalações dos softwares utilizados. Em relação às ferramentas Kettle e Apache Cassandra, a versão atual (até a escrita desse artigo) do Kettle 6.0, não é compatível com a versão atual do Cassandra (3.4). Para que pudesse ser realizada a integração dos dois softwares foi necessário instalar o Cassandra 2.0.17.

Esses softwares exigem uma grande quantidade de memória disponível e dedicada à sua execução. Em outras palavras, computadores simples de laboratório não são

suficientes para que a arquitetura proposta seja testada.

5. Trabalhos Relacionados

O artigo de [Cheng et al. 2015] apresenta o *City Data and Analytics Platform* (CiDAP), uma arquitetura para trabalhar com plataformas para *big data*. CiDAP processa tanto dados históricos quanto de tempo-real, enquanto expõe dados para várias aplicações. A arquitetura proposta projeta um ambiente para *big data* e descreve as ferramentas que foram utilizadas para a construção da plataforma. Os testes para validar a plataforma CiDAP foram realizados na cidade de Santander na Espanha, uma das maiores cidades inteligentes para testes chamada SmartSantander.

Por outro lado, o artigo de [Ma and Liang 2015] apresenta uma arquitetura distribuída de plataforma para *big data* para cidades inteligentes. Nessa arquitetura os autores utilizaram ferramentas como Kafka, Storm e Spark para a implementação de transmissão em tempo real de dados estruturados e não estruturados e bases de dados. Adicionalmente, desenvolveram um sistema de análise estatística para dados em tempo real e para dados *offline*.

No artigo de [Khan and Kiani 2012] é proposta uma arquitetura baseada em nuvem computacional para armazenar e processar informações fornecidas pelos habitantes de uma cidade inteligente (através do processo de ciência cidadã). A proposta inclui a possibilidade do cidadão tanto fornecer dados, através do uso de smartphones, quanto acessar informações contextuais, pré-definidas pelo usuário, a partir de um sistema integrado.

Os trabalhos mencionados apresentam uma arquitetura desenvolvida para ambientes e com ferramentas específicas. A proposta desse artigo se diferencia, pois apresenta um modelo não condicionado à ferramentas ou ambientes, dando assim maior autonomia ao desenvolver na escolha da arquitetura.

6. Considerações Finais e Trabalhos Futuros

Este artigo apresentou um modelo e uma arquitetura de infraestrutura para armazenamento, busca e análise eficientes para grandes volumes de dados. Para a implementação inicial da proposta foi utilizado o projeto ParticipACT Brasil como estudo de caso. Este projeto visa unir informações fornecidas tanto pelos cidadãos, através de campanhas de *crowdsensing*, quanto por empresas públicas e privadas, através da disponibilização voluntária de seus dados. O modelo e a arquitetura propostos estruturou de maneira eficiente o ambiente de armazenamento e análise de dados. Pôde-se observar também, que o modelo ofereceu certa autonomia para a escolha de outras ferramentas.

Como trabalhos futuros, pretende-se utilizar outras fontes de dados, como a companhia de água e saneamento, de distribuição de energia elétrica, dados do censo, bem como outras ferramentas no modelo proposto para ratificar sua autonomia e eficiência. Quanto ao aspecto de rastreabilidade dos resultados das análises pretende-se desenvolver um portal que apresente para o usuário final os dados e a forma de análise utilizada para sua manipulação. Pretende-se também realizar um estudo comparativo das ferramentas, de modo que seja possível determinar qual a ferramenta mais apropriada para o ambiente e os requisitos específicos, realizar análises mais completas dos dados, além de inserir ao modelo restrição temporal para que as respostas sejam fornecidas dentro de um tempo pré-determinado.

Referências

Albrecht, A. and Naumann, F. (2008). Managing etl processes. *NTII*, 8:12–15.

Bicocchi, N., Cecaj, A., Fontana, D., Mamei, M., Sassi, A., and Zambonelli, F. (2013). Collective awareness for human-ict collaboration in smart cities. In *Enabling Technologies: Infrastructure for Collaborative Enterprises, 2013 IEEE 22nd International Workshop on*, pages 3–8. IEEE.

Cassandra (2016). Apache cassandra. http://www.planetcassandra.org/. Acessado: Março.

Cheng, B., Longo, S., Cirillo, F., Bauer, M., and Kovacs, E. (2015). Building a big data platform for smart cities: Experience and lessons from santander. In *Big Data, 2015 IEEE International Congress on*, pages 592–599. IEEE.

Conrad, C. C. and Hilchey, K. G. (2011). A review of citizen science and community-based environmental monitoring: issues and opportunities. *Environmental monitoring and assessment*, 176(1-4):273–291.

Deka, G. C. (2014). A survey of cloud database systems. *IT Professional*, (2):50–57.

Gomes, E., Dantas, M., Macedo, D. D. J., Rolt, C. D., Brocardo, M., and Foschini, L. (2016). Towards an infrastructure to support big data for a smart city project. In *25th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises*. Artigo aceito pelo evento.

Inacio, E. C. and Dantas, M. A. (2014). A survey into performance and energy efficiency in hpc, cloud and big data environments. *International Journal of Networking and Virtual Organisations*, 14(4):299–318.

Jara, A. J., Genoud, D., and Bocchi, Y. (2014). Big data in smart cities: from poisson to human dynamics. In *Advanced Information Networking and Applications Workshops, 2014 28th International Conference on*, pages 785–790. IEEE.

Khan, Z. and Kiani, S. L. (2012). A cloud-based architecture for citizen services in smart cities. In *Proceedings of the 2012 IEEE/ACM 5th International Conference on Utility and Cloud Computing*, pages 315–320. IEEE Computer Society.

Kitchin, R. (2013). Big data and human geography opportunities, challenges and risks. *Dialogues in human geography*, 3(3):262–267.

Ma, S. and Liang, Z. (2015). Design and implementation of smart city big data processing platform based on distributed architecture. In *Intelligent Systems and Knowledge Engineering, 2015 10th International Conference on*, pages 428–433. IEEE.

Moreno-Cano, V., Terroso-Saenz, F., and Skarmeta-Gomez, A. F. (2015). Big data for iot services in smart cities. In *Internet of Things, 2015 IEEE 2nd World Forum on*, pages 418–423. IEEE.

ParticipACT (2016a). Participact. http://participact.unibo.it/. Acessado: Abril.

ParticipACT (2016b). Participact brasil. http://labges.esag.udesc.br/participact/. Acessado: Abril.

Pentaho (2016). Pentaho data integration. http://www.pentaho.com/. Acessado: Março.

R (2016). R project. https://www.r-project.org/. Acessado: Março.