PROCESSAMENTO DE BIG DATA URBANO

CIDADES INTELIGENTES PRODUZEM GRANDES QUANTIDADES DE DADOS EM PERÍODOS MUITO CURTOS DE TEMPO. PARA OBTER INFORMAÇÕES EM TEMPO REAL E EXTRAIR CONHECIMENTO DESSES DADOS, SÃO NECESSÁRIAS TÉCNICAS EFICIENTES DE ARMAZENAMENTO E PROCESSAMENTO.

por **Kelly Rosa Braghetto, Daniel Cordeiro e Alfredo Goldman**

Big Data Urbano

Uma grande quantidade de dados provenientes das mais variadas fontes é gerada continuamente no contexto de uma cidade. As fontes mais comuns são dispositivos eletrônicos com capacidade de sensoriamento e poder computacional, que capturam dados (muitas vezes, automaticamente e periodicamente) e os armazenam ou transmitem por meio de uma rede, formando assim a Internet das Coisas. Exemplos disso são os sensores ambientais, os GPSs, as câmeras de segurança, os *smartphones*, etc. Outros dados sobre a cidade também são gerados em redes sociais e sistemas de governo eletrônico. A análise desses dados é fundamental para identificar deficiências das cidades e amparar políticas públicas que visem à qualidade de vida dos cidadãos e ao uso sustentável da infraestrutura e dos recursos naturais.

Fontes e estruturas heterogêneas, coletas em alta frequência, grande volume, alto valor social e econômico: essas características fazem dos dados de cidades um exemplo perfeito de Big Data. Se por um lado é importante destacar o potencial do Big Data urbano para alavancar serviços em cidades inteligentes, por outro é essencial ressaltar que o seu uso envolve diversos desafios. Eles incluem, por exemplo, o tratamento de problemas relacionados a privacidade dos cidadãos, validade temporal e espacial dos dados, imprecisão dos dispositivos de coleta e transparência das análises (para evitar possíveis manipulações de opinião). Sobrepondo-se a esses desafios, tem-se ainda o complexo problema de armazenar e processar, de forma eficiente, enorme quantidade de dados.

Se por um lado é importante destacar o potencial do Big Data urbano para alavancar serviços em cidades inteligentes, por outro é essencial ressaltar que o seu uso envolve diversos desafios.

Armazenar, Processar, Usar

O processamento de Big Data pode ser feito em lotes ou em fluxos de dados. No processamento em lotes (*batch processing*), dados previamente coletados e armazenados são processados, o que pode levar horas no caso de lotes grandes. Já no processamento de fluxos de dados (*stream processing*), os dados são processados à medida que chegam à aplicação, gerando resultados com baixa latência. Arcabouços[1] populares para processamento em lotes são o Hadoop e o Spark, enquanto que o Storm, o Sanza e o Flink são bastante usados no processamento de fluxos de dados.

Para processamentos mais complexos e que envolvem uma quantidade muito grande de dados, é necessária uma arquitetura que permita o armazenamento resiliente, escalável e que possa ser processado de forma eficiente.

Algumas aplicações requerem o uso combinado de processamento de lotes e de fluxos de dados. Por exemplo, um sistema de monitoramento de trânsito precisa identificar e reportar com rapidez a ocorrência de acidentes, bem como cruzar informações históricas que permitirão prever as zonas mais perigosas e evitar novas situações de risco. Diferentes arquiteturas de software (como a Lambda e a Kappa) foram propostas para lidar com esse tipo de demanda.

Cada camada de uma plataforma para cidades inteligentes pode utilizar um tipo de sistema de informação diferente. Por exemplo, um *context broker*, responsável pela coleta, redistribuição de dados e disparo de novos eventos, normalmente usa um banco de dados NoSQL para minimizar a latência ao acesso a dados recentes. Para processamentos mais complexos e que envolvem uma quantidade muito grande de dados, é necessária uma arquitetura que permita o armazena-

mento resiliente, escalável e que possa ser processado de forma eficiente. Uma solução popular é o uso do sistema de arquivos distribuídos HDFS (Hadoop Distributed File System) em conjunto com modelos de programação distribuídos como o MapReduce (usado no Apache Hadoop) ou baseados em *dataflows* (como o proposto pelo Apache Spark). Os resultados de processamentos complexos podem ser oferecidos de forma conveniente por sistemas gerenciadores de bancos de dados relacionais.

Próximos passos

Vários são os desafios de pesquisa que emergem do gerenciamento de dados de cidades. **Segurança e privacidade:** apesar de os dados serem gerados por usuários anônimos, podem-se inferir ou restaurar informações pessoais por meio de mineração de dados. Além disso, os dados nessas plataformas são gerados de forma colaborativa; será preciso desenvolver um modo de autenticar os dispositivos que criam e processam os dados antes que eles possam operar sobre a plataforma. **Mineração:** a grande variedade, velocidade e volume dos dados impõem novos requisitos aos algoritmos de mineração de dados e salientam a necessidade de novos modelos de programação distribuída de alto desempenho. **Visualização:** visualizar os dados coletados é difícil devido ao seu volume e dimensionalidade. A maior parte das soluções atuais estão aquém do esperado em termos de funcionalidades, escalabilidade e tempo de resposta. **Integração:** os dados devem ser vistos de forma uniforme, apesar de terem fontes e estruturas variadas; essa integração de dados continua um problema em aberto.

Referências

1 Todos os arcabouços citados são projetos da Fundação Apache (https://www.apache.org/).

KELLY ROSA BRAGHETTO | É professora doutora do Departamento de Ciência da Computação do IME -USP e pesquisadora associada ao INCT da Internet do Futuro para Cidades Inteligentes. Atua nas áreas de Bancos de Dados, Gerenciamento de Workflows e Métodos Formais para Projeto e Análise de Sistemas.

DANIEL CORDEIRO | É professor doutor da Escola de Artes, Ciências e Humanidades da Universidade de São Paulo. Ele recebeu o título de doutor em Mathématiques et en Informatique pela Université de Grenoble, França, e de mestre em Ciência da Computação pela Universidade de São Paulo. Seus principais interesses incluem Computação de Alto Desempenho e Teoria do Escalonamento.

ALFREDO GOLDMAN | É professor associado da Universidade de São Paulo. Possui mestrado em Matemática Aplicada pela Universidade de São Paulo (1994) e doutorado em Informatique et Systèmes - Institut National Polythecnique De Grenoble (1999). Atua nos seguintes temas: Computação paralela e distribuída, escalonamento e métodos ágeis de desenvolvimento de software.