

Cidades Inteligentes & Big Data

Tecnologias de Big Data como habilitadoras das Cidades do Futuro - Uma abordagem top-down

Anderson dos Santos Paschoalon

Universidade Estadual de Campinas,
Faculdade de Engenharia Elétrica e Computação,
DCA Departamento de Computação e Automação
INTRIG Information & Networking Technologies Research & Innovation Group
anderson.paschoalon@gmail.com

Resumo Este trabalho tem por objetivo discutir os temas de cidades inteligentes e big data, inseridos dentro do mesmo contexto, e mostrar os inúmeros pontos em que eles se interrelacionam. É discutido como as tecnologias de big data tratam-se se um habilitados fundamental para que cidades inteligentes possam ser de fato implantada. Ambos os conceitos são formalmente introduzidos, sendo o conceito de cidades inteligentes discutido inserido dentro da arquitetura e mapa tecnológico de cadeia de valor do big data. Por fim, é apresentado um exemplo prático de implementação de uma infraestrutura de big data de uma cidade inteligente: o CiDAP, em Santander, na Espanha.

Keywords: Big data, cidades inteligentes, big data *analytics*, gerenciamento de big data, armazenamento de big data, *machine learning*, mineração de dados, NoSQL, CouchDB, GFS, Hadoop, HDFS, Infraestrutura de big data, modelos de programação, sistemas de arquivos distribuídos, IoT, IoE, IPaaS, APaaS, *middleware*, Civitas, *civitas plug*, redes de sensores, dispositivos móveis, transporte inteligente, dados de máquina, dados abertos, CiDAP

1 Introdução

O advento das cidades inteligentes é um fato quase que inexorável para o nosso futuro. Não é mais uma questão se e como, mas quando. Toda a tecnologia necessária para a aplicação dos conceitos que já estão sendo estudados pela academia há um tempo razoável, já existem. Todas as ferramentas para a etapa de aquisição de dados, processamento de informações e inteligência já existem, mas aplicadas em outros campos. O maior desafio da atualidade é ser capaz de organizar as tecnologias existentes, melhorá-las para o cenário atual, e implantá-las. Para que a implantação seja possível, é necessário que haja tanto uma demanda por setores da sociedade, como pessoas e empresas; quanto um esforço conjunto da iniciativa pública, privada e da academia em aplicar tais ideias e tecnologias no mundo real.

Trata-se aqui como uma cidade inteligente, uma infraestrutura digital de computação e telecomunicações que permeia toda a cidade, coletando informações do ambiente e da sociedade, e sendo capaz de reagir de maneira inteligente, respondendo proativamente as informações captadas. Tudo isso em prol de maximizar a utilidade e o bem-estar das pessoas que vivem na cidade, através dos novos serviços, e do uso de tecnologia cognitiva nos serviços já existentes. Este trabalho possui o foco em analisar tecnologias de big data dentro do contexto das cidades inteligentes. De fato, big data pode ser considerado um dos conceitos chaves indispensáveis para que a implantação efetiva das cidades inteligentes possa se tornar real.

Sistemas computacionais convencionais, apesar de ter se tornado algo munido, são sistemas bastante complexos, construídos por meio de várias camadas de abstração. No nível mais baixo o hardware físico, barramentos e discos rígidos. Em seguida um firmware, responsável por programar diretamente o hardware. Sobre este é desenvolvido um sistema operacional, responsável por abstrair e unificar todos os recursos, fornecendo interfaces simples às aplicações que sobre ele irão trabalhar. O objetivo de uma cidade inteligente é criar um ambiente semelhante a esse que permeie toda a cidade. Na camada física, teríamos sensores, câmeras, dispositivos móveis, e dados públicos, servindo como fonte de informação. Sobre ele, uma camada de interoperabilidade, responsável por consolidar todas essas tecnologias, e da mesma forma ser capaz de fornecer serviços de inteligência a todos eles. Tecnologias de big data, viriam então como responsáveis por lidar com todo o armazenamento dessas informações, e oferecer metodologias de processamento e extração de informação de quantidades enormes de dados. Dessa forma, aplicações e serviços poderiam ser desenvolvidos, utilizando-se dessa informação coletada, por meio de simples APIs.

Esse trabalho segue uma abordagem top-down com relação a uma das representações das tecnologias de big data, chamada cadeia de valor (*value chain*). Essa abordagem realiza um mapa tecnológico, adotando como referência o ciclo de vida de um dado dentro desse tipo de infraestrutura. São esses estágios, a geração de dados, aquisição, armazenamento e processamento.

Dentro desse contexto, inicialmente, na seção 2, será definido formalmente o conceito de big data, discutindo-se os principais pontos que envolvem o tema, dando-se um panorama geral sobre o assunto. Em seguida, na seção 3 será discutido o processo de análise e extração de informação, incluindo principais metodologias e tecnologias. Na seção 4 serão discutidas as principais técnicas de armazenamento e gerenciamento de big data, como bases de dado NoSQL, sistemas operacionais distribuídos, e modelos de programação. Todos os assuntos discutidos nessas três últimas seções são agnósticos a aplicações. Essencialmente se aplicam a qualquer tipo de sistema de big data. Na seção 5 será inicialmente introduzido o conceito de Cidades inteligentes, sua definição formal, motivações e arquitetura. Em seguida será discutido o processo de aquisição de dados dentro de uma cidade inteligente. Como estudo de caso será apresentado o *middleware* Civitas. Na seção 6, será discutido o processo de geração de dados dentro de uma cidade inteligente, fechando a cadeia valor do big data. Por fim, na seção

7, todo o ciclo descrito durante esse trabalho será revisitado, no estudo de caso da implementação pratica de uma cidade inteligente, em Santander na Espanha, focando-se na arquitetura da infraestrutura de big dada da cidade, chamado CiDAP.

2 O Big Data

2.1 Definição de Big Data

O constante crescimento de dados que vem ocorrendo ao longo das ultimas décadas, e que se intensificou de maneira muito forte nos últimos anos fez com que o termo big data fosse cunhado, com o intuito de abrigar uma gama de diferentes conceitos em apenas expressão. Apesar dele ter se tornado recorrente, o seu significado literal traz, de fato, muita pouca informação a respeito de seu verdadeiro significado. big data não significa apenas “quantidade de dados muito grande”, ou “quantidade massiva de dados”. No entanto, não há apenas uma única definição monolítica do que significa big data. Sua definição dependerá muito de sua aplicação, e de seu contexto.

Aqui serão apresentadas as 3 principais definições, de acordo com [1], sendo a primeira delas a mais subjetiva, no entanto a de mais simples entendimento, e mais recorrentemente citada. A segunda é comumente chamada de “definição 4V”, e a terceira trata-se do padrão definido pelo NIST (*The National Institute of Standards and Technology*) [2].

1. **Definição comparativa:** “*conjunto de dados cujo tamanho vai além da capacidade de ferramentas de base de dados convencionais para realizar operações de captura, armazenamento, gerenciamento e análise*”. Essa é uma definição subjetiva, pois não envolve métrica objetiva alguma. No entanto, captura um conceito evolucionário, de forma que em dado período de tempo é possível classificar o que é, e o que não é big Data. Foi criada em 2011, em um relatório da Mckinsey [1].
2. **Definição atributiva:** “*A tecnologia de big data descreve uma nova geração de tecnologias e arquiteturas, projetadas para economicamente extrair valor de grandes volumes de uma grande variedade de dados, de forma que estas possibilitem a captura, descoberta e/ou análise de dados em alta velocidade*”. Como essa definição salienta quatro importantes características do big data, sendo elas volume, variedade, valor e velocidade, ela é muitas vezes referenciada como definição “4V”. É muitas vezes também referenciada como definição “3V” [3], referenciando apenas as medidas de magnitude (volume, variedade e velocidade).
3. **Definição arquitetural:** “*big data é o ambiente onde os volumes de dados, velocidade de aquisição, ou a representação dos dados limitam a habilidade de se realizar análises efetivas ao se utilizar abordagens relacionais tradicionais, ou que requerem um uso significante de escalabilidade horizontal para um processamento eficiente*”.

Como é possível notar, as 3 definições citadas, apesar de diferentes, não são conflitantes. A primeira, trata big data como um conjunto de dados que se diferenciam de dados tradicionais. A segunda, trata big data como sendo um conjunto de tecnologias que visa atingir certos objetivos, que tecnologias tradicionais não alcançariam. A terceira, por sua vez, da destaque as características de arquitetura necessárias para processamento eficiente de quantidades massivas de dados. Por esse motivo, esse trabalho irá utilizar todas as definições, que irão variar de acordo com o contexto.

Para ilustrar as diferenças entre big data, e dados tradicionais, ressaltar alguns pontos importantes

- Primeiramente o volume de dados. Este é um ponto chave. Enquanto base de dados tradicionais são adequadas para se trabalhar com volumes de dados da ordem de gigabytes, o volume de big data se encontra entre a ordem de terabytes e petabytes. Além disso, bases dados tradicionais tendem a possuir possuem uma frequência baixa de atualização desses dados (por hora, pro dia), enquanto em big data, ela costuma ser muito mais frequente (por exemplo, a cada segundo).
- Em segundo lugar, dados tradicionais são armazenados de maneira estruturada. big data, por possuir grande uma variedade de dados, possui dados organizados tanto de forma estruturada (bases relacionais), semi-estruturada (páginas web, html e logs), e não estruturada (vídeos, imagens, texto, documentos e arquivos binários), sendo dados não estruturados representantes da vasta maioria de fontes de big datas comuns [1], como por exemplo do Facebook Twitter, ou Youtube. Além disso, no big data os dados são pouco integrados, e armazenados de maneira distribuída. Dados tradicionais são armazenados de maneira centralizada.
- Terceiro, a velocidade de análise e processamento de big data, deve coincidir com a velocidade de produção de dado. Por exemplo, imaginando um sistema de RFID, ou de detecção de fraude, os dados devem sair dos emissores, entrarem dentro do ambiente de processamento, e serem processados o mais rápido possível (preferencialmente em tempo real) de maneira a agregarem o máximo de valor possível a informação. Para esse tipo de aplicação, um tempo de processamento demasiadamente grande, poderia tornar a correta operação do serviço inviável, ou sua utilidade nula.
- Por fim, na mesma proporção que a quantidade de dados pode ser muito grande, a densidade de valor do dados em geral é baixa. Para se resolver esse tipo de problema, técnicas de mineração são utilizadas para extração de informações úteis de grandes quantidades de dados são necessárias.

2.2 Seu Histórico e Big Data Hoje

Antes de iniciar a discussão de detalhes a respeito do big data, será dada um breve histórico. A figura 1 mostra de forma ilustrativa a evolução da quantidade de dados armazenados demandadas ao longo dos anos.

Com relação a quantidade de volume armazenado, pode-se dividir o big data nas seguintes fases:

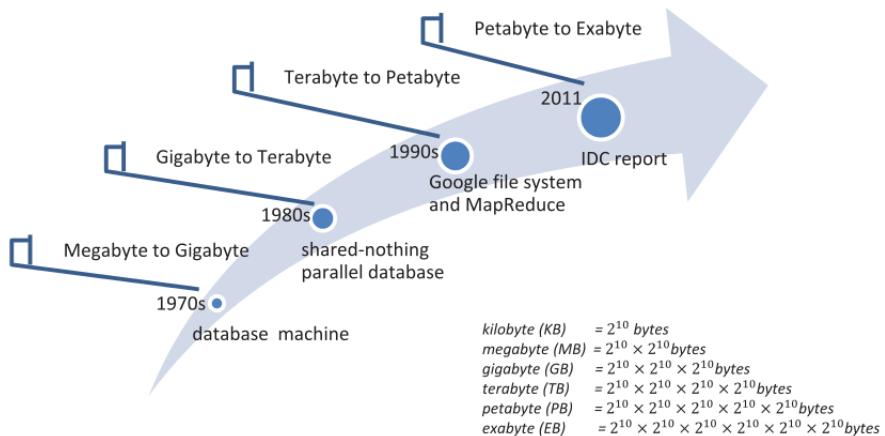


Figura 1. Evolução dos volumes de dados armazenados ao longo dos dados. Ao longo de cerca de 40 anos, o tamanho das bases de dados demandadas passaram de megabytes, para peta e exabytes. Esta figura captura o conceito evolutivo de big data, proposto em sua primeira definição. Fonte da imagem: [1]

- **Megabyte para Gigabyte:** Essa fase se deu entre os anos 1970 e 1980, quando as primeiras bases de dados comerciais começaram a surgir. De início as pesquisas focavam em desenvolver sistemas de hardware e software integrados, a fim de fornecer um melhor desempenho. Porém, após um certo período de tempo, com a evolução dos computadores de propósito geral, tais sistemas deram lugar a soluções puramente baseadas apenas em software, sendo executadas sobre hardware “de prateleira”.
- **Gigabyte para Terabyte:** Ao final dos anos 1980, a popularização da tecnologia digital fez com que o tamanho das bases de dados aumentasse consideravelmente. Para aumentar a capacidade de processamento, foi proposta a paralelização do processamento dos dados, e modelos relacionais foram bastante explorados.
- **Terabyte para Petabyte:** Ao final dos anos 1990, com a popularização da Internet, novamente houve um salto no volume de dados armazenados. Google criou então o Google File System, e o modelo de programação MapReduce, que permitia a paralelização e computação distribuída de aplicações em larga escala, de maneira automática. Um sistema rodando GFS e MapReduce é capaz de escalar conforme a demanda, não tendo limitação da quantidade de dados que é capaz de processar. Houve também o desenvolvimento de base de dados NoSQL, que são livres de esquemas, rápidas, altamente escaláveis, e confiáveis.
- **Petabyte para Exabyte:** Durante essa fase surge o modelo moderno de big data. Com a explosão de novos serviços *online* oferecidos, e causou uma subsequente explosão da demanda, a quantidade de dados semi-estruturados,

e principalmente não-estruturados aumentou muito. Também aumenta-se o interesse por extração de valor de grandes quantias de dados.

Da mesma forma, é possível classificar historicamente o big data com relação as tendências de geração de dados, ou seja, com relação às principais fontes de dados gerados em cada nova época de expansão da armazenagem de dados. E possível classificar os padrões de geração de dados nos seguintes estágios:

- **Estágio 1:** O primeiro estagio começou no início dos anos 1990, quando a tecnologias digitais e bases de dados já eram amplamente utilizadas, principalmente por estabelecimentos comerciais (bancos, redes de lojas) e pelo governo. os dados eram em sua grande maioria estruturados.
- **Estágio 2:** Este estágio se iniciou ao final dos anos 1990, e início dos anos 2000. Primeiramente com a popularização das ferramentas de busca, como Google, e posteriormente com o possibilidade de criação de conteúdo pelos próprios usuários, como em blogs, fóruns e redes sociais. Os dados já era em sua maioria semi-estruturados e não-estruturados.
- **Estágio 3:** A terceira fase teve inicio com a emergência dos dispositivos móveis, como *smart phones*, *tablets*, e redes de sensores, o que introduziu os conceitos de geração de dados centrados em pessoas, *location awareness*, e relevantes ao contexto.

Podemos classificar os estágios de geração de dados anteriores, como sendo o primeiro o de geração passiva de dados, o segundo de geração ativa de dados, e o terceiro o de geração automática de dados.

2.3 Fontes de Big Data

Todas as três formas de geração de dados (passiva, ativa e automática) continuam sendo, todas, as principais fontes de big data atualmente. Os dados também podem ser classificados de acordo com a categoria dos mesmos. As três principais atualmente são:

- **Dados de Negócios:** Dês do primeiro estágio do big data, o uso de tecnologia da informação no setor de negócios vem possibilitando, cada vez mais as margens de lucro das empresas. Estima-se que o volume de dados de negócios produzidos por companhias ao redor do mundo vem dobrando a cada 1,2 anos [1]. Incluindo-se transações de negócios para negócios e negócios para consumidores, é estimado que as transações pela internet movimentam em torno de 450 bilhões de dólares por dia [1]. Calcula-se que o Walmart lida com cerca de 1 milhão de transações de consumidores todas as horas. No intuito de acompanhar essa demanda, é necessário que análises de dados em tempo real sejam efetivas. Esse tipo de big data corresponde a dados da ordem de até Petabytes, e em sua maioria, os dados são armazenados de forma estruturada e semi-estruturada.

- **Dados de Rede:** Esse tipo de dado é relacionado tanto a criação proativa de dados por parte dos usuários, e geração de dados automáticos (englobando portanto os estágios 2 e 3). Como exemplo, pode-se citar o mecanismo de busca da Google. Estima-se que ainda em 2008 ele processava cerca de 20 PB de dados todos os dias [1]. Com relação a redes móveis, em 2010 cerca de 4 bilhões de pessoas pelo mundo estavam usando móveis, dentre os quais 12% destes eram *smart phones*. Com relação ao campo da Internet of Things, atualmente mais de 30 milhões de sensores de rede estão funcionando em meios de transporte público, automóveis, indústrias e no setor de varejo. Estes sensores estão crescendo a uma taxa de cerca de 30% ao ano [1]. Este tipos de dados, são em geral semi-estruturados, ou em grande maioria não-estruturados. Porem, ao mesmo tempo requerem um rápido tempo de resposta, alem de terem que suportar um alto número de usuários concorrentes.
- **Dados Científicos:** Paralelamente as categorias acima citadas, há também um crescente aumento no tamanho das bases de dados científicas sendo analisadas atualmente, dado o aumento do poder computacional, e novas tecnologias. Podem ser citadas como áreas de maior destaque, a biologia computacional astronomia e física de alta energia. estima-se que o laboratório CERN gerava em 2008 uma quantidade de dados igual a 2PB por segundo, tendo gerado ao final do ano cerca de 10 PB de dados processados.

Até o momento foram discutidas as definições de big data, seu histórico, e suas principais fontes na atualidade. No restante desta seção serão discutidos aspectos da arquitetura do big data.

Inicialmente será introduzido o conceito de cadeia de valor, que essencialmente descreve o ciclo de vida, e constrói um mapa tecnológico do big data, a partir do qual esse trabalho foi estruturado.

Em seguida será discutido brevemente o modelo de arquitetura em camadas.

2.4 Computação com o Big Data: Cadeia de Valor e Mapa Tecnológico

Uma forma comum de descrição de um sistema big data, bastante adotada na industria, é através do conceito de cadeia de valor, ou *value chain*. Essa abordagem essencialmente “conta” o ciclo de vida de um dado em um sistema de big data. Ela é dividido em quatro diferentes etapas: Geração de dados, aquisição de dados, armazenamento de dados e análise de dados. Cada uma dessas etapas é tratadas a seguir. A figura 2 ilustra o mapa tecnológico do big data, baseando-se na cadeia de valor.

Cada uma dessas etapas é discutida a seguir:

- **Geração de dados:** Esta etapa diz respeito em como os dados são gerados. Tratando-se de big data, é esperado que os conjuntos de dados aqui tratados sejam diversos, complexos, pouco estruturados. Podemos citar como exemplos, sensores, vídeos, setor de segurança do governo, dados científicos, monitoramento de ambientes, comercio *online*, etc.

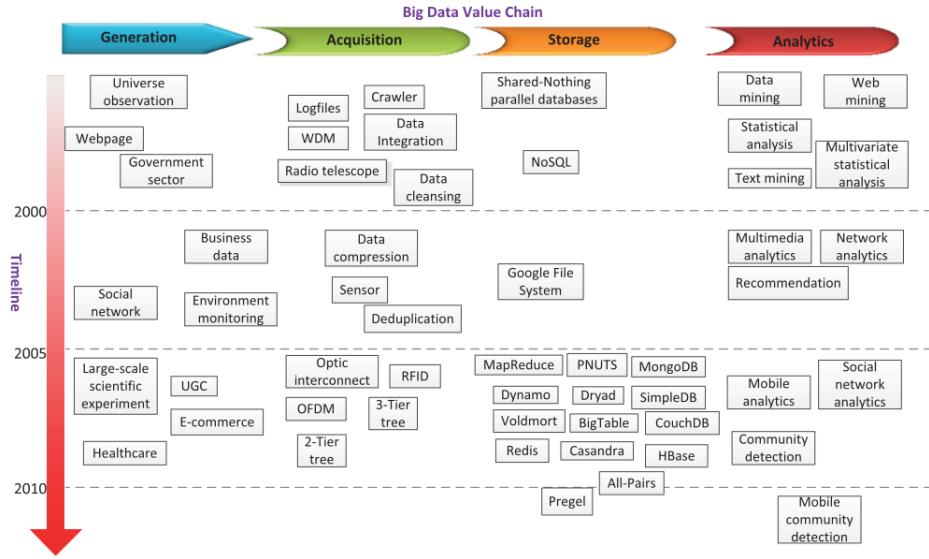


Figura 2. Mapa tecnológico do big data, conhecido como cadeia de valor(*value chain*). Isso porque essa arquitetura é capaz de classificar as tecnologias de big data, por meio do ciclo de vida de um dado. Fonte da imagem: [1].

- **Aquisição de dados:** Se refere ao processo de obtenção de informação. Esta etapa é dividida em coleta de dados, transmissão de dados, e pré-processamento de dados. Em um cenário de big data, os dados podem vir das mais variadas fontes, com os mais variados formatos , dentre eles: textos, arquivos binários, vídeos, texto formatado, páginas web, arquivos de log, vídeos, bases de dados estruturadas. O processo de coleta de dados se refere à tecnologia dedicada de coleta de dados, especializada em adquiri-los de um ambiente específico. Após a coleta, é necessário haver algum mecanismo de rápido transporte de dados, para que este seja rapidamente armazenado, e utilizado pelas aplicações de análise. Por fim, os dados capturados podem conter grande quantidade de dados sem valor ou redundantes. Para não haver um desperdício desnecessário de recursos, se faz necessário uma etapa de pré-processamento. Uma alternativa, por exemplo, é a compressão de dados.
- **Armazenamento de dados:** Esta etapa diz respeito a armazenagem e gerenciamento de conjuntos de dados de larga escala. Um sistema de armazenagem de dados pode ser dividido em duas partes: infraestrutura de hardware, e infraestrutura de software. A infraestrutura de hardware deve ser capaz de organizar-se de maneira elástica, escalando-se de acordo com a demanda. Por outro lado, a infraestrutura de software é responsável por manter os conjuntos de dados de larga escala.
- **Análise de dados:** fornece métodos analíticos para inspecionar, transformar, e modelar dados a fim de extrair deles valor. Apesar da grande variedade de dados diferentes que cada cenário pode possuir, e com diferentes tipos de

requisitos, técnicas emergentes de análise podem ser classificados em seis áreas específicas: análise de dados estruturados, análise de texto, análise de multimedia, análise web, análise de rede e análise de mobile.

2.5 Arquitetura em Camadas do Big Data

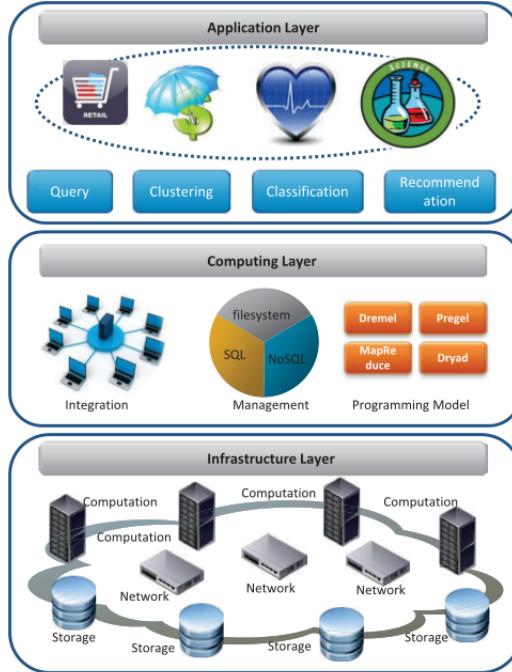


Figura 3. Arquitetura do big data, em camadas. Essa arquitetura da ênfase na disposição dos recursos computacionais. Fonte da imagem: [1]

De maneira alternativa, sistemas de big data podem ser decompostos em uma estrutura de camadas. Neste caso a divisão é feita em 3 diferentes níveis: camada de infraestrutura, camada de computação e camada de aplicação. Um diagrama da divisão da arquitetura em camadas, pode ser visto na figura 3. Este tipo de divisão diz traz um enfoque principal na disposição dos recursos computacionais, não focando tanto nas tecnologias.

- **Camada de Infraestrutura:** Esta camada é representada pela infraestrutura de *cloud*, composta de recursos computacionais e de comunicação físicos, sendo essa estrutura habilitada por meio de tecnologias de virtualização.
- **Camada de Computação:** Funciona sobre a camada de infraestrutura, e encapsula diversas ferramentas em uma camada de interoperabilidade, ou

middleware. Essas ferramentas oferecem serviços de integração de dados, gerenciamento de dados e modelos de programação. Um ferramenta de integração deve ser responsável por coletar dados dispersos de forma distribuída, e integrá-los de maneira unificada para o processamento. O gerenciamento de dados significa a capacidade de prover dados armazenados de maneira persistente e eficiente. Tal serviço é oferecido por sistemas operacionais distribuídos e bases de dados SQL e NoSQL. Já os modelo de programação oferecem abstrações lógicas que facilitam a análise de dados e desenvolvimento de aplicações. Elas são fornecidas por modelos como MapReduce, Dryad, Premel, dentre outros.

- **Camada de Aplicação:** utiliza a interface fornecida pelos modelos de programação da camada de computação, para a implementação de diversos tipos de funções de análise de dados. Dentre esses tipos de funções, inclui-se consulta, análise estatística, *clustering* e classificação.

3 Análise do Big Data: BDS (*Big Data Analytics*)

Análise de big data é o processo através do qual algoritmos rodando nas plataformas de suporte, tem por objetivo descobrir dados ocultos, tais como padrões escondidos, ou correlações desconhecidas [1].

Nesta seção ainda serão discutidos métodos de análise e extração de informação com valor de big data. Serão discutidos os paradigmas de análise de big data, classes de propósitos, categorias, principais métodos, e os tipos de análise de dados que podem ser realizados.

No contexto de cidades inteligentes, este campo ainda pode ser tratado de maneira agnóstica a tecnologia. Nesse sentido, uma vez o dado estando armazenado em um *datacenter* da cidade inteligente, a informação seria requisitada, por uma aplicação externa, e esta devolvida, após realizações de operações comuns a outros contextos de big data.

3.1 Paradigmas de Análise de Big Data

Primeiramente serão definidos aqui os dois principais paradigmas para análise de big data. São eles:

- **Processamento em *streaming*(rajada):** Esse tipo de paradigma de análise é necessário quando o valor e a utilidade do dado depende de sua coleta e produção ser completada em um curto espaço de tempo. Nessa situação, os dados devem ser processados tão rápido quanto possível, para a derivação dos resultados, havendo apenas uma pequena parcela dos dados capturados efetivamente sendo armazenados. Como exemplos de sistemas *open source* para processamento *streaming*, temos o Kafka, Storm, e o S4. [1]
- **Processamento em *batch*(fornada):** Nesse tipo de processamento, os dados são primeiramente armazenados, e só então processados. Nesse paradigma, o modelo MapReduce se tornou o grande dominante [1].

Dentro do contexto de cidades inteligentes, uma situação em que um sistema utilizando o paradigma *streaming* seria necessário, poderia ocorrer, por exemplo, em uma rede de sensores que coleta dados sobre o trânsito, e produz soluções em tempo real, indicando por exemplo, qual rota está com um menor tráfego de carros naquele momento, ou modificando a período dos sinais verde e vermelho para um cruzamento.

Uma situação análoga, que estaria mais enquadrada no paradigma *batch*, ocorreria, por exemplo em uma aplicação que calcula as melhores rotas para um determinado destino. Essa informação já está previamente disponível antes da requisição. Bastaria apenas coletar blocos de dados pré-processados para cada região, e então com a junção de todos, uma boa rota poderia ser calculada.

3.2 Propósito e Categorias

A análise de dados pode ter diferentes tipos de propósitos, podendo-se citar por exemplo a extração e interpretação de um dado, checagem de legitimidade, tomada de decisão, diagnóstico de falha, ou realização de previsões. Mas é possível se categorizar a análise de dados em três campos principais:

- **Análise Prescritiva:** visa utilizar dados antigos, a fim de descrever o que ocorreu. Está associada com inteligência de negócios, e visibilidade de sistemas.
- **Análise Preditiva:** tem por objetivo fazer prognósticos futuros, utilizando técnicas estatísticas, como regressão linear e mineração de dados.
- **Análise Prescritiva:** tem o foco na tomada de decisão, por exemplo, gerando soluções ótimas.

3.3 Métodos Comuns

Apesar do propósito poder variar muito, de acordo com o tipo de aplicação, há três tipos principais de métodos para a análise de dados em big data:

- **Visualização de dados:** A ideia é extrair dados e dispor-los na forma de gráficos, para que as pessoas possam interpretar os dados.
- **Análise estatística:** Tem por objetivo utilizar teorias e métodos estatísticos sobre os dados, e dessa forma extrair informação, visando tanto descrição ou inferência (via estatística descritiva e inferencial).
- **Mineração de dados:** é o processo computacional dedicado a encontrar padrões de dados em grandes quantias de dados. Como exemplos de algoritmos temos o C4.5, k-means e SVM.

3.4 Analíticas de Big Data, pela Natureza do Dado

Apesar da imensa variedade de dados possível em big data, é possível classificar os tipos de analíticas de dados 6 tipos principais, de acordo com a natureza de grupo de dados. São eles: análise de dados estruturados, texto, web, multimídia, rede e mobile.

- **Análise de dados estruturados:** Este é o campo em que historicamente há o maior número de pesquisas e trabalhos. Como exemplo de metodologias amplamente aplicadas em análise de grande quantidade de dados estruturados, temos técnicas de *machine learning*, que visa extrair padrões estatísticos, e mineração de dados, que procura encontrar padrões. Como exemplo de técnica de *machine learning* que vem se destacando muito nos últimos anos, pode ser citada o *deep learning*. [1] [4]
- **Análise de texto:** Esse tipo de informação costuma ser armazenada na forma de emails, documentos corporativos, páginas de web, e mídia social. E defendida a ideia de que análise de textos tem um potencial comercial superior a análise de dados estruturados. As metodologias de mineração aplicadas nesse contexto, se baseiam em técnicas de processamento de linguagem natural, ou técnicas NLP (*natural language processing*). Técnicas de *Opinion Mining* visam extrair informação a respeito do público em geral a respeito de preferência de produtos, *marketing*, movimentos políticos, e eventos. [1]
- **Análise web:** Métodos de análise de web são desenvolvidos apoiando-se em outras metodologias de extração de informação, dentre elas as de bases de dados, NLP, hiperlinks, mineração de dados semi-estruturados (páginas de html) e de dados não-estruturados (áudio, vídeo, imagens e texto). Como áreas importantes de aplicação, podem ser citados o comércio eletrônico, privacidade e segurança. [1]
- **Análise de multimídia:** Entende-se aqui por multimídia áudio, vídeo e imagens. Tais técnicas referem a extrair conhecimento e entendimento de semântica por traz desse tipo de conteúdo. Essas analíticas se referem, principalmente a tarefas de sumarização, anotação, indexação e recuperação e recomendação. Sumarização se refere ao ato de extração de textos (por exemplo extração de frases ditas em um vídeo). Anotações se referem a adição de rótulos que descrevam o conteúdo presente na multimídia. Indexação e recuperação diz respeito a descrição, armazenamento e organização de conteúdo, a fim de facilitar pessoas localizarem o conteúdo. Recomendação diz respeito a sugestão de algum conteúdo a usuários, baseando-se nas preferências individuais. [1] [4]
- **Análise de rede:** diz respeito a análise de redes sociais, através da estrutura de ligação entre dados por meio de grafos, podendo focar, por exemplo na predição de ligações, detecção de comunidades, evolução de redes sociais, etc. [1]
- **Análise de mobile:** Refere-se a análise de dados gerados por dispositivos móveis em geral. Tal agrupamento é resultado de características peculiares, comuns a todos, como por exemplo riqueza de redundância, ruído na informação, sensibilidade de atividade, e consequência de mobilidade. [1]

4 Armazenamento e Gerenciamento de Big Data

Na última seção foram estudados métodos e técnica de análise de big data. Nesta seção, continuado o caminho inverso da cadeia de valor do big data em direção às

cidades inteligentes, será discutida a etapa de armazenamento e gerenciamento de dados em big data.

Esta discussão será dividida em 3 blocos principais: inicialmente serão discutidas metodologias de armazenamento físico de dados. Em seguida, serão discutidos os modelos e *frameworks* de gerenciamento de dados. Por fim, será discutido de maneira breve a plataforma Apache Hadoop, um *frameworks open source* que consolida diversas funcionalidades de processamento de big data, e vem se tornando uma das soluções mais difundidas e adotadas na atualidade.

4.1 Infraestrutura de armazenamento

Dentre as infraestruturas físicas de armazenamento para big data, três se destacam: DAS (*Direct Attached Storage*), NAS (*Network Attached Storage*), SAN (*Storage Area Network*). Um diagrama de cada um dos casos pode ser visto na figura 4.

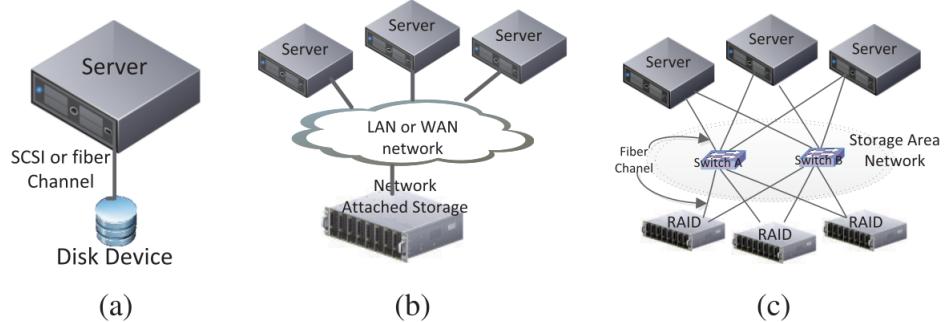


Figura 4. Representações dos diferentes tipos de infraestrutura físicas que podem ser utilizadas no armazenamento de big data: (a) DAS (*Direct Attached Storage*), (b)NAS (*Network Attached Storage*), (b)SAN (*Storage Area Network*). Essencialmente, DAS funciona como uma extensão do hardware do servidor, no NAS, o *storage* funciona como um servidor de arquivos, o SAN oferece um serviço de armazenamento em bloco, criando o efeito de que ele está sendo feito localmente. Fonte da imagem: [1].

- **DAS:** Os dispositivos de armazenamento são conectados diretamente ao servidor de processamento. Do ponto de vista do servidor de processamento, o espaço de armazenamento se comporta como uma extensão de seu hardware.
- **NAS:** É um tipo de armazenamento feito ao nível de arquivo, no qual arquivos são providos aos clientes do serviço. Nesse caso, os dispositivos de armazenamento funcionam como servidores de arquivos para os servidores de processamento.
- **SAN:** Constitui-se a uma rede de computadores e dispositivos de armazenamento que oferece uma área comum de armazenamento ao nível de bloco, criando o efeito de que o armazenamento está sendo feito localmente

4.2 Framework de Gerenciamento de dados

Os frameworks de gerenciamento de dados podem ser classificados em três camadas diferentes, como apresentado na figura 5: Sistemas de arquivos distribuídos, Base de dados e modelos de programação.

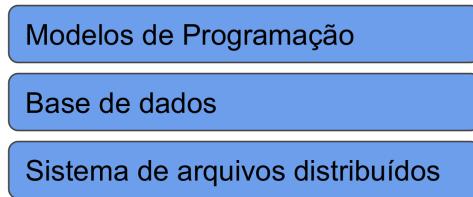


Figura 5. Arquitetura dos sistemas de armazenamento e gerenciamento.

Sistemas de arquivos distribuídos devem ser capazes de operar grandes quantidades de dados dispersos em diferentes regiões, sendo capaz de oferecer tolerância a falhas, e alta performance para um elevado numero de clientes. Após longos anos de pesquisa, se tornaram hoje uma tecnologia madura. Dentre sistemas adotados temos o GFS, HDFS, Haystack e o TFS.

Para armazenamento de big data, bases de dados tradicionais não são uma opção viável, dado o problema da variedade de dados e escalabilidade. Nesse sentido, bases de dado NoSQL vem se tornando um padrão, devido a suas diversas características positivas. Dentre elas, ser livre de esquemas, suportar replicação de forma simples, possuir uma API relativamente simples, e possuir consistência eventual¹ e suportar grandes volumes de dados. Por isso é dito que as bases de dados NoSQL seguem o modelo BASE (*Basic Available, Soft state, Eventually consistent*) [6]:

- **Basically Available:** NoSQL partitiona os dados, e o replica em diferentes locais, de maneira a fazer qualquer falha ser parcial. Ou seja, o sistema está sempre disponível, mesmo que uma parcela dos dados fiquem indisponíveis por um curto período de tempo.
- **Soft state:** NoSQL permite que os dados sejam inconsistentes, e delega o tratamento delas para os desenvolvedores de aplicações.
- **Eventually consistent:** apesar das aplicações precisarem lidar com inconsistências instantâneas, sistemas NoSQL garantem que em algum tempo no futuro os dados serão consistentes.

¹ Consistência eventual diz respeito a um resultado da ciência da computação chamado teorema CAP (*Consistency, Availability, Partition tolerance*). Ele diz que é impossível um sistema distribuído oferecer mutuamente consistência, disponibilidade e tolerância a falhas de elementos do sistema (tolerância a partição). Consistência eventual significa que a propriedade de consistência mútua entre todos os nós existentes no sistema sera atingida em algum momento no tempo. Ou seja, mesmo que não imediata, em algum momento(eventualmente), todo o sistema obterá consistência [5]

Há uma quantidade relativamente grande de bases de dados NoSQL disponíveis. Elas podem ser agrupadas de acordo com a metodologia de armazenagem dos dados, orientada a chave/valor, documento, coluna, linha, objetos e grafos.

- **Chave/valor:** Dynamo, Voldemort, Redis.
- **Documento:** CouchDB, SimpleDb, MongoDB.
- **Coluna:** BigTable, Cassandra, Hbase, Hypertable.
- **Linha:** PNUTS.
- **Objetos:** Db4o, Caché.
- **Grafos:** DEX, Neo4j, InfiniteGraph, OrientDB.

Apesar de base de dados NoSQL serem atrativas por inúmeras razões, diferentemente das bases de dados relacionais, não são capazes de suportar expressões declarativas, como operações de *join*, além de oferecerem um suporte limitado a requisições e operações de análise. Os modelos de programação, por outro lado, tem por objetivo oferecer implementações lógicas, e facilitar a análise de dados. Como modelos de programação paralela tradicionais não podem ser escalados para dimensões de big data (centenas e milhares de servidores *commodity*, espalhados por uma grande área); novos modelos de programação vem sendo propostos, com o objetivo de fornecer esse serviço, e ao mesmo tempo preencher essa fraqueza do NoSQL. Dentre os principais modelos, destacam-se:

- **Modelo de Processamento Genérico:** MapReduce, Dryad;
- **Modelo de Processamento em Grafo:** Pragel, GraphLab;
- **Modelo de Processamento em Rajada:** Storm, S4.

Dentre esses modelos de programação, o Mapreduce vem se tornando um dos modelos dominantes, em especial para aplicações do tipo *batch*. A ideia central por traz do MapReduce é quebrar o big dada em diversas porções menores de dados. Cada uma dessas porções é analisada separadamente, em paralelo, obtendo-se assim resultados parciais, através de operações localizadas de filtragem e ordenação, realizados pela função `Map()`. Os recursos computacionais destinados ao processamento desses dados são alocados próximos a origem, a fim de se evitarem *overheads*. O resultado final é obtido agregando-se todos os resultados parciais, produzindo-se um sumário da operação, por meio da função `Reduce()`. Um diagrama do modelo de processamento do MapReduce é apresentado na figura 6. [1] [7] [4] [8]

5 Aquisição de Big Data em Cidades Inteligentes

Até o momento, víhamos tratando de conceitos puramente relacionados ao big data, seguindo uma abordagem top-down, com relação a cadeia de valor de big data. Estas tecnologias discutidas até aqui devem ser oferecidas para cidades inteligentes na forma de serviços. Tanto o armazenamento quanto o processamento dos dados são feitos de maneira independente e agnóstica ao serviço que coletou os dados, e a aplicação que está realizando a requisição

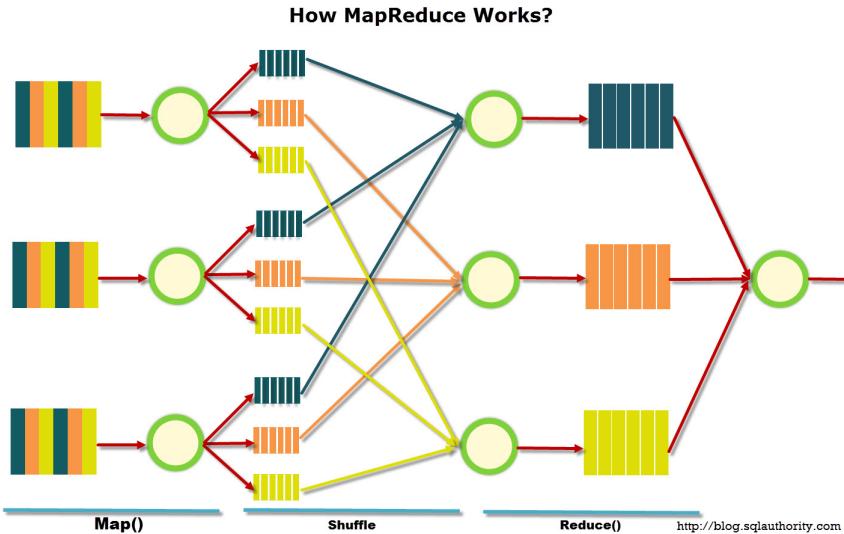


Figura 6. Diagrama que representa o modelo de processamento do MapReduce. Trata-se de uma modelo de programação e um *framework* para computação distribuída. Seu funcionamento é baseado na execução de duas funções principais: o `Map()`, que tem por objetivo “quebrar” o big data em trechos pequenos de dados, e realizar a análise desses segmentos de maneira individual a paralela, aplicando operações de filtragem e ordenação. A função `Reduce()` tem por objetivo agregar esses resultados. Fonte da imagem: [7]

dos dados processados. Nesse sentido, efetivamente não há grandes diferenças entre a implementação de um sistema de big data para uma cidade inteligente, da implementação de outro sistema big data qualquer (tendo em vista que as tecnologias devem ser direcionadas as demandas particulares de cada ambiente)

A partir de agora, iremos abordar os demais pontos de big data dentro do contexto de cidades inteligentes. Esta seção irá tratar da aquisição e coleta de dados dentro de uma cidade inteligente. Mas para isso, antes de tudo, será discutido o que é que uma cidele inteligente. Será apresentada uma definição do conceito, que será utilizada nesse trabalho, e quais as suas motivações e objetivos. Em seguida será tratado efetivamente do processo de aquisição e coleta de dados dentro de uma cidade inteligente. Nesse contexto será discutido o *middleware* Citas. Em uma cidade inteligente, o *middleware* corresponde a plataforma de interoperabilidade entre os sistemas a ela conectados, e portanto tem a função de reunir os dados coletados, e encaminha-los para o armazenamento.

5.1 Cidades Inteligentes

5.1.1 O que são cidades inteligentes

Juntamente com o aparecimento das novas tecnologias emergentes relacionadas

a telecomunicações e tecnologia da informação, novas possibilidades de serviços e acesso a recursos aparecem constantemente. Nesse sentido, é desejável que assim como em inúmeras áreas de nossa vida tais tecnologias possam ser aplicadas de maneira a gerenciar de forma mais inteligente serviços encontrados e oferecidos dentro de uma cidade. Ou seja, é desejável um sistema integrado, capaz de coletar informações das mais diversas fontes, armazenadas, e processá-las para as transformarem em informações úteis que ajudem na tomada de decisão.

Não há um conceito único unificado do que é uma cidade inteligente. Será considerada aqui uma cidade inteligente, a cidade que usufrui de um espaço digital capaz de melhorar e aumentar os serviços oferecidos, oferecendo a possibilidade de interagir com a sociedade de maneira automática, oferecendo aos cidadãos serviços inteligentes, com o objetivo de facilitar o dia a dia e melhorar a qualidade de vida do ambiente urbano. Esta será a definição adotada ao longo desse trabalho.

Como é muitas vezes dito referenciado na literatura a implementação de cidades inteligentes não é uma questão de “se”, mas sim uma questão de quando. As cidades inteligentes permitirão que se utilizem tecnologias das redes de comunicações, de sensores sem fio distribuídos, e sistemas de processamento e gerenciamento de maneira a resolver problemas atuais e futuros, bem como criar novos serviços. [9]

As tecnologias de cidades inteligentes visam integrar e analisar quantidades massivas de dados para antecipar, mitigar, e prevenir problemas. Os dados podem ser usados, por exemplo, para direcionar de forma inteligente o tráfego urbano, evitando acidentes; identificar pontos focais de criminalidade, fomentando assim a redução de crimes; e conectar cidadão durante o trabalho, ou fora dele, dentro do ambiente urbano. Cidades inteligentes também devem fornecer de maneira proativa serviços, notificações e informações para os cidadãos, como por exemplo indicar uma vaga de estacionamento, uma nova loja, ou mesmo monitorar a poluição do ar. Elas também devem conectar os cidadãos com os governos locais, de maneira a encorajar maior participação direta, interação e colaboração. É necessário também que essas soluções sejam economicamente e ambientalmente sustentáveis. [9]

5.1.2 Motivações e tendências

Inúmeras motivações e tendências vêm emergindo de maneira a criar e fomentar o conceito de Cidades inteligentes. [9] cita os seguintes:

- Competição global por talentos: com o objetivo de se manterem competitivas dentro de um cenário global, as cidades precisarão ser capazes de atrair e manter pessoas talentosas e capacitadas. Com seu ambiente de inovação, as cidades inteligentes devem desenvolver um ambiente que favoreça a inovação e a indústria. Deve ser um ambiente favorável a start-ups, e profissionais jovens. Dessa maneira elas próprias devem se beneficiar do contínuo desenvolvimento econômico.
- O aumento da população urbana tende a estressar as infraestruturas e recursos urbanos: Todas as grandes cidades sofrem do problema de congestão

onamento de trânsito, pois as vias públicas não foram construídas com a capacidade de suportar o tráfego urbano atual. Construções de novas infraestruturas podem apenas resolver parcialmente o problema. A mesma lógica se aplica a outros recursos, como escolas, hospitais, segurança pública, e fornecimento de água. Dessa forma, é necessário que as cidades utilizem as novas tecnologias com o propósito de gerenciar de forma inteligente os recursos existentes, bem como planejar o crescimento futuro.

- Mudanças climáticas fazem o uso eficiente de energia um problema urgente: É previsto que as mudanças climáticas mundiais venham a causar grandes impactos sobre a sociedade em breve. É estimado que 67% dos gases do efeito estufa são produzidos em cidades atualmente. Esse valor deve aumentar para 74% até 2030. Dessa forma, nota-se que as cidades têm um papel vital na eficiência energética, e na redução da emissão dos gases do efeito estufa.
- A forma com que as pessoas esperam lidar e ter acesso aos serviços governamentais está mudando rapidamente. Pessoas novas esperam ter a cidade em seus bolsos, seja para chamar um táxi; encontrar saber o horário do próximo ônibus; encontrar um restaurante ou evento próximo; reportar algum dano público, como uma lâmpada de rua queimada; candidatar-se a sua carteira de motorista, ou mesmo mandar um *tweet* para seu governante ou representante governamental.
- Proliferação de tecnologias e expansão da quantidade de dados gerados: a fim de se oferecer todos os tipos de serviços mencionados, uma grande e vasta infraestrutura de geração continua de dados se faz necessária. Dentre as fontes de dados, podem ser citadas câmeras de vídeo; transponder de de vias públicas; sensores, ligados a pontes, estacionamentos, ruas, canos de água, luzes de ruas, são capazes de gerar uma quantidade enorme de dados, que precisa ser tratada.

A figura 7 mostra de forma ilustrativa a união desses conceitos. Este conceito é chamado pela Cisco de *Internet of Everything*, ou IoE, que essencialmente é uma generalização do conceito de IoT (*Internet of Things*). Este conceito traduz a ideia de ubiquidade de serviços.

É estimado que atualmente menos de 1% do que poderia estar conectado, está de fato ligado a internet ou sistemas inteligentes [9]. Isso abre potencial para uma quantidade de dados gerados ainda maior. É estimado que até 2020 cerca de 212 bilhões de “coisas” estejam conectadas a internet. Além disso, estima-se também que até 2017 64% das conexões a internet sejam feitas por dispositivos moveis.

Ambos são pontos chaves e necessários para as cidades inteligentes, e se tratados, tem um grande potencial no que diz respeito a melhoria de serviços públicos, melhorar a qualidade de vida e possibilitar a descoberta de novas informações sobre as cidades, até então desconhecidas. Por exemplo, a polícia americana estimava que 80% das vezes que tiros eram ouvidos, o serviço de polícia era acionado. Com a instalação de sensores de sons em cidades, verificou-se que os números chegavam a apenas 10% em San Francisco e 22% em Oakland.

É também desejável que as cidades inteligentes se comportem de maneira automática com os cidadãos. Por exemplos, atualmente as pessoas precisam

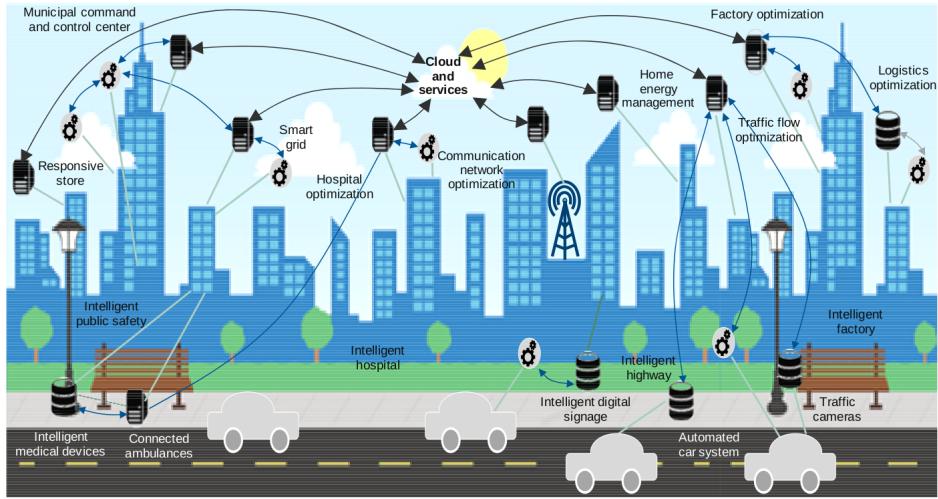


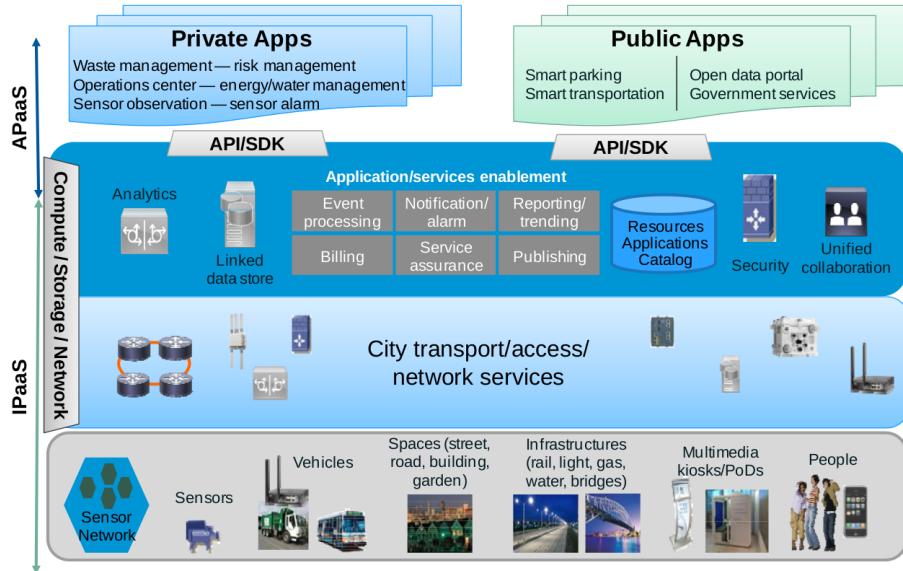
Figura 7. Figura conceitual representando uma cidade inteligente. A idéia da ubiquidade de serviços de informação pode ser resumida no conceito de IoE *Internet of Everything*: tantos serviços, quanto possível, devem estar conectados a rede, de maneira a serem criadas plataformas de inteligência remota aos mesmos. Fonte de imagem: [9].

conectar-se proativamente à internet com seus dispositivos móveis, ou então entrar em contato com serviços de emergência como ambulâncias, ou de segurança pública, no caso de alguma ocorrência. É desejável por exemplo que sensores conectados a uma pessoa idosa possam contatar o serviço de emergência de maneira automática, ou que o serviço de segurança seja acionado quando tiros forem detectados. Teríamos uma situação de muito maior eficiência, caso todos esses serviços pudessem ser atendidos de forma automatizada, com alguma inteligência por trás. [9]

Nesse sentido, serão discutidos a seguir nessa seção, propostas de arquiteturas de cidades inteligentes. Como núcleo das mesmas se faz necessária uma camada de interoperabilidade entre as diferentes fontes de dados, e os recursos computacionais disponíveis. É necessário que exista serviços unificados e integrados de aquisição de dados, para que estes possam ser corretamente armazenados, e sobre elas aplicadas técnicas analíticas de big data já discutidas, a fim de se obter valor dessa informação, propiciando os inovadores serviços de cidades inteligentes. Tal camada de interoperabilidade representada pelo *middleware* de uma cidade inteligente, no contexto de big data, correspondem a etapa de aquisição de dados.

5.2 Arquitetura de uma cidade inteligente

A figura 8, retirado de [9], mostra um diagrama conceitual da arquitetura em camadas de uma cidade inteligente. Na figura, IPaaS e APaaS significam respec-



Source: Cisco, 2013

Figura 8. Arquitetura de uma cidade inteligente. Indo do nível mais alto para o nível mais baixo, temos inicialmente a camada de aplicações e serviços, que em conjunto com as APIs e SDKs oferecidas pelo nível inferior, corresponde a Aplicação como um serviço (APaaS). Em seguida temos a camada de armazenamento e processamento de dados, a camada de aquisição e transporte e a camada de geração de dados, diretamente congruentes com a própria cadeia de valor de big data. Esses três últimos níveis formam a camada de Plataforma de integração como um serviço (IPaaS). Fonte da imagem: [9]

tivamente *Integration Platform as a Service* e *Application Platform as a Service*, ou seja Plataforma integrada como um serviço, e plataforma de aplicações como um serviço. Isso quer dizer, há portanto uma infraestrutura inteligente que permeia toda a cidade, e que permite a criação de aplicações de alto nível capazes de oferecer toda a gama de serviços possíveis, mencionados na seção anterior.

No contexto de big data deste trabalho, o terceiro nível, indo de baixo para cima (o primeiro do nível IPaaS) diz respeito tanto as etapas de armazenamento quanto processamento de dados. Este nível deve ser capaz de obter todas as informações coletadas do nível inferior, e armazena-las apropriadamente. Deve também ser capaz de aplicar as técnicas de processamento de big data discutidos, de forma a extrair informação útil dessa quantidade gigantesca de dados gerados por uma cidade inteligente. Neste contexto, essa camada deve prover APIs e SDKs a fim de facilitar o desenvolvimento de aplicações de alto nível, oferecidas pela cidade inteligente.

As camadas inferiores dizem respeito à coleta e geração de dados em uma cidade inteligente, respectivamente. O restante dessa seção será dedicada à camada

de coleta de dados para a geração de big data dentro de uma cidade inteligente, ou seja, ao nível dois da arquitetura. A proxima seção abordará a geração de dados neste ambiente, ou seja, o primeiro nível.

5.3 Coleta de dados em cidades inteligentes: *Middleware Civitas*

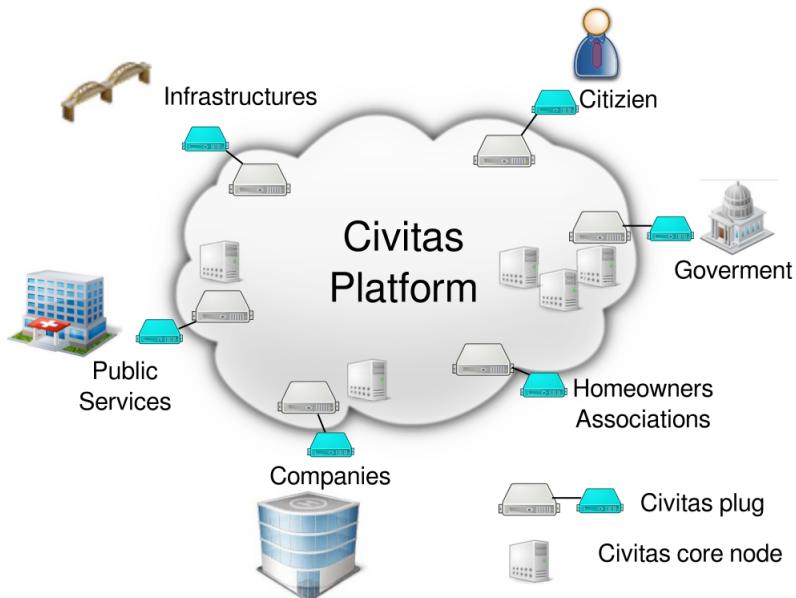


Figura 9. Representação do *middleware Civitas*. Através do *civitas plug*, diversos tipos de aparelhos diferentes, como dispositivos móveis, sensores, e servidores públicos e privados, poderiam se conectar a infraestrutura da cidade inteligente. [10]

Os dados coletados de uma cidade inteligente podem ter diversas fontes, como por exemplo sensores espalhados pelas cidades, logs de máquina, vídeos de câmeras, conteúdos de redes sociais; bem como dados públicos abertos, como mapas da cidade, horários de ônibus, localização de serviços e negócios, como lojas, e restaurantes, e muitos outros tipos de informação. [11]

Deve haver portanto um meio integrado capaz de coletar dados de todas essas diferentes fontes, transporta-los por uma infraestrutura de rede, de forma a possibilitar que toda essa informação coletada se transforme efetivamente em big data, para então ser armazenada e processada. Mas, no caso de uma cidade inteligente, como a proatividade de toda a infraestrutura também é um requisito necessário, necessário que essa camada seja capaz de resgatar dados processados, e informação ou comandos para parte dessas fontes, de maneira que os equipamentos e serviços possam responder de maneira inteligente.

Dentre as tentativas de preencher essa lacuna, temos a plataforma Civitas [10]. Um diagrama é apresentado na figura 9. Na figura o Civitas é apresentado como sendo o núcleo da infraestrutura de informação da cidade inteligente. Nesse sentido ele tem a função de orquestrar as diferentes entidades conectadas, dentre elas cidadãos, empresas, infraestruturas públicas, serviços públicos e governamentais. [10]

```
module Civitas{
    module Traffic{
        enum State {Red, Yellow, Green};
        interface Semaphore{
            void setState(State st);
        };
        ...
    }
}
```

Figura 10. Exemplo de interface de para um semáforo de transito, no *middleware* Civitas. Apesar da grande variedade de diferentes tipos de dispositivos, com diferentes sistemas operacionais e hardware, conectados à cidade inteligente, a interface do dispositivo segue o padrão C POSIX, de maneira a facilitar a integração dos dispositivos. Fonte da imagem: [10]

Para que todas essas diferentes entidades estejam mutuamente conectadas ao Civitas, todas elas devem usar um elemento chamado *Civitas plug*. Ele trata-se de é um dispositivo, devidamente certificado, que visa fornecer e consumir informações tendo como origem e destinos elementos da cidade inteligente. Ele pode ser visto como uma forma de certificar que todos os consumidores e produtores da cidade inteligente possuem as credenciais apropriadas para realizar a comunicação com a infraestrutura da mesma. O *Civitas plug* pode ser instalado em diversos aparelhos diferentes, como por exemplo *smart phones*, roteadores residenciais, servidores de empresas, sensores, bases de dados governamentais, e muitos outros. Dependendo do aparelho, a instalação do civitas pode variar dês de um simples aplicativo, até um software com uma quantidade maior de recursos adicionais. [10]

Uma coisa importante a ser salientada, é que a partir do momento que o *Civitas plug* é instalado em um aparelho, ele se torna automaticamente parte da cidade inteligente, já que do ponto de vista da plataforma da cidade inteligente, todos eles se transformam em objetos do *middleware*. O Middleware funciona de maneira orientada a objeto, o que torna possível a interoperabilidade entre diferentes tipos de objetos pertencentes a diferentes tecnologias, já que esta camada de abstração abstrai tais detalhes de cada componente, por meio de uma interface comum. Mantendo-se em mente o modelo de orientação a objeto, torna-se possível o gerenciamento de todos os dados coletados e armazenados pelo Civitas da mesma forma que isto é feito em uma *cloud* de TI. [10]

O Civitas possui também entidades internas próprias chamadas de *Civitas core nodes*. Eles trabalham da mesma maneira que servidores, sendo em geral suportados por instituições públicas e governamentais, onde uma grande variedade de serviços é desenvolvida. Esses nós são capazes de fornecer, portanto, serviços para cada um dos objetos conectados a cidade inteligente. Entidades privadas também são capazes de contribuir serviços e nós próprios. [10]

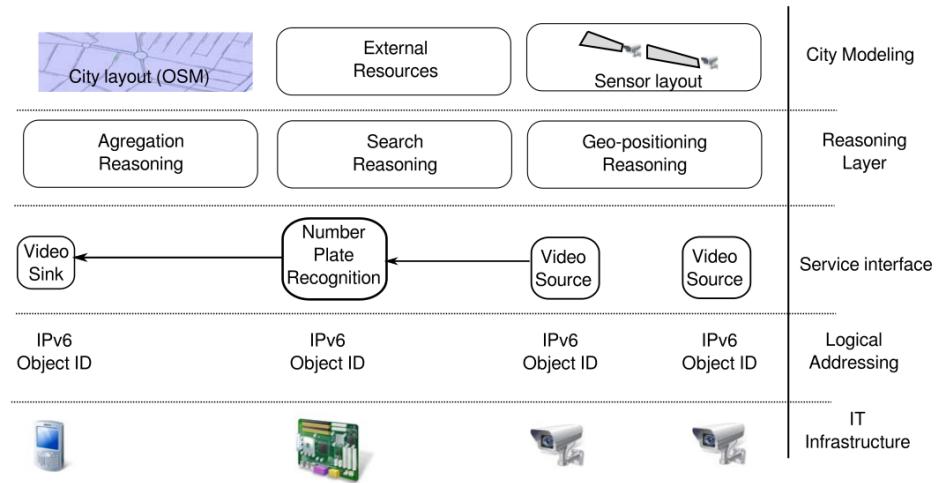


Figura 11. Arquitetura em camadas da infraestrutura do Civitas. Fonte da imagem: [10].

Resumidamente, tem-se como princípios de projeto do civitas [10]:

- Tudo é um objeto de software: dês de sensores de 8 bits, espelhados pela cidade, a poderosas FPGAs, e dispositivos mobile. Todos são vistos como objetos de software, acessados por uma interface comum, e invocados por um mesmo, e eficiente, protocolo. Do ponto de vista do desenvolvedor, cada objeto é identificado por um endereço IPv6 e um ID.
- item A interface é feita de maneira simples, de forma a permitir que os extitCivitas plug não se diferenciem entre plataformas, possibilitando a interoperabilidade.
- Plataformas de processamento ad-hoc.
- Suporte nativo de *streams* de áudio e vídeo.
- Inteligência verdadeira sobre a plataforma, possibilitando raciocínio e tomada de decisão.
- Independência do formato e traçado da cidade.
- Padronização: diversos padrões foram adotados nos diferentes níveis, de forma a sempre possibilitar a interoperabilidade.

No civitas, todas as interfaces de software seguem um modelo do padrão POSIX. Por exemplo, qualquer objeto destinado a operar um semáforo de trânsito deve implementar a interface especificada na figura 10. Um *overview* em camadas do civitas é apresentado na figura 11. [10]

6 Geração de dados em cidades inteligentes

Esta seção será dedicada a ultima etapa da cadeia de valor de big data no contexto de cidades inteligentes. Até o momento foi discutido como processar dados de big data, como armazena-los, como qual o mecanismo de coleta de big data em uma cidade inteligente, restando portanto o processo de geração de dados para esse contexto. Nesta seção serão discutidos brevemente as seguintes fontes possíveis de big data de cidades inteligentes: Dispositivos de borda e redes de sensores, dados de máquina e dados abertos

6.1 Dispositivos de borda sem fio

Aqui, dispositivos de borda sem fio, assim como definidos por [12], é entendido como qualquer dispositivo capaz de se comunicar com a rede via uma rede de acesso a rádio, como dispositivos móveis em geral, redes de sensores e câmeras.

O conceito de conceito de cidades inteligentes, é inclusivo com relação ao conceito de computação móvel de borda, definido por [12] como sendo servidores *cloud* sendo executados na borda da rede e executando serviços que seriam inviáveis para infraestruturas tradicionais.

Nesse sentido, o conceito de cidades inteligentes generaliza a computação móvel de borda, como sendo uma parte dos serviços oferecidos pela cidade inteligente. Como exemplos de casos de uso, podemos ter:

- Serviços de posicionamento geográfico em tempo real, independentes do uso de GPS, por meio do uso de algoritmos localização geográfica que usam como parâmetro o cálculo da distância entre estações radio-base sem fio locais [12] (figura 12).
- Entrega de acesso a internet para dispositivos móveis com alta vazão, e baixa latência [12].
- Serviços de analítica de vídeos de câmeras espalhadas pelas cidades. Câmeras sem espalhadas pelo ambiente urbano são capazes de capturar grandes quantidades de vídeos. Com analíticas apropriadas, se torna possível o reconhecimento de certos eventos, como acidentes, crimes, crianças perdidas, bagagens abandonadas, etc [12] (figura 13).
- Serviço de cache local para serviços de áudio e vídeo *online*, a fim de se melhorar o desempenho e minimizar a latência para usuários de dispositivos móveis [12].
- Serviços de redes de sensores, capazes de oferecer análises em tempo real das condições urbanas, como densidade de tráfego nas ruas, o que possibilitaria, por exemplo um balanceamento e redistribuição inteligentes de rotas,

utilizando para isso os semáforos urbanos [9]. Outras possibilidades são a monitoração de crimes, clima, controle de ruído, e emissão de CO₂, apenas para citar alguns exemplos. Esse tipo de fonte de dados em geral requer requisitos próprios especiais, como por exemplo, baixa utilização de recursos computacionais (pelo tamanho reduzido e baixo custo do dispositivo), e baixo consumo de energia, já que muitas vezes devido a sua localização, é inviável a utilização de fontes diretas de energia.

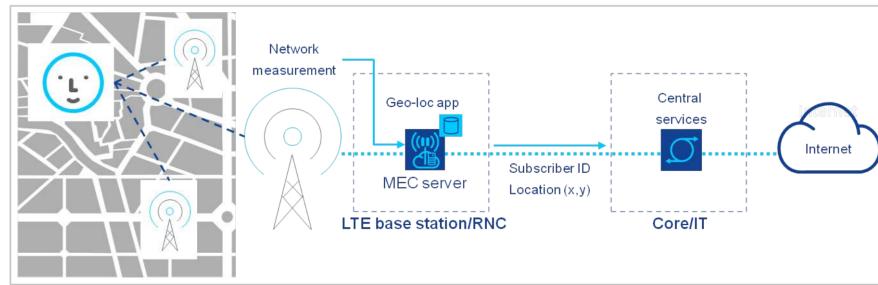


Figura 12. Exemplo de caso de uso de computação de borda de rede, que pode ser viabilizado por uma infraestrutura de cidade inteligente. A localização espacial é feita por meio da triangulação entre torres, sem a necessidade de GPS. Fonte da imagem: [12]

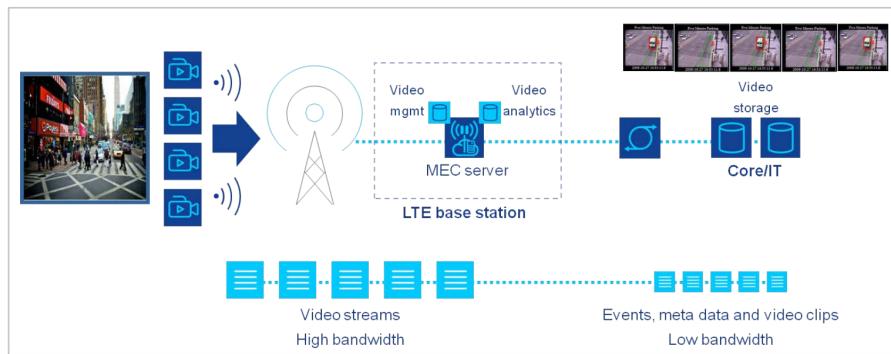


Figura 13. Exemplo de caso de uso em segurança urbana, utilizando câmeras espalhadas pela cidade. Fonte da imagem: [12].

6.2 Dados de Máquina

Estas categorias de dados incluem dados gerados de maneira automática, sem interferência humana no processo, desempenhando um papel fundamental para

um desempenho eficiente de uma cidade inteligente. Dentre as fontes de dados automáticos, além dos exemplos já citados como sensores e câmeras de vídeo, temos também logs de servidores e equipamentos de rede gerados de maneira automática [13].

Dados de logs possuem uma natureza bastante diversa, podendo representar registros de requisições HTTP [13], quanto respostas a erros de operação, falhas de hardware, problemas nas configurações de equipamentos e da rede. Esse tipo de dado, assim como os de redes de sensores, possuem natureza semi-estruturada, sendo em geral gerados em forma de par chave-valor.

Tradicionalmente, quando é verificado algum tipo de erro ou problema no ambiente de rede, a correção de tais falhas é realizada com ajuda dos logs, porém de maneira semi-automatizada, utilizando-se buscas por meio de expressões regulares [14]. Para superar esse problema, técnicas de análise automática de logs de computador vem sendo desenvolvidas [14], com uso de técnicas de agrupamento e aprendizado de máquina, a fim de se realizar a filtragem de dados relevantes.

Em um ambiente de uma cidade inteligente, dado a grande quantidade de equipamentos dispersos, e consequentemente de aumento da chance de falhas da geração de logs, a aplicação de técnicas de auto-adaptação e auto-otimização, a partir da análise desse tipo de dados fazem imprescindíveis [3] .

6.3 Dados Abertos

Por fim, serão discutidos fontes de dados abertos. Dados abertos, são assim definidos pela Open Knowledge Foundation: “dados são abertos quando qualquer pessoa pode livremente usá-los, reutilizá-los e redistribuí-los (sujeito a, no máximo, a exigência de creditar a sua autoria e compartilhar pela mesma licença).” [15]. Dentro do contexto das cidades inteligentes, esses dados podem incluir dados públicos como mapas da cidade, horários de transporte público, nomes de ruas, endereços de localizações, dentre outros. Mas podem incluir também dados privados que se estejam disponíveis para o livre acesso, como por exemplo, dados de setores de serviços, como preços e promoções, dados de redes sociais, linhas de táxi, dentre outros. A captação desse tipo de dado, e a extração de informação útil é de grande importância para uma cidade continuamente integrada, capaz de oferecer serviços inteligentes, e continuamente integrados com as necessidades dos usuários.

7 Santander, Espanha

Nesta ultima seção será realizada uma revisão rápida dos conceitos cobertos, através da análise da implantação de uma cidade inteligente real, feita em Santander, na Espanha, descrito em detalhes em [11].

Através do projeto chamado SmartSantander, cerca de 15000 sensores, conectados a cerca de 1200 nós, foram instalados por uma área de cerca de 13,4



Figura 14. Exemplo de aplicação mobile, que poderia fazer um efetivo uso de uma infraestrutura de uma cidade inteligente, melhorando a acurácia das previsões. A aplicação MTA Bus Time trata-se de um serviço oferecido em Nova York. Fonte da Imagem: [16]

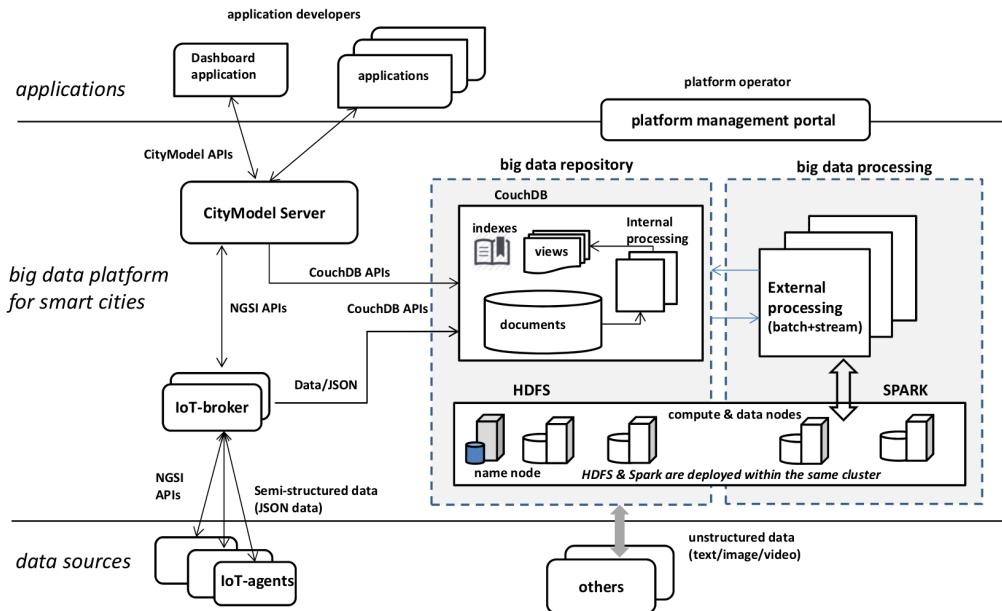


Figura 15. Infraestrutura de big data da CiDAP, em Santander, Espanha. Nela, podemos ver todas a localização lógica de cada um dos componentes da cadeia de valor do big data: o setor de análise(*big data processing*), armazenamento(*big data repository*), aquisição(*IoT-broker*, *IoT-agents*, *others*) e geração de dados(*data sources*). Fonte da imagem: [11].

milhas quadradas na cidade. Uma grande porção desses sensores estão escondidos dentro de caixas brancas instaladas na infraestrutura da cidade, como por exemplo lâmpadas das cidades prédios, e polos de serviços públicos, enquanto outros encontram-se enterrados dentro do pavimento, como por exemplo senso-

res de estacionamento. Nem todos os sensores são estáticos, muitos encontram-se instalados na rede de transporte, como por exemplo em ônibus, táxis e carros de polícia. Cada um desses sensores é responsável pela coleta de dados em tempo real, de diferentes tipos de parâmetros ambientais, como luz, temperatura, ruído, e CO_2 .

Sobre essa infraestrutura física, foi desenvolvida a plataforma de big data chamada CiDAP. Ela é capaz de acessar tanto dados históricos como os produzidos em tempo real, enquanto pode expor os resultados de várias aplicações. Um diagrama da arquitetura da plataforma CiDAP é apresentado na figura 15. A entidade responsável por realizar a comunicação com os sensores espalhados pela cidade e se dá por meio dos IoT-agents, que geram metadados no formato JSON, e que por sua vez se comunicam com o elemento chamado IoT-broker. A comunicação se da tanto em direção a infraestrutura de big data, quanto por meio de instruções recebidas do servidor CityModel. Este servidor implementa as APIs das aplicações e visa atender requisições tanto de acesso aos dados processados pela infraestrutura de big data, quanto fornecer controles sobre sensores, de maneira análoga a descrita na figura 8. O armazenamento de big data é feito em bases de dados NoSQL CouchDB, que oferece um *framework* do tipo MapReduce.

Em comparação com o *middleware* Civitas discutido anteriormente, o CiDAP não possui um componente único de entrada e saída de dados na infraestrutura de big data. Parte dessa tarefa é executada pelo IoT-broker em conjunto com os IoT-agents, sendo outras formas de dados capturadas por meio de infraestruturas de redes tradicionais.

8 Conclusão

Neste trabalho foram discutidos tópicos em big data, e suas tecnologias, dentro do conceito de cidades inteligentes. Inicialmente os principais conceitos de big data, como definição, e arquitetura foram apresentados, e discutidos. Em seguida, o conceito de cidades inteligentes e suas tecnologias correspondentes foram abordados, tendo como referência a cadeia de valor, ou *value-chain* do big data. Primeiramente foram discutidos conceitos de análise, e armazenamento e gerência de big data. Em seguida, foi definido o conceito de cidades inteligentes, foram discutidas motivações e arquitetura, introduzindo-se nesse contexto o processo de coleta e geração de big data, nesse ambiente. Por fim, foi realizado um estudo prático e real da arquitetura de uma cidade inteligente real, Santander, na Espanha, cuja infraestrutura de big data chama-se CiDAP.

Referências

1. H. Hu, Y. Wen, T.-S. Chua, and X. Li, “Toward scalable systems for big data analytics: A technology tutorial,” *Access, IEEE*, vol. 2, pp. 652–687, 2014.
2. “Nist big data,” <http://bigdatawg.nist.gov/home.php>, 2015, [Online; acessado 20/11/2015].

3. “Big data analytics - ericsson white paper,” <http://www.ericsson.com/res/docs/whitepapers/wp-big-data.pdf>, 2013, [Online; acessado 05/10/2015].
4. “Big data analytics - advanced analytics in oracle database, an oracle white paper,” <http://www.oracle.com/technetwork/database/options/advanced-analytics/bigdataanalyticswp0aa-1930891.pdf>, 2013, [Online; acessado 03/10/2015].
5. “Cap theorem,” https://en.wikipedia.org/wiki/CAP_theorem, [Online; acessado 05/10/2015].
6. “Oracle nosql database, an oracle white paper,” <http://www.oracle.com/technetwork/database/nosqldb/learnmore/nosql-database-498041.pdf>, 2011, [Online; acessado 03/10/2015].
7. “Map-reduce,” https://erlerobotics.gitbooks.io/erle-robotics-python-gitbook-free/content/caches,_message_queues,_and_map-reduce/map-reduce.html, [Online; acessado 20/11/2015].
8. “Big data for everyone hunk™: Splunk analytics for hadoop, cito white paper,” <https://www.splunk.com/content/dam/splunk2/pdfs/white-papers/big-data-for-everyone.pdf>, 2013, [Online; acessado 03/10/2015].
9. “Smart cities and the internet of everything: The foundation for delivering next-generation citizen services,” https://www.cisco.com/web/strategy/docs/scc/ioe_citizen_svcs_white_paper_idc_2013.pdf, 2013, [Online; acessado 03/10/2015].
10. F. Villanueva, M. Santofimia, D. Villa, J. Barba, and J. Lopez, “Civitas: The smart city middleware, from sensors to big data,” in *Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS), 2013 Seventh International Conference on*, July 2013, pp. 445–450.
11. B. Cheng, S. Longo, F. Cirillo, M. Bauer, and E. Kovacs, “Building a big data platform for smart cities: Experience and lessons from santander,” in *Big Data (BigData Congress), 2015 IEEE International Congress on*, June 2015, pp. 592–599.
12. “Mobile-edge computing – introductory technical white paper,” https://portal.etsi.org/Portals/0/TBpages/MEC/Docs/mobile-edge_computing_-_introductory_technical_white_paper_V1%2018-09-14.pdf, 2014, [Online; acessado 03/10/2015].
13. B. Devlin, “Unlocking machine-generated data bridging the structure chasm between hadoop and relational,” <http://assets.teradata.com/resourceCenter/downloads/WhitePapers/Unlocking%20Machine-Generated%20Data%20EB6983.pdf?processed=1>, 2013, [Online; acessado 03/10/2015].
14. “Analisar dados de máquina — a melhor maneira de avançar,” <http://www8.hp.com/h20195/v2/GetPDF.aspx/4AA5-7132PTL.pdf>, 2015, [Online; acessado 03/10/2015].
15. “What is open?” <https://okfn.org/opendata/>, [Online; acessado 20/11/2015].
16. “Mta bus time,” <http://bustime.mta.info/>, [Online; acessado 20/11/2015].