

Uma abordagem arquitetural

B-Health Analytics

José Gomes, Instituto Politécnico Cávado e Ave, Portugal, josenoversa@sapo.pt

Hélder Quintela, Instituto Politécnico Cávado e Ave, Portugal, haquintela@gmail.com

Resumo

A análise de dados de saúde por meio de tecnologias analíticas é uma das principais áreas de tecnologias de informação para a saúde (Healthcare IT), pelos impactos positivos na gestão da saúde pública, qualidade de atendimento, adequação de tratamentos e bem-estar da população. A reutilização de dados registados durante o atendimento e tratamento de pacientes é essencial para investigação clínica, suporte à tomada de decisão, gestão de recursos, monitorização epidemiológica, fármaco-vigilância.

Os dados registados no atendimento e tratamento (em muitos casos ao longo da vida) são potencialmente de grande dimensão em número de registos, de elevada complexidade e de uma grande variedade de tipos (i.e., estruturados, não estruturados, áudio, imagem, vídeo). Por estes motivos existe um interesse crescente no desenvolvimento de sistemas de Big Data aplicados à área da Saúde. O conceito de Big Data surge com o avanço significativo da produção exponencial de dados em quantidade, velocidade e variedade no tipo de dados.

O objetivo deste trabalho consiste na conceptualização da arquitetura de um sistema analítico baseado em tecnologias Big Data adaptado para a área da Saúde, e desenvolver um protótipo que permita validar os conceitos.

Palavras-chave: Business Intelligence; Healthcare Business Intelligence; Data Warehouse; Big Data; openEHR

Abstract

The analysis of health data through analytical technologies is one of the main areas of information technologies for health (Healthcare IT), by positive impacts on public health management, quality of care, adequacy of treatments and well-being of the population. Re-use of data recorded during the care and treatment of patients is essential for clinical research, decision-making support, resource management, epidemiological monitoring, pharmacovigilance.

The data recorded in the care and treatment (during lifetime) are potentially large in number of records, high complexity and a large variety of types (i.e., structured, unstructured, audio, image, video). For these reasons there is a growing interest in the development of Big Data systems applied to health care. The concept of Big Data comes up with the significant advance of the exponential production of data on quantity, velocity and variety in the type of data.

The objective of this work consists in the conceptualization of an analytical system architecture based on Big Data technologies adapted to the area of health, and develop a prototype that allows to validate the concepts.

Keywords: Business Intelligence; Healthcare Business Intelligence; Data Warehouse; Big Data; openEHR

1. INTRODUÇÃO

Nos dias de hoje existe uma maior produção de dados (tendencialmente crescente em quantidade, velocidade e variedade devido ao número de Sistemas de Informação de Suporte e à integração crescente de *IoT* –

Internet of Things). Existe um crescimento na produção de dados que são armazenados em Bases de Dados, dados esses que provêm de diferentes sistemas de informação, desde publicações em redes sociais, a compras online, a registo clínico de diagnóstico, prescrição médica de medicamentos ou meios complementares de diagnóstico, ou numa simples operação de apoio comercial e de gestão de uma empresa (Contributor, 2014). A possibilidade de correlacionamento de dados heterogêneas (ao nível de fontes e tipos) é hoje uma realidade e proporciona às organizações a capacidade de tomar decisões menos empíricas e mais sustentadas (baseadas na evidência), reduzindo o risco e aumentando potencialmente a sua sustentabilidade e rentabilidade, ao mesmo tempo que lhes coloca novos desafios na gestão e exploração de dados essenciais ao negócio (Hendler, 2014).

Os dados, e a análise de dados são fundamentais para suporte a decisões estratégicas, táticas e operacionais. No entanto a quantidade de dados que as organizações hoje necessitam tratar faz com que o manuseamento de milhares de *terabytes* de dados tenham que ser suportados por novos conceitos no âmbito dos Sistemas de Informação (Bughin, Chui, & Manyika, 2010).

Ao longo do tempo, as organizações têm implementado Sistemas de Informação principais como suporte à sua atividade, fazendo deles um apoio operacional do seu negócio, faltando, contudo, uma ferramenta de integração para uma visão analítica do conteúdo dessas operações, uma ferramenta/aplicação de suporte para a análise dos dados. Com o crescimento da produção de dados de uma forma exponencial (90% dos dados existentes foram criados nos últimos 2 anos por empresas que armazenam em média 80% dos seus dados, mas que analisam 12%) (Quintela, 2015) as entidades carecem de uma estratégia e de uma plataforma de análise de dados para obter benefícios da análise dos mesmos: o aumento da produtividade e a satisfação plena da necessidade de obter uma resposta para um problema, com maior velocidade de processamento e a integração de uma maior variedade de dados provenientes de diversas fontes.

O valor dos dados em qualquer área de negócio têm agora um papel-chave na visão estratégica no conjunto da economia e da sociedade, e esse valor é importante pela influência que terá na tomada de decisões. No entanto, a quantidade, velocidade e variabilidade de dados que as organizações hoje precisam de processar e tratar implica a existência de adoção de sistemas baseados em novos conceitos, modelos e tecnologias para que o resultado final seja útil.

Na área da saúde, a produção de dados é complexa devido ao ambiente onde são produzidos e à combinação de múltiplos sistemas de suporte de todos os seus intervenientes: sistemas de gestão de pacientes, sistemas de informação clínicos, sistemas de informação de laboratório, sistemas de informação de radiologia, etc....

Esta miríade de sistemas que necessitam funcionar de forma integrada e orquestrada nas instituições de saúde implica a adoção de um novo conceito tecnológico que passa por englobar novas definições para que seja possível extrair conhecimento de negócio essencial quer para a gestão económico-financeira quer para a gestão de pacientes: *Business Intelligence*, *Business Analytics*, *Data Mining* e *Big Data*. Mas o que é que estes conceitos podem acrescentar na análise dos dados nos cuidados de saúde?

Uma das definições de *Big Data*, é o conceito de 3 V's, velocidade, volume e variedade (Laney, 2001). No entanto vários autores têm apresentado outras definições e descrições, e uma dessas definições do conceito de *Big Data*, inclui 5 V: velocidade, volume, variedade, veracidade e valor (Anuradha, 2015). Estes conceitos de veracidade e valor são muito importantes na área da saúde, onde as decisões tomadas pelos diversos atores (administradores, médicos, enfermeiros, técnicos) têm impacto económico e social muito elevado.

De acordo com Dwayne Spradlin, (Contributor, 2014) "o poder de aceder e analisar enormes conjuntos de dados pode melhorar a nossa capacidade de antecipar e tratar doenças. Estes dados podem ajudar a reconhecer os indivíduos que estão em risco de graves problemas de saúde. A capacidade de tratar grande volume de dados num sistema de saúde também pode reduzir o custo dos cuidados de saúde em toda a linha"

Tendo presente a necessidade de existência de sistemas de *Healthcare Business Intelligence* e da complexidade dos dados a incluir e analisar justifica-se o desenvolvimento de um sistema analítico para a área da saúde baseado em *Big Data*, que enderece e resolva problemas relacionados com a integração e análise de grandes quantidades de dados provenientes de fontes heterogéneas. Aliás, neste trabalho uma das preocupações principais reside na adopção de um standard que permita a gestão, armazenamento, exploração de dados clínicos.

2. ABORDAGEM

O desenvolvimento de um modelo de arquitetura para um sistema analítico seguindo uma abordagem 2.0 (abordagem *Big Data*) irá permitir apoiar a análise de dados de saúde, tendo como base um protótipo para validar a arquitetura.

O sistema irá permitir a integração de dados de diferentes sistemas de informação (ex. °, ALERT, GLINTT, MedicineOne, etc.), de diferentes tipos de organizações de cuidados de saúde (ex. °, centros de saúde, hospitais, laboratórios, centros de tratamento, etc.) codificados em diferentes padrões (ex.º ICD – Classificação Internacional de Diagnósticos, SNOMED, LOINC), para garantir uma análise integrada e escalável.

Na área da Saúde a criação de plataformas centralizadas para consolidação de dados é uma necessidade premente, e que deve ser estratégica no atual conceito de gestão de Sistemas de Informação, de forma a potenciar a utilização dos dados registados para monitorização, controlo, planeamento, a vários níveis, Gestão da atividade contratualizada e prestada, *Benchmark*¹, vigilância epidemiológica, qualidade dos cuidados prestados, apoio à investigação médica, à investigação para otimização de processos em saúde, etc., políticas ativas de acompanhamento de doenças crónicas (e.g., Diabetes, Hipertensão, etc...), criação de

¹ *Benchmark* - Processo contínuo e sistemático que permite a comparação das performances das organizações e respetivas funções ou processos face ao que é considerado 'o melhor nível', visando não apenas a equiparação dos níveis de performance, mas também a sua ultrapassagem (DG III – Indústria da Comissão Europeia, 1996).

*Balance Score Cards*² por exemplo por tipologia de organização e tipo de cuidados prestados configurável a diferentes níveis para diferentes perfis de utilização e de acesso: visão departamental, visão organizacional, visão regional, visão nacional, etc..

A utilização de uma arquitetura e implementação de Sistemas Analíticos Centralizados ao nível das entidades reguladoras do Sistema Nacional de Saúde (que deverá recorrer atualmente a tecnologias de *Cloud Computing* e *Big Data*) deverá consolidar os dados dos prestadores de cuidados de saúde (Hospitais, Cuidados Primários, Cuidados Continuados), de forma transversal e com diferentes níveis de granularidade, disponibilizando o seu acesso (através da definição de políticas de acesso aos seus dados) a entidades tão diversas quanto: ACSS³, ARS⁴, Saúde 24⁵, ERS⁶, etc...., e favorecendo a sua disponibilização aos utilizadores de modo efetivo para a tomada de decisão baseada na evidência e em tempo útil.

A definição de uma política de gestão de dados, e a sua concretização através do desenvolvimento de plataformas de *Business Intelligence* permitirá, melhorar o acesso aos dados (acessibilidade e tempo de disponibilização), incrementar a qualidade dos registos, contribuir para a adoção e utilização crescente de standards para registo de dados, promover a interoperabilidade entre sistemas, promover a integração de dados de entidades privadas prestadoras de cuidados de saúde e que tenham contratualização com o setor público (e.g., Santas Casas da Misericórdia), o que implica a definição de normas de integração, promover colaboração, otimizar a gestão de recursos e otimizar políticas de saúde com base em informação atual e acessível de forma fácil.

3. ARQUITETURA

A. *Descrição da Arquitetura.*

Neste trabalho propõe-se o desenvolvimento de um sistema baseado em *Big Data* que permita uma integração adequada dos dados recolhidos para processamento e consolidação, independente do tipo de standards codificados.

Para responder adequadamente aos desafios é importante a utilização de modelos, ferramentas e tecnologias disponíveis em *open source* (código aberto), em particular o *openEHR*, para a modelação da arquitetura de dados, *MapReduce* para mapeamento e processamento de dados em larga escala, volume e velocidade, e de

² *Balance Score Cards* - é um sistema de planeamento e gestão estratégica, que é amplamente utilizado no mundo dos negócios e da indústria, governo e organizações sem fins lucrativos em todo o mundo para alinhar as atividades de negócios para a visão e estratégia da organização, melhorar a comunicação interna e externa e monitorar desempenho da organização em relação às metas estratégicas (Kaplan & Norton, 1996).

³ ACSS – Acrónimo de Administração Central do Sistema de Saúde, IP), que é um Instituto Público, criado em 2007, integrado na administração indireta do Estado, dotado de autonomia administrativa, financeira e patrimonial próprio, que prossegue as atribuições do Ministério da Saúde, sob a sua superintendência e tutela e têm jurisdição sobre o todo o território continental.

⁴ ARS – Administração Regional de Saúde que tem por missão garantir à população o acesso à prestação de cuidados de saúde, adequando os recursos disponíveis às necessidades e cumprir e fazer cumprir políticas e programas de saúde na sua área de intervenção.

⁵ Saúde 24- Linha de apoio à saúde criada pelo Ministério da Saúde para facilitar o acesso prestações de cuidados médicos sem ter de se deslocar a um centro de saúde ou hospital.

⁶ ERS - A Entidade Reguladora da Saúde (ERS) é uma entidade pública independente que tem por missão a regulação da atividade dos estabelecimentos prestadores de cuidados de saúde.

um sistema de base de dados baseado em estruturas *noSQL*, o *MongoDB*, para o armazenamento de um grande volume de dados para posterior exploração dos mesmos.

A horizontalidade da integração entre os prestadores de cuidados de saúde e o repositório de dados será efetuada através de uma API (Interface de Programação de Aplicações) que permitirá a receção dos dados em qualquer formato e standard, indiferente do tipo de sistema, quer seja sistema de saúde, unidades de saúde, hospitais, clínicas e laboratórios clínicos, de redes sociais, de armazenamento em *Cloud*, de meios de comunicação social ou de aplicações de dispositivos móveis. Para assegurar isto é essencial a utilização de um standard transversal para a representação de dados clínicos que na sua origem estão codificados em múltiplos standards, propondo-se a adoção do *openEHR* que permitirá uma maior escalabilidade, flexibilidade e disponibilidade na exploração de dados.

A utilização de padrões de modelação e extensões baseadas em *openEHR*, irá permitir a criação de arquétipos, com uma estrutura idêntica à apresentada na Figura 7 - Exemplo de Arquétipo, um exemplo de arquétipo sobre os dados de uma paciente clínico, onde os dados são representados estruturados e detalhados, sendo possível classificá-los por tipo e referência, o que os torna em modelos eletrónicos computáveis (Bacelar & Correia, 2015) que tornam as informações clínicas compartilháveis e uniformes necessárias ao provimento de cuidado de qualidade em saúde, e aceite como padrão Europeu desde 2007 e considerados pela ISO 18308 (Lloyd & Beale, 2006).

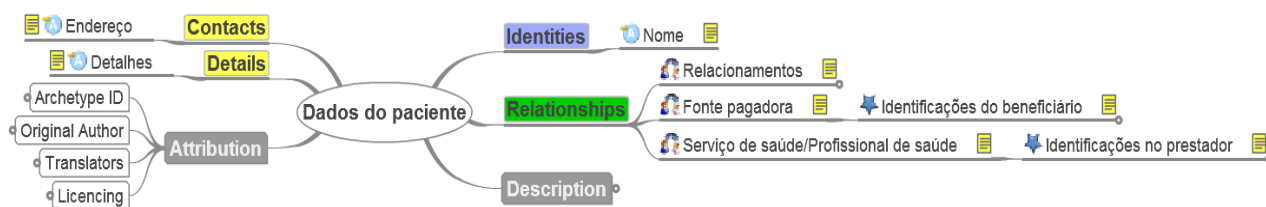


Figura 7 - Exemplo de Arquétipo

A construção desta plataforma permitirá não só a consolidação de dados armazenados utilizando o standard para estruturação dos mesmos, mas também torná-la numa ferramenta necessária para a extração de conhecimento após o mapeamento e tratamento em *MapReduce* sendo a sua disponibilização um contributo para apoio e suporte a ferramentas e sistemas *BI* e *Data Mining* já existentes.

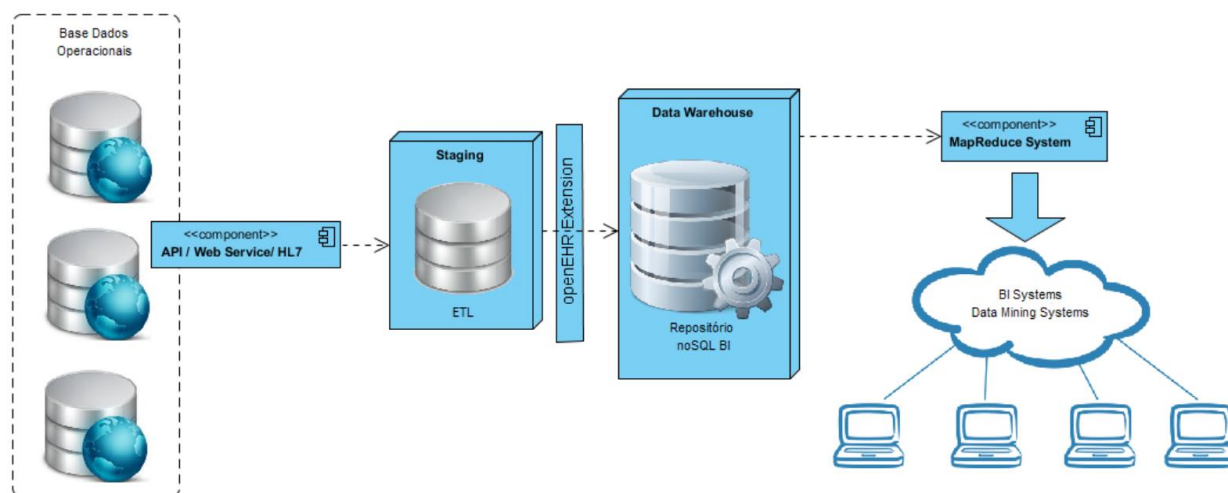


Figura 8 - Modelo conceptual

B. Ferramentas e Técnicas (Técnicas Big Data)

Em primeiro lugar a utilização do *openEHR* vai permitir definir um conjunto de regras e um modelo que garanta um padrão de armazenamento dos dados recebidos dos diferentes sistemas de informação de saúde. Esse padrão engloba referências, linguagens e arquiteturas num conceito de *open source*, disponibilizando assim uma ferramenta para os sistemas de informação na saúde com um nível semântico e funções de análise para complemento de apoio à decisão.

Em segundo lugar as bases de dados *noSQL*, nomeadamente o *MongoDB*, que se entende como um sistema de base de dados não-relacional e de certa forma distribuída, que permite o armazenamento de grande volume de dados, com maior velocidade e variedade de tipos de dados. As bases de dados *noSQL* são muitas vezes referidas como bases de dados em nuvem, bases de dados não relacionais, base de dados de *Big Data* e foram desenvolvidos para dar resposta à definição de *Big Data* e à consequência do grande volume de dados que estão a ser gerados, armazenados e analisados por utilizadores modernos (*user-generated data*) e suas aplicações (*machine-generated data*) (Planet Cassandra, n.d.).

O motivo pelo qual a escolha e a importância na utilização das bases de dados *mongoDB*, consiste no facto de terem sido já testadas na velocidade de processamento de dados, no armazenamento de grandes volumes de dados e na variedade de dados. Sendo também considerado uma vantagem o tipo de dados, que podem ser estruturados, semiestruturados ou não estruturados o que no caso deste projeto é essencial para o tipo de dados a armazenar.

Por último a escolha entre o *Apache Spark* e o *Apache Hadoop*, para o mapeamento e a consolidação dos dados armazenados.

A escolha do *Apache Spark* em detrimento do *Apache Hadoop* que tinha sido referenciado como uma primeira escolha, deve-se aos recentes dados de comparação entre as duas linguagens publicados.

Segundo os resultados da *The Apache Software Foundation*, o *Apache Spark* tem um processamento maioritariamente em memória que para este tipo de processamento de dados em *Big Data* pode ser 100 vezes mais rápido ou 10 vezes mais rápido em disco (The Apache Software Foundation, n.d.).

Tanto o *Apache Spark* como o *Apache Hadoop* utilizam o conceito de *MapReduce*, que é uma técnica apresentada como linguagem de programação para permitir o processamento distribuído de grandes volumes de dados. Como pode estar num ambiente distribuído, os dados estão divididos por várias máquinas, cada uma contendo uma parte dos dados, para o processamento ser feito em cada máquina de forma independente para obter um melhor resultado final. (Dean & Ghemawat, 2010).

O processamento computacional ocorre com os dados armazenados num sistema de ficheiros ou dentro de uma base de dados, que leva um conjunto de valores de chave de entrada e produz um conjunto de valores de chave de saída (Ekanayake & Pallickara, 2008).

4. CONCLUSÃO E TRABALHO FUTURO

A arquitetura proposta está enquadrada num trabalho de mestrado que tem por objetivo o desenvolvimento de uma plataforma analítica que permitirá responder à cada vez maior necessidade de consolidação de dados na área da saúde para a sua análise, com nas componentes de gestão económico-financeiras e clínicas.

Com a criação deste sistema pretende-se promover a interligação, comunicação, partilha e disponibilização de informação para os diferentes intervenientes na prestação de cuidados de saúde procurando-se contribuir em áreas como: atendimento, redução de custos e fundamentalmente previsão e planeamento na prestação de cuidados de saúde à população.

Após este trabalho de conceptualização de arquitetura do sistema iniciar-se-á o trabalho de prototipagem do mesmo para validação e implementação piloto. Acreditamos que os desafios que serão explorados neste trabalho associados à construção deste conceito de plataforma integradora permitirá no futuro a sua utilização em outras áreas onde a extração de conhecimento resultante da recolha e armazenamento de grandes volumes de dados é fulcral para tomadas de decisão.

5. REFERÊNCIAS

- Anuradha, J. (2015). A Brief Introduction on Big Data 5Vs Characteristics and Hadoop Technology. *Procedia Computer Science*, 48(C), 319–324. <http://doi.org/10.1016/j.procs.2015.04.188>
- Bacelar, G., & Correia, R. (2015). As bases do openEHR, (SEPTEMBER 2015), 42. <http://doi.org/10.13140/RG.2.1.3248.9687>
- Bughin, J., Chui, M., & Manyika, J. (2010). Clouds, big data, and smart assets: Ten tech-enabled business trends to watch. *McKinsey Quarterly*, 75–86. <http://doi.org/10.1109/MC.2012.358>
- Contributor, C. (2014). CastlightVoice: How Big Data Will Help Save Healthcare - Forbes.
- Dean, J., & Ghemawat, S. (2010). MapReduce: Simplified Data Processing on Large Clusters, 1–18. <http://doi.org/10.1145/1327452.1327492>
- DG III – Indústria da Comissão Europeia, I. da C. E. (1996). O que é o Benchmarking.
- Ekanayake, J., & Pallickara, S. (2008). MapReduce for Data Intensive Scientific Analysis. *Fourth IEEE International Conference on eScience*, 277–284.
- Hendler, J. (2014). Data integration for heterogenous datasets, Mary Ann Liebert. Inc, Vol 2, No 4., DOI:

10.1089/big.2014.0068

Kaplan, D. R., & Norton, D. D. (1996). Balanced Scorecard Basics.

Laney, D. (2001). META Delta. *Application Delivery Strategies*, 949(February 2001), 4.
<http://doi.org/10.1016/j.infsof.2008.09.005>

Lloyd, D., & Beale, T. (2006). OpenEHR Release 1. ISO 18308 Conformance Statement, (09), 51.

Planet Cassandra. (n.d.). NoSQL Databases Defined & Explained | Planet Cassandra.

Quintela, H. (2015). Big Data Caminho para o Sucesso no mercado global. *One World*.

The Apache Software Foundation. (n.d.). Apache Spark™. Retrieved from <http://spark.apache.org/>