UNIVERSIDADE FEDERAL DE SERGIPE
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

ANTONIO FERNANDO CRUZ SANTOS

SALUTEM: ARQUITETURA DE *BIG DATA* PARA ÁREA DA SAÚDE

SÃO CRISTÓVÃO/SE
2017

UNIVERSIDADE FEDERAL DE SERGIPE
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

ANTONIO FERNANDO CRUZ SANTOS

SALUTEM: ARQUITETURA DE *BIG DATA* PARA ÁREA DA SAÚDE

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação (PROCC) da Universidade Federal de Sergipe (UFS) como parte de requisito para obtenção do título de Mestre em Ciência da Computação.

Orientadora: Profa. Dra. Adicinéia Aparecida de Oliveira

SÃO CRISTÓVÃO/SE
2017

ANTONIO FERNANDO CRUZ SANTOS

SALUTEM: ARQUITETURA DE *BIG DATA* PARA DA SAÚDE

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação (PROCC) da Universidade Federal de Sergipe (UFS) como parte de requisito para obtenção do título de Mestre em Ciência da Computação.

BANCA EXAMINADORA

Profa. Dra. Adicinéia Aparecida de Oliveira, Presidente
Universidade Federal de Sergipe (UFS)

Prof. Dr. Douglas Dyllon Jerônimo de Macedo, Membro Interno
Universidade Federal de Santa Catarina (UFSC)

Prof. Dr. Paulo Caetano da Silva, Membro Externo
Universidade Salvador (UNIFACS)

SALUTEM: ARQUITETURA DE *BIG DATA* PARA SAÚDE

Este exemplar corresponde à dissertação de Mestrado, sendo o Exame de Defesa do Mestrando Antonio Fernando Cruz Santos para ser aprovado pela Banca Examinadora.

São Cristóvão, 08 agosto de 2017.

_____

Profa. Dra. Adicinéia Aparecida Oliveira

Orientadora

_____

Prof. Dr. Douglas Dyllon Jerônimo de Macedo

Membro Interno

_____

Prof. Dr. Paulo Caetano da Silva

Membro Externo

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA CENTRAL
UNIVERSIDADE FEDERAL DE SERGIPE

Santos, Antonio Fernando Cruz

S237s Salutem : arquitetura de big data para área da saúde / Antonio Fernando Cruz Santos ; orientador Adicinéia Aparecida de Oliveira. – São Cristóvão, 2017.

89 f.

Dissertação (mestrado em Ciências da computação)– Universidade Federal de Sergipe, 2017.

1. Programas de computador. 2. Big data. 3. Arquitetura de software. 4. Saúde – Previsão. I. Oliveira, Adicinéia Aparecida de. II. Título.

CDU 004.45:61

AGRADECIMENTOS

Primeiramente gostaria de agradecer a Deus, por mais esse engrandecimento pessoal e profissional. A minha família, meus pais, irmãos e sobrinhos, em especial minha mãe, grande incentivadora da minha vida, te amo muito. A minha noiva e futura esposa Dayse, por me dar suporte e incentivo, principalmente nas horas mais difíceis. A minha orientadora, Adicinéia, que hoje considero mais que orientadora, uma amiga, muito obrigado por tudo. Aos profissionais do Hospital Universitário de Sergipe, em especial a Profa. Dra. Angela Maria da Silva, ao Prof. Dr. Antônio Marco Prado Nunes e a Farmacêutica Flávia Oliveira da Costa, por ajudarem na realização deste trabalho. Finalmente, sou grato aos meus amigos e colegas, que até mesmo sem terem conhecimento, me ajudaram com palavras de encorajamento para seguir a diante nessa jornada.

RESUMO

Não é de hoje que os dados são considerados um dos principais ativos das organizações. Nos últimos anos, decorrente a crescente geração de dados, em diversos formatos e variadas fontes, a importância em tornar os dados em informações e em conhecimento, são características basilares de *big data*. No ambiente da saúde, o conceito *big data* está se popularizando, trazendo inúmeros benefícios para a área, como: predição de doenças, melhoria no tratamento dos pacientes, entre outros. No entanto, a aplicação de soluções de *big data* pode não ser simples, pois apresenta inúmeros desafios, como a segurança dos dados dos pacientes e profissionais de saúde, a questão da unicidade semântica empregada, entre outros. Por um lado, os dados, em contextos de saúde, são sensíveis e por esse motivo devem ser salve guardados contra acessos indevidos. Por outro, os dados que são compartilhados pelos profissionais de saúde devem manter uma mesma interpretação para que não gere interpretações dúbias. Nesse sentido, este trabalho propõe a arquitetura *Salutem,* uma arquitetura de *big data* para a área da saúde, com vistas a simplificar sua aplicação, mas também contornar os problemas da segurança e semântica dos dados. Como forma de avaliação, a arquitetura proposta foi aplicada em um estudo de caso, onde foi possível avaliar sua utilização. Os resultados obtidos foram satisfatórios, haja vista que foram entregues serviços para os consumidores com êxito.

Palavras-chave: *Big data*, arquitetura, predição, saúde.

ABSTRACT

It is not new that data is considered one of the main assets of organizations. In recent years, due to the growing data generation, in several formats and varied sources, the importance of transforming data into information and knowledge are basic characteristics of the big data. In the healthcare environment, the big data is becoming popular, bringing innumerable benefits to the area, such as: prediction of diseases, improvement in the treatment of patients, among others. However, the application of solutions may not be simple, because it presents challenges that must be taken into account, such as the safety of patients 'and health professionals' data, as well as the semantic uniqueness employed. On one hand, the data, in healthcare context, are sensitive and for this reason must be protected against improper access. On the other hand, the data that are shared by the health professionals must maintain the same interpretation, so it does not generate dubious interpretations. In this sense, this dissertation presents a big data architecture for healthcare, called Salutem, in order to simplify its application, but also to overcome the problems of data security and semantics. As a form of evaluation, the proposed architecture was applied in a case study, by which was possible to evaluate its use. The results were satisfactory, given that consumer services were delivered successfully.

Keywords: *Big data*, architecture, prediction, health.

LISTA DE FIGURAS

Figura 1: *Big data* em números. 23
Figura 2: Principais dimensões de *big data*. 24
Figura 3: Fases de análise de *big data*. 26
Figura 4: Proposta de arquitetura conceitual de *big data* em saúde. 39
Figura 5: Arquitetura extensão da NIST. 40
Figura 6: Proposta de arquitetura para monitorar pacientes com *big data* e IoT. 41
Figura 7: Arquitetura BISE. 42
Figura 8: Arquitetura H-DRIVE. 43
Figura 9: Arquitetura *Salutem*. 47
Figura 10: Camada de Aquisição. 48
Figura 11: Camada de Armazenamento. 51
Figura 12: Camada de Análise. 53
Figura 13: Processo CRISP-DM. 55
Figura 14: Camada de Serviço. 56
Figura 15: Camada de Segurança. 58
Figura 16: Ontologia. 60
Figura 17: Exemplo de atores da *Salutem*. 61
Figura 18: Taxonomia de referência da arquitetura *Salutem*. 62
Figura 19: Processo dos exames laboratoriais. 66
Figura 20: Ontologia ontoHemograma. 68
Figura 21: Ontologia ontoHemograma série branca. 68
Figura 22: Ontologia ontoHemograma série vermelha. 69
Figura 23: Extrato do arquivo CSV dos hemogramas. 71
Figura 24: Extrato do arquivo CSV da camada de armazenamento. 72
Figura 25: Página principal do Hemo *Data*. 73
Figura 26: Informações hemoglobina. 74
Figura 27: Solução de Controle de Acesso (CA). 75

LISTA DE QUADROS

Quadro 1: Definições de *big data*. 21
Quadro 2: Quantitativo de artigos selecionados na execução do protocolo. 33
Quadro 3: Análise propostas de arquitetura. 35
Quadro 4: Resumo das arquiteturas de *big data* em saúde selecionadas. 44

LISTA DE SIGLAS E ABREVIATURAS

ACID	Atomicidade, Consistência, Integridade e Durabilidade
AIDS	*Acquired Immune Deficiency Syndrome*
API	*Application Programming Interface*
BPMN	*Business Process Model and Notation*
CA	Controle de Acesso
CRISP-DM	*Cross Industry Standard Process for Data Mining*
CRM	*Customer Relationship Management*
CSV	*Comma-Separated Values*
DNA	Ácido Desoxirribonucleico
EHR	*Electronic Health Record*
ERP	*Enterprise Resource Planning*
ETL	*Extraction-Transform-Load*
GPS	*Global Positioning System*
IEC	*International Electrotechinica Commission*
HDFS	*Hadoop Distributed File System*
HTTP	*Hypertext Transfer Protocol*
HU	Hospital Universitário
IoT	*Internet of Things*
ISO	*International Organization for Standardization*
JSON	*Javascript Object Notation*
KDD	*Knowledge-Discovery in Databases*
kNN	*k-Nearest Neighbor*
LABCLIN	Laboratório de Análises Clínicas
NoSQL	*Not only* SQL
OMS	Organização Mundial da Saúde
OSI	*Open Source Iniciative*
PaaS	*Platform as a Service*
PIB	Produto Interno Bruto
REST	*REpresentational State Transfer*
SGBD	Sistema Gerenciador de Banco de Dados
SGPTI	Setor de Gestão de Processos e Tecnologia da Informação

SQL	*Structured Query Language*
SUS	Sistema Único de Saúde
TI	Tecnologia da Informação
TOGAF	*The Open Group Architecture Framework*
UFS	Universidade Federal de Sergipe
WHO	*World Health Organization*
XML	*eXtensible Markup Language*

SUMÁRIO

1 INTRODUÇÃO 15

1.1 PROBLEMÁTICA E MOTIVAÇÃO 16

1.2 HIPÓTESE 18

1.3 OBJETIVOS 18

1.3.1 OBJETIVO GERAL 18

1.3.2 OBJETIVOS ESPECÍFICOS 18

1.4 METODOLOGIA DE PESQUISA 18

1.5 ORGANIZAÇÃO DO DOCUMENTO 19

2 BIG DATA 21

2.1 DEFINIÇÃO 21

2.2 FASES DE IMPLANTAÇÃO DE *BIG DATA* 25

2.3 APLICAÇÕES DE *BIG DATA* 27

2.4 DESAFIOS DE *BIG DATA* 29

2.5 *DATA ANALYTICS* 30

2.6 CONSIDERAÇÕES FINAIS DO CAPÍTULO 31

3 ARQUITETURA PARA *BIG DATA* 32

3.1 ARQUITETURA 32

3.2 PROTOCOLO DE REVISÃO SISTEMÁTICA DE LITERATURA 32

3.3 ARQUITETURAS PARA *BIG DATA* 34

3.4 ARQUITETURAS DE *BIG DATA* PARA SAÚDE 39

3.5 ANÁLISE DAS ARQUITETURAS PARA A ÁREA DA SAÚDE 43

3.6 CONSIDERAÇÕES FINAIS DO CAPÍTULO 45

4 ARQUITETURA SALUTEM 46

4.1 ESTRUTURA DA ARQUITETURA *SALUTEM* 46

4.2 CAMADA DE AQUISIÇÃO 48

4.3 CAMADA DE ARMAZENAMENTO 50

4.4 CAMADA DE ANÁLISE 53

4.5 CAMADA DE SERVIÇO 56

4.6 CAMADA DE SEGURANÇA 57

4.7 COMPONENTE ONTOLOGIA 59

4.8 ATORES 61

4.9 TAXONOMIA DE REFERÊNCIA DA ARQUITETURA *SALUTEM* 62
4.10 CONSIDERAÇÕES FINAIS DO CAPÍTULO 63

5 ESTUDO DE CASO: HEMO *DATA* 64
5.1 O HOSPITAL UNIVERSITÁRIO E O LABORATÓRIO DE ANÁLISES CLÍNICAS 64
5.2 PROCESSO DE REALIZAÇÃO DE HEMOGRAMA 66
5.3 SERVIÇOS REQUISITADOS 67
5.4 ONTOLOGIA ONTOHEMOGRAMA 67
5.5 PROVEDORES DE DADOS 70
5.6 CAMADA DE AQUISIÇÃO 70
5.7 CAMADA DE ARMAZENAMENTO 71
5.8 CAMADA DE ANÁLISE 72
5.9 CAMADA DE SERVIÇOS 73
5.10 CAMADA DE SEGURANÇA E CONSUMIDORES 74
5.11 ANÁLISE DA APLICAÇÃO DA *SALUTEM* NO ESTUDO DE CASO 75
5.12 CONSIDERAÇÕES FINAIS DO CAPÍTULO 77

6 CONSIDERAÇÕES FINAIS E TRABALHOS FUTUROS 78
6.1 PRINCIPAIS CONTRIBUIÇÕES 79
6.2 LIMITAÇÕES DA PESQUISA E TRABALHOS FUTUROS 79

REFERÊNCIAS 81

CAPÍTULO 1

INTRODUÇÃO

Segundo a Organização Mundial da Saúde (OMS), o gasto com a saúde correspondeu a 9,9% do Produto Interno Bruto (PIB) do planeta em 2014, já Brasil, o gasto com a saúde correspondeu a 8,3% do PIB (WHO, 2017). Apesar de o volume ser expressivo, segundo Relatório Executivo da IBM Global *Business Services* (Korster e Seider 2010), o gasto com sistema de saúde é altamente ineficiente, em que são desperdiçados anualmente cerca de US$ 2,5 trilhões. Encontrar alternativas que otimizem os gastos na saúde, seja na tomada de decisão pelos gestores, seja na melhora dos cuidados ao paciente, é crítico. Uma dessas possibilidades é a análise de dados, e consequentemente, das informações e do conhecimento existentes no ecossistema da saúde.

Segundo Sondergaard (GARTNER GROUP, 2011), a informação será o óleo do século 21. Isto pode ser observado pela busca constante das organizações por mais dados e, consequente, transformações em informações que podem ser diferenciais e oportunidades de negócio. Hoje em dia, as organizações armazenam cada vez mais dados advindos de diversas fontes, que possuem uma enorme variedade e complexidade. Por consequência, levando a uma mudança na estratégia de gestão de dados, a qual foi denominada *big data* (GARTNER GROUP, 2011).

De acordo com (DAVENPORT, 2014), *big data* é um termo genérico para dados que não podem ser contidos nos repositórios usuais, há grande quantidade de dados que não cabem num único servidor, ou ainda, há dados não estruturados que não se adequam a um banco de dados organizado em linhas e colunas. Os dados armazenados, que até então eram somente estruturados (principalmente em tabelas), passam a ser também não estruturados, como por exemplo, arquivos de vídeos, postagens em redes sociais, entre outros (DAVENPORT, 2014; HANSMANN; NIEMEYER, 2014; JUNIOR *et al.*, 2015; VIEIRA *et al.*, 2012).

Segundo Groves *et al.* (2013), estima-se um aumento de US$100 bilhões de lucro anualmente por parte das empresas, se estratégias de *big data* forem adotadas. Embora a popularidade de *big data* em todos os segmentos de negócio seja recente, seu uso está se tornando um diferencial. No contexto da saúde, o volume de dados é grande, exige rapidez e

são demasiadamente complexos para que os profissionais da área processem e interpretem em tempo hábil e, assim, proporcionem valor para o negócio (ANDREU-PEREZ *et al.*, 2015; KUPWADE PATIL; SESHADRI, 2014; RAGHUPATHI; RAGHUPATHI, 2014).

Na literatura são relatados diversos problemas na aplicabilidade de *big data* em saúde, entre eles estão: privacidade e segurança dos dados, padronização dos dados, compartilhamento de dados, entre outros. Também na literatura existem diversas propostas de soluções para tais problemas, como arquiteturas, metodologias, *frameworks*, entre outras.

1.1 Problemática e Motivação

O ambiente da saúde tem gerado historicamente uma grande quantidade de dados motivados pela manutenção de registros dos pacientes, conformidade e requisitos dirigidos por regulamentações e a assistência ao paciente. Embora muitos desses dados ainda estejam disponíveis em papel, nas últimas décadas está ocorrendo um progresso na digitalização dos registros médicos, além das bases de dados eletrônicas em saúde. Impulsionada por requisitos obrigatórios e a potencial melhoria na qualidade da prestação de cuidados em saúde e reduzindo custos, a grande quantidade de dados (*big data*) procura dar suporte a uma ampla gama de serviços médicos e de saúde, incluindo o apoio à decisão, a vigilância epidemiológica, a gestão da saúde pública, entre outros (GROVES *et al.*, 2013; RAGHUPATHI; RAGHUPATHI, 2014).

Segundo Raghupathi e Raghupathi (2014), em 2001 o sistema de saúde dos EUA atingiu 150 *exabytes* em dados, sendo que ao continuar nesse ritmo de crescimento, logo atingirá os *yottabytes*. No contexto de *big data*, não se deve ater somente ao "*big*", mas também nas outras características envolvidas com o termo, como a variedade de formatos. Estas variedades de dados manipulados podem ser estruturados, semiestruturados e não estruturados, sendo estes encontrados com facilidade no âmbito da saúde, como por exemplo, as imagens de raio-x e de ultrassonografias, mensagens feitas pelos pacientes em redes sociais, que também podem ser operadas por soluções *big data* e posteriormente analisadas, entre outros (GROVES *et al.*, 2013).

Outro desafio para *big data* em saúde é a questão da segurança dos dados. Os registros eletrônicos dos pacientes, por exemplo, são suscetíveis a acesso inadequado, integridade de dados comprometida ou distribuição não autorizada. Quando tais registros são armazenados e processados em *big data* o problema ainda persiste, sendo necessário então proteger os dados

principalmente de acessos indevidos e também possibilitar níveis de acesso por usuário, nos casos de compartilhamento de dados com a comunidade de saúde (ANDREU-PEREZ *et al.*, 2015; RAO; SUMA; SUNITHA, 2015).

As ontologias permitem o compartilhamento de dados mantendo-se a semântica. As ontologias constituem uma conceituação formal de um domínio específico, que através de um vocabulário comum fornece explicitamente os significados dos dados, aumentando a capacidade de reutilização independentemente do tipo e formato dos dados. Essa característica é importante no setor da saúde, pois assim, é possível que o mesmo dado seja entendível da mesma forma, por todos que o compartilham (KIONG; PALANIAPPAN; YAHAYA, 2011; MEZGHANI *et al.*, 2015).

Outro ponto relevante ao se trabalhar com dados em saúde diz respeito à qualidade destes. De acordo com Tekieh e Raahemi (2015), os dados de saúde são coletados de diversas fontes, com configurações totalmente diferentes e projetos de banco de dados que tornam os dados complexos, sujos, incompletos, com diferentes padrões de codificação, entre outros. Para Zillner e Neururer (2016), para garantir comparabilidade de conjuntos de dados (qualidade), são necessários processos que garantam alta qualidade de dados através de documentação padronizada, bem como análise sistemática da saúde e dados de resultados da população específica de pacientes.

Também, ao se trabalhar no âmbito da saúde, existe o problema da variedade e quantidade de *stakeholders* envolvidos, ou seja, muitos atores manipulam os dados com interesses e necessidades diversas. Segundo Cano *et al.* (2017), existe uma necessidade crescente em coordenar os interesses entre diferentes partes interessadas, como: gestores, médicos, pacientes, entre outros. Isto gera conflitos que podem afetar na coleta, análise e manipulação dos dados.

Para que os potenciais benefícios de *big data* na saúde possam ser obtidos, de forma rápida e segura, este trabalho propõe uma arquitetura para *big data* em saúde, com vistas a dar o suporte para que as organizações possam implantar soluções de *big data* para ofertar serviços necessários como: auxiliem na tomada de decisão por profissionais de saúde, concedam a segurança da informação com verificação de usuário e níveis de permissões, garantam a semântica dos dados, melhorem a velocidade na obtenção de respostas, facilitem as pesquisas em saúde, entre outros.

1.2 Hipótese

Este trabalho trata da hipótese que o uso e a implantação de soluções de *big data*, podem ser facilitados com o suporte de uma arquitetura específica para o domínio da saúde.

1.3 Objetivos

Nesta seção é apresentado o objetivo geral deste trabalho bem como os objetivos específicos que contribuirão para o desenvolvimento deste.

1.3.1 Objetivo Geral

O objetivo geral deste trabalho é a proposição da arquitetura *Salutem*, que tem o intuito de facilitar o desenvolvimento, implantação e utilização de soluções de *big data* na área da saúde.

1.3.2 Objetivos Específicos

Para alcançar o objetivo geral, alguns objetivos específicos foram atendidos, a saber:

- Mapear arquiteturas de *big data* em diferentes contextos, analisando suas estruturas e aplicabilidade, de modo a identificar suas principais características e desafios;
- Propor uma arquitetura de b*ig data* para saúde, que atenda as características e desafios identificados; e,
- Avaliar a arquitetura proposta por meio de um estudo de caso.

1.4 Metodologia de Pesquisa

Pode-se destacar as seguintes modalidades de pesquisa científica utilizadas nesta proposta:

- Quanto à natureza de pesquisa: Pesquisa aplicada, a qual tem a necessidade de produzir conhecimento com a finalidade de contribuir para fins práticos (BARROS; LEHFELD, 2000).
- Quanto à abordagem do problema: Pesquisa qualitativa, pois é baseada na interpretação dos resultados e na atribuição de significados descritivos

(MIGUEL, 2007). Na pesquisa qualitativa, diferentemente da quantitativa, o pesquisador busca compreender os fenômenos observando-os, interpretando-os e descrevendo-os (MELO *et al.*, 2012).

- Quanto aos objetivos: Pesquisa exploratória, a qual visa proporcionar maior familiaridade com o problema com objetivo de torná-lo explícito ou construir hipóteses (SILVA; MENEZES, 2005).
- Quanto à natureza das fontes: Pesquisa bibliográfica, a qual utilizará livros, artigos publicados, websites, dissertações e teses relacionados ao tema deste trabalho.
- Quanto aos procedimentos técnicos: Estudo de caso, análise de modo detalhado de um único caso com objetivo de adquirir conhecimento a partir do fenômeno estudado (VENTURA, 2007).

Para auxiliar a realização deste trabalho, também foi adotado o método de revisão sistemática de literatura, proposta por (KITCHENHAM, 2004). Através da revisão sistemática é possível realizar o levantamento das principais referências dentro do contexto e objetivos deste trabalho.

1.5 Organização do Documento

Este documento está organizado em seis capítulos, contando com esta introdução. Os tópicos a seguir descrevem os assuntos abordados em cada um dos capítulos restantes, sendo eles:

- Capítulo 2: apresenta os principais conceitos de *big data*, *data analytics*, desafios e fases de *big data*.
- Capítulo 3: apresenta os resultados da revisão sistemática de literatura, cujo objetivo foi identificar propostas de arquiteturas de *big data*.
- Capítulo 4: descreve a arquitetura *Salutem*, apresentando suas principais características e o detalhamento de suas camadas.
- Capítulo 5: apresenta o estudo de caso, que ilustra a utilização da arquitetura, o qual analisou os exames de hemogramas realizados no HU-UFS de 2004 a 2016, com vistas a prover serviços demandados.

- Capítulo 6: apresenta as considerações finais do trabalho, identificando as principais contribuições, as dificuldades e limitações encontradas, bem como os trabalhos futuros.

CAPÍTULO 2

BIG DATA

Neste capítulo apresenta a definição do que vêm ser *big data*, as fases que normalmente contemplam *big data*, algumas aplicações de *big data* e seus desafios.

2.1 Definição

Big data é um conceito abstrato que possui outras características além da ideia de dados massivos. Embora a importância de *big data* ter sido reconhecida, ainda existem muitas definições para o termo na visão de muitos autores. Em linhas gerais *big data* são conjuntos de dados que não puderam ser percebidas, adquiridos, gerenciados e processados por ferramentas tradicionais de TI (Tecnologia da Informação) dentro de um tempo tolerável. Existem outras definições para o termo *big data* na visão de outros autores, haja vista ser um termo novo e conceituação abstrata, nessa perspectiva o Quadro 1 mostra algumas dessas definições (CHEN; MAO; LIU, 2014a; SCHNEIDER, 2012).

Quadro 1: Definições de *big data*.

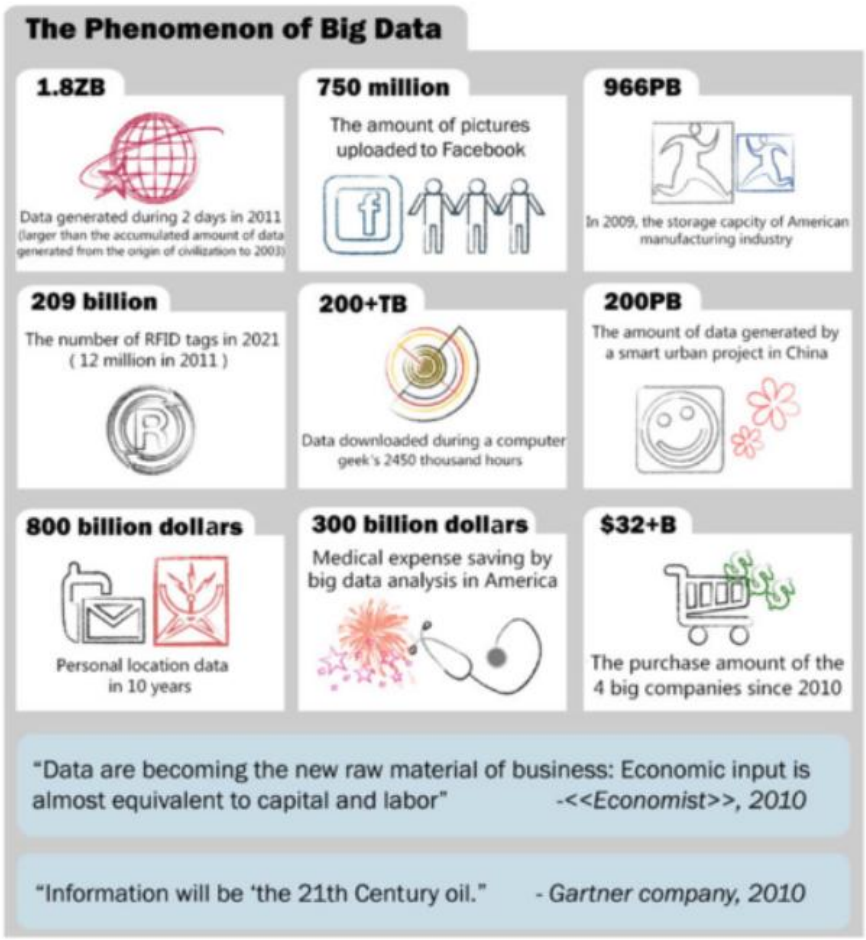
Autor(es)	Definição
(CHEN; MAO; LIU, 2014a)	São conjuntos de dados que não puderam ser percebidas, adquiridos, gerenciados e processados por ferramentas tradicionais de Tecnologia da Informação (TI) dentro de um tempo tolerável.
(SCHNEIDER, 2012)	É um termo descrito por pelo menos três tendências distintas e inter-relacionadas que são: captura e gerencia de muitas informações, trabalhar com muitos novos tipos de dados, e exploração das massas de dados e dos novos tipos de dados com novos estilos de aplicações.
(KRISHNAN, 2013)	São volumes de dados disponíveis em

	diferentes graus de complexidade, geradas em velocidades diferentes e vários graus de ambiguidade, que não podem ser processados usando tecnologias tradicionais, métodos de processamento, algoritmos, ou quaisquer soluções comerciais.
(HURWITZ *et al.*, 2013)	E definido como qualquer tipo de fonte de dados que tem pelo menos três características comuns: volume extremamente grande de dados, velocidade extremamente alta dos dados e variedade extremamente ampla de dados.

Apesar das definições serem distintas entre si, observa-se semelhanças no que tange as características, como: de volume, velocidade e variedade de dados.

O crescimento do volume de dados em diversos segmentos e tecnologia apresentam números consideráveis. Alguns exemplos são apresentados na Figura 1.

Figura 1: *Big data* em números.



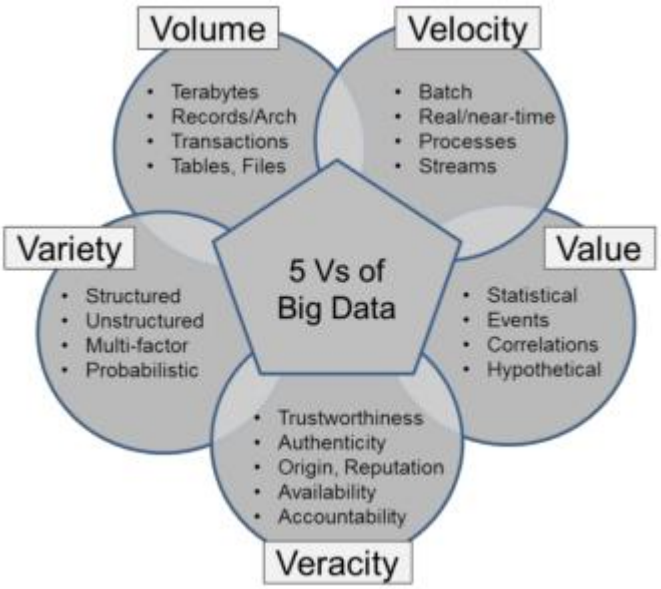
Fonte: Chen *et al.* (2014).

Para (CHEN *et al.*, 2014), o crescimento na geração de dados importados para *big data* se deve principalmente a explosão de tecnologias que armazenam os dados através de sensores, transações comerciais, interações de mídia social, uploads de vídeo e áudio e sinais GPS (*Global Positioning System*) de telefones móveis. Levando em consideração números é estimado que a cada minuto usuários enviem mais de 204 milhões de mensagens, Google receba mais de 2 milhões de consultas, os usuários do Facebook compartilhem mais de 684 mil posts, consumidores gastem mais de U$270 mil em lojas na Web, usuários do Twitter enviem 100 mil tweets, mais de 570 novos websites são criados, e usuários do Instagram compartilham 3600 mil novas fotos. Todas essas fontes de dados devem contribuir para que se tenha 35 *zetabytes* de dados em 2020 (SANTOVENA, 2013).

Big data é caracterizado sobre a ótica dos Vs, sendo eles: volume, variedade e velocidade (Figura 2). Esses Vs, também conhecidos como dimensões, são características que definem o que é um *big data.* Existem autores que adicionam outras dimensões, como valor,

veracidade, entre outras (CHANDARANA; VIJAYALAKSHMI, 2014; LIU; YANG; ZHANG, 2013; RAMESH, 2015).

Figura 2: Principais dimensões de *big data*.



Fonte: Demchenko *et al.* (2013).

O volume é a dimensão mais evidente quando se fala em *big data*, justamente pela grande quantidade de dados disponibilizados para serem analisados e processados. Antes o aumento no volume de dados se dava principalmente pelo aumento no número de transações e seus detalhes em relação à granularidade. Esse aumento é pequeno levando em consideração a dimensão volume em soluções de *big data*. O volume em *big data* começou com o crescimento das interações dos usuários na Web, os chamados weblogs, haja vista a adoção e por consequência popularização da Web. Através desses logs era possível compreender o comportamento dos usuários que utilizam a Web. Depois vieram as redes sociais, tais como Tweeter e Facebook, que combinadas com o aumento da capacidade e tipos de dispositivos para Internet causaram um grande salto na produção de dados. Todo esse volume de dados precisa ser armazenado e, por conseguinte, analisado para retirada de informações pertinentes. Resumindo a dimensão volume significa que, com a geração e a coleção de massas de dados, a escala de dados torna-se cada vez maior (CHEN; MAO; LIU, 2014b; RAMESH, 2015).

A dimensão variedade de dados é tão diversificada quanto as fontes de dados e seus formatos. Os dados a serem analisados advinham principalmente de sistemas transacionais,

onde os dados eram armazenados de forma altamente estruturada. No entanto, com o passar do tempo, outros tipos de dados começaram a ser armazenados e também analisados. Entre esses outros tipos de dados estão os chamados dados semiestruturados, e também os não estruturados, como é o caso de arquivos de áudio, vídeo, textos, entre outros. Essa variedade de dados torna a análise cada vez mais desafiadora, ainda mais combinando a dimensão variedade com a dimensão volume (DEMCHENKO *et al.*, 2013; RAMESH, 2015).

A velocidade é um fator crucial para se trabalhar com *big data*. Essa dimensão está ligada a velocidade em que os dados devem ser armazenados e, por conseguinte, analisados de modo que possa gerar o máximo de valor comercial, no menor tempo possível ou tempo real (CHEN; MAO; LIU, 2014a; DEMCHENKO *et al.*, 2013).

Já a dimensão valor define o valor que os dados podem acrescentar para a organização. Ou seja, os dados por si só não geram informações pertinentes para o usuário, sendo necessário identificar, transformar e extrair dos dados tais informações, através de análises, que agregam valor para o negócio (DEMCHENKO *et al.*, 2013; LIU; YANG; ZHANG, 2013).

E, por fim, a dimensão veracidade, a qual inclui dois aspectos: a consistência dos dados (certeza) que pode ser definida pela sua confiança estatística e a confiabilidade dos dados, que é definida por uma série de fatores, incluindo a origem dos dados, a coleta e os métodos de processamento, em adição uma infraestrutura fácil e confiável. De forma geral, a veracidade garante que os dados utilizados são confiáveis, autênticos e estejam protegidos contra acessos não autorizados (DEMCHENKO *et al.*, 2013).

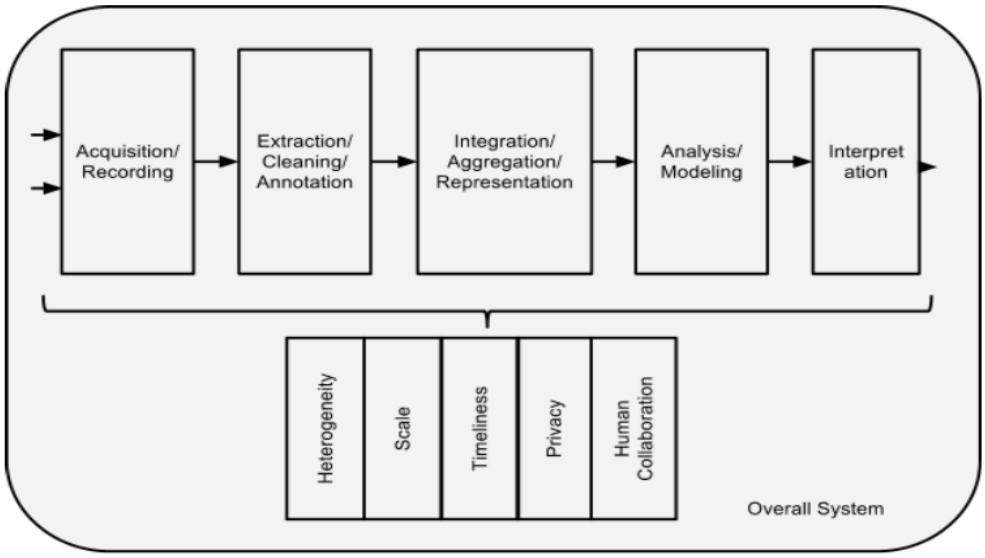
Como pode ser visto, *big data* possui características bem definidas para ser considerado como tal. De acordo com (ZIKOPOULOS; EATON, 2011), soluções em *big data* são ideais: para analisar não somente dados estruturados, mas também semiestruturados e não estruturados, a depender da variedade de suas fontes; para analisar todos os dados, ou boa parte deles, contra uma amostra; e, para análise exploratória e iterativa quando as medidas de negócio não são predeterminadas.

2.2 Fases de Implantação de *Big Data*

Segundo Labrinidis e Jagadish (2012), a análise de dados de *big data* envolve várias fases, tendo cada uma delas suas particularidades e desafios. Na Figura 3 são mostradas as fases pertinentes à análise, como proposta pelos autores, sendo elas: aquisição e gravação;

extração, limpeza e anotação; integração, agregação e representação; análises e modelagem; e interpretação.

Figura 3: Fases de análise de *big data*.



Fonte: Labrinidis e Jagadish (2012).

Na fase de aquisição e gravação todos os dados são adquiridos das mais variadas fontes e formatos para então serem armazenados. Um ponto importante é o fato de nem todos os dados armazenados sejam interessantes, sendo, portanto, passíveis de serem filtrados e comprimidos (LABRINIDIS; JAGADISH, 2012).

Já a fase de extração, limpeza e anotação se preocupam em deixar os dados num formato tal que possam ser trabalhados a posteriori. Na maioria das vezes, os dados disponibilizados para serem armazenados estão em formatos que impossibilitam sua análise. Também, nem sempre os dados estão em sua forma limpa, ou seja, na forma que os usuários possam ter o mesmo entendimento. Por exemplo, o dado sexo, pode estar disponibilizado de forma abreviada com a primeira letra significativa, ou mesmo as três primeiras letras, ou de forma completa. Isso é um problema para análise, haja vista que pode tornar todo o processo falho e consequentemente sem a geração da informação correta (LABRINIDIS; JAGADISH, 2012).

Na fase de integração, agregação e representação são responsáveis, como deixa claro seu nome, pela integração, agregação e representação dos dados. Os dados obtidos de diversas fontes são muitos heterogêneos, o que dificulta qualquer pessoa a encontrar possíveis informações ou mesmo reutilizá-las, caso encontre. A análise de dados em grande escala deve ser um processo totalmente automatizado, em que as diferenças na estrutura e na semântica

dos dados devam ser expressas de forma compreensível para os computadores. A integração de todos os dados é uma tarefa árdua, que necessita ser bem feita, para que seja possível chegar a uma análise eficaz e com o resultado esperado (LABRINIDIS; JAGADISH, 2012).

Ainda de acordo com Labrinidis e Jagadish (2012), a penúltima fase é a de análise, também conhecida como *data analytics*. Esta fase é encarregada de encontrar informações pertinentes para a aplicação através da mineração dos dados e algoritmos de aprendizado de máquina. Além disso, o *data anaytics* trabalhando em conjunto com *big data* formam grandes redes de informações heterogêneas, com redundância de informação que podem ser exploradas: para compensar a falta de dados e cruzar casos conflitantes, para validar relações de confiança, para divulgar os clusters (os grupos) inerentes e para descobrir relações e modelos ocultos. O *data analytics* será melhor detalhado na seção 2.5 deste capítulo.

E, por fim, a fase de interpretação se preocupa em realizar a intepretação correta a partir da análise feita. De nada adianta apenas fornecer os resultados para o usuário, sendo então importante mostrar como tais resultados formam derivados baseados nas entradas. Normalmente o tomador de decisão precisa interpretar a análise para retirar informações pertinentes para o negócio, sendo também necessário realizar suposições e refazer análises. Os resultados devem ser mostrados de uma forma intuitiva e de rápido acesso para que, assim, seja de fácil assimilação e compreensão pelos usuários (LABRINIDIS; JAGADISH, 2012).

2.3 Aplicações de *Big Data*

Big data pode ser aplicado em diversos segmentos de mercado, como por exemplo: mídias sociais, indústria financeiras, redes varejistas, saúde, entre outros.

A utilização de mídias sociais vem crescendo muito como pode ser percebido pelo grande número de usuários ativos e quantidades de dados gerados pelos mesmos. Entre as mídias mais populares estão o Facebook e o Twitter. As principais estratégias ao se analisar mídias sociais incluem: marketing de produto, promoção de marca, identificação de novas oportunidades de vendas, atendimento ao cliente etc. Uma forma de realizar o marketing de produtos nas mídias sociais é direcionando propagandas para os usuários dessas mídias que podem influenciar outros usuários, principalmente os usuários que possuem muitas pessoas conectadas a eles (MOHANTY; BHUYAN; CHENTHATI, 2015).

Na indústria financeira muitos dados são gerados através de serviços diversos, tais como, conta corrente e poupança, cartão de crédito e débito, internet banking, empréstimos entre outros, porém muitos desses dados não são analisados, pois não são vislumbradas as

informações que podem ser obtidas. Esse fato vislumbra uma oportunidade em compreender e identificar as necessidades dos clientes, além de oferecer novos serviços e consequentemente gerar mais rentabilidade. O que a indústria vem fazendo é a gestão de riscos e detecção de fraudes para minimizar a exposição dos bancos. Outro interesse da indústria financeira é a retenção de clientes, ou seja, não perder clientes para outras instituições (MOHANTY; BHUYAN; CHENTHATI, 2015).

Na indústria varejista não é diferente, principalmente as grandes redes, como Walmart e Amazon, que fazem uso de *big data*. Uma das estratégias utilizadas é a recomendação de produtos a clientes com base no que o mesmo está ou vem comprando. Os dados são obtidos de múltiplos canais e com base no histórico de compras do cliente e também dos padrões de buscas, são geradas recomendações de alta qualidade o que consequentemente pode resultar num ganho de receitas para a organização (MOHANTY; BHUYAN; CHENTHATI, 2015).

Outro setor que pode se beneficiar de *big data* é a área saúde. *Big data* pode revolucionar a saúde, entre outras, através de: otimização e redução de custos, alertando qual departamento necessita ser reorganizado, ou mesmo avaliar e monitorar em tempo real qualidade dos serviços, o desempenho das unidades médicas, e as necessidades em recursos humanos e em equipamentos médicos; melhor compreensão da evolução de doenças através da análise de grandes massas de dados sobre vírus e Ácido Desoxirribonucleico (DNA), e assim ajudar médicos e pesquisadores em encontrar novas formas de prevenção de doenças hereditárias e genéticas; melhor tomada de decisão por parte do médico ao analisar os históricos dos resultados das cirurgias dos pacientes em combinação com a análise dos sintomas atuais ou mesmo dos registros históricos do mesmo e, assim, encontrar a melhor intervenção e tratamento para os pacientes; melhor prevenção através da análise dados clínicos públicos e privados para evitar que epidemias se espalhem, isso de forma preditiva, ou seja, antes que aconteça o fato. O principal desafio do setor da saúde é a integração das diferentes fontes de dado, o que demanda um grande esforço de médicos e profissionais da saúde de instituições privadas e públicas. Outro desafio é conectar e integrar múltiplas plataformas médicas, haja vista que essa integração exige a gestão segura dos registros médicos e o respeito a privacidade dos pacientes de acordo com as leis e regulamentos vigentes (BENJELLOUN; LAHCEN; BELFKIH, 2015).

2.4 Desafios de *Big data*

Para se trabalhar com *big data* e colocá-lo em prática, surgem também numerosos desafios. O grande volume de dados que cresce a cada dia com a velocidade para a geração de *big data* que necessita ser muito rápido e também com a variedade dos dados que estão expandindo cada vez mais (CHANDARANA; VIJAYALAKSHMI, 2014). Alguns dos desafios são:

- Representação dos dados: representar os dados tem como objetivo tornar os dados mais significativos para a análise e interpretação do usuário. Dessa forma, uma representação inadequada irá reduzir o valor dos dados e pode também obstruir a análise dos mesmos de forma eficaz (CHEN; MAO; LIU, 2014a);
- Privacidade e segurança: por exemplo, ao se analisar dados pessoais em conjunto com outras massas de dados, podem gerar inferências particulares sobre determinada pessoa que podem ser secretas ou mesmo não permitida à divulgação. Ou mesmo, quando se trabalha com *big data* em sistemas financeiros, pode ser que haja o dado do número do cartão de crédito das pessoas analisadas. Sendo assim, o fator segurança é importante ao se trabalhar com *big data* (CHEN; MAO; LIU, 2014a; KATAL; WAZID; GOUDAR, 2013);
- Armazenamento e processamento: o problema de armazenamento dos dados produzidos todos os dias, haja vista que não existe espaço suficiente para guardar todos os dados. Uma possível solução seria o armazenamento na nuvem, mas existe o problema da realização de upload de arquivos muitos grandes, além da demora, para finalizar tal tarefa. Também seria muito difícil conseguir carregar grandes massas de dados em tempo real e posteriormente processá-los de forma rápida, já que normalmente as nuvens trabalham de forma distribuída (KATAL; WAZID; GOUDAR, 2013).
- Análise semântica: existem aplicações de *big data* que necessitam compartilhar os dados entre seus atores de forma padronizada e que não gere interpretações ambíguas. Uma possível solução seria a utilização de ontologias para que se mantenha uma unicidade de vocabulário e consequentemente intepretações em contextos de aplicações (KIONG; PALANIAPPAN; YAHAYA, 2011; MEZGHANI *et al.*, 2015).

2.5 *Data Analytics*

Com *big data* e o aumento exponencial de dados armazenados, as empresas não conseguem processá-los para extrair valor em sua totalidade. O que se vê normalmente são as empresas analisarem apenas subconjuntos de seus dados, o que pode não corresponder necessariamente a uma vantagem em relação aos seus concorrentes. Através do *data analytics*, as organizações podem aprender mais sobre seus negócios, operações e, consequentemente, traduzir esse conhecimento em melhoria na tomada de decisão e/ou no seu desempenho (CHANDARANA; VIJAYALAKSHMI, 2014; MOHAMED *et al.*, 2014).

O *data analytics*, também conhecido como *data science*, lida com a extração de conhecimento a partir de grandes volumes de dados, independentemente da estrutura dos mesmos (vídeos, texto, XML etc.). Segundo Godbole *et al.* (2015), o *data analytics* também faz parte do processo de *Knowledge-Discovery in Databases* (KDD). Existem quatro modelos para o *data analytics* (GODBOLE; LAMB, 2015):

- Preditivo: que procurar localizar relações de casualidade e padrões entre as variáveis explicativas e as variáveis dependentes, ou seja, se esforça para prevê determinado conhecimento. Exemplo: Próxima preferência de um cliente, falha de sistema (BABU; SASTRY, 2014; LAU; YEW, 2015);
- Descritivo: explora dados históricos para responder "o que aconteceu/está acontecendo?". Exemplo: preferência de produtos, segmentação de clientes com base em características sócio demográficas (BABU; SASTRY, 2014; LAU; YEW, 2015; SHAO; SHIN; JAIN, 2014);
- Diagnóstico: procura responder "por que isso aconteceu/está acontecendo?", ou seja, ajuda a identificar as causas que conduzem a tal acontecimento (LAU; YEW, 2015; SHAO; SHIN; JAIN, 2014);
- Prescritivo: procura responder "o que deve ser feito para que isso aconteça? ". Concentra-se em identificar políticas e entradas que levarão a performance desejada. Exemplo: gerenciar pacientes para identificar possíveis obesos e focar num tratamento (LAU; YEW, 2015; SHAO; SHIN; JAIN, 2014).

A utilização do data analytics traz consigo a necessidade de aumentar a agilidade no acesso aos dados e no processamento dos mesmos para gerar análises mais sofisticadas. Tais análises têm, por sua vez, a necessidade de buscar: algoritmos de aprendizado de máquina, e

análises estatísticas, técnicas de mineração de dados etc., para serem mais eficientes (AHMED *et al.*, 2015; LAU; YEW, 2015).

2.6 Considerações finais do capítulo

Este capítulo apresentou os conceitos em *big data*, assim como, as fases envolvidas em sua implantação e algumas áreas de aplicações, como por exemplo, na área da saúde. também foi detalhado o *data analytics*, técnica que possibilita a análise e transformação de grande volume de dados, brutos ou não, em informações com valor. E, por fim, foi discutido alguns desafios pertinentes a utilização de *big data.*

Estes desafios estão relacionados às caractertísitcas das soluções para *big data* e impactam na implantação e utilização pelas organizações. Assim, a adoção de uma arquitetura deve levar em consideração tais características. No próximo capítulo será discorrido sobre arquitetura de *big data*, as propostas encontradas na literatura e as características atendidas.

CAPÍTULO 3

ARQUITETURA PARA *BIG DATA*

Segundo (BASS; CLEMENTS; KAZMAN, 2012) existem 13 pontos que determinam a importância de uma arquitetura, como por exemplo, a análise de uma arquitetura permite realizar uma previsão da qualidade de um sistema, uma arquitetura documentada melhora a comunicação entre as partes interessadas, entre outros. Com relação a soluções de *big data* não é diferente, sendo importante definir uma arquitetura que ajude aos usuários a chegarem aos resultados de forma mais rápida. Tendo isto como premissa, este capítulo descreve as arquiteturas para *big data* encontradas na literatura assim como a análise destas.

3.1 Arquitetura

De acordo a norma ISO/IEC/IEE 42010, arquitetura de um sistema constitui o que é essencial sobre o sistema considerado em relação ao seu ambiente. Uma arquitetura de sistema pode expressar componentes, ou elementos do sistema; como elementos do sistema são organizadas ou inter-relacionadas; princípios e padrões de organização ou *design*; e princípios que regem a evolução do sistema ao longo do seu ciclo de vida.

Para o *The Open Group Architecture Framework* (TOGAF), que é um *framework* de arquitetura corporativa, arquitetura de sistema é a descrição formal de um sistema, ou um plano detalhado do sistema a um nível de componente para orientar a sua implementação; e a estrutura de componentes, suas inter-relações e os princípios e diretrizes que regem seu *design* e evolução ao longo do tempo (HARRISON, 2013).

Com o intuito de identificar e mapear as arquiteturas apresentadas na literatura para *big data* foi realizada uma revisão sistemática de literatura, conforme protocolo detalhado na próxima seção.

3.2 Protocolo de Revisão Sistemática de Literatura

Adotou-se como forma de pesquisa, a revisão de literatura proposta por Kitcheham *et al.* (2004) e definido inicialmente o protocolo a ser seguido. No protocolo primeiramente foram definidas as variáveis de interesse deste trabalho, que foram:

1. País de publicação.
2. Dimensões (Volume, Velocidade, Variedade, Veracidade e Valor) encontradas.
3. Aplicação (saúde, setor financeiro, cidades inteligentes etc.).
4. Ferramentas utilizadas, funções das mesmas e se *open source* ou não.
5. Objetivo do estudo (proposta de arquitetura, proposta de f*ramework*, *survey*, caracterizadores, entre outros).
6. Descrição de estudo de caso.

A revisão teve início em março de 2016, pesquisando em banco de dados bibliográficos. A *string* de busca foi montada com o objetivo de obter o maior número de estudos sobre *big data*, com foco principal em propostas de arquiteturas, plataformas e tecnologias. Somente artigos em inglês foram incorporados. Não foi estabelecido nenhum critério em relação à data de publicação.

As bases utilizadas para a pesquisa foram: IEEE Explorer e Spring Link, sendo a string de busca: *Big Data AND* (*Architecture OR Platform OR Technologies*).

Após a consulta inicial, retornaram 5311 artigos, encontrados nas duas bases consultadas (ver Quadro 2).

Quadro 2: Quantitativo de artigos selecionados na execução do protocolo.

Etapa	IEEE	Springer Link
Após a consulta da *string* de busca	2051	3260
Não são artigos	104	101
Acesso restrito	0	102
Não são artigos completos	586	26
Não contém *big data* no título ou resumo ou palavras-chaves	168	2645
Removidos após a leitura do título	304	130
Removidos após a leitura do resumo	522	157
Resultado	367	99

Na primeira fase foram retiradas as publicações que não eram artigos. Depois foram retirados os artigos que não estavam disponibilizáveis, ou por serem pagos ou mesmo patentes. Continuando as remoções, os estudos que não estavam completos, ou seja, que estavam na forma de short *paper,* foram excluídos. Também foram excluídos os artigos que não continham a palavra *big data* no título ou resumo ou nas palavras-chave, o que resultou numa remoção quantitativa. Isso acontece haja vista que as *engine* de busca de cada base são diferentes na filtragem das publicações.

Na segunda fase, foram retiradas as publicações a partir de leituras, primeiramente do título e a posteriori do resumo. Essa fase foi a mais trabalhosa, já que foi necessária uma leitura minuciosa para incluir ou não os artigos. Foram retirados principalmente artigos que possuíam essência voltada para parte de infraestrutura e não possuíam foco principal em *big data*. Ao final dessa fase restaram 466 artigos. Nesta fase foi construído um *software* para ajudar na análise das variáveis de interesse deste trabalho, e consequentemente armazenar os dados extraídos num banco de dados. Dessa forma, foi possível analisar as publicações mais facilmente e de forma mais rápida.

Dos 466 artigos, 56 propunham uma arquitetura que será discutida na próxima seção.

3.3 Arquiteturas para *Big Data*

Com base na revisão sistemática realizada, foram identificados 56 artigos que propõem arquiteturas para *big data*. Uma análise destes trabalhos é apresentada no Quadro 3, onde foram identificadas as características mais discutidas e indicadas como relevantes para uma arquitetura de *big data* pelos autores, a saber:

- Analisa os dados, ou seja, realiza o processamento de dados para obtenção de informações e/ou conhecimento para o negócio;
- Armazena os dados, ou seja, se armazena os dados coletados e/ou analisados;
- Coleta os dados, ou seja, se adquirem os dados a partir de fontes de dados diversas;
- Análise semântica dos dados, ou seja, os dados seguem algum padrão para que a intepretação dos mesmos seja comum a todos os envolvidos;
- Disponibiliza serviços, ou seja, disponibiliza ao final da análise dos dados algum tipo de serviço;
- Garante segurança dos dados, ou seja, se arquitetura dispõe de alguma forma de segurança para os dados.

Quadro 3: Análise propostas de arquitetura.

Artigo	Coleta	Análise de Dados	Armazenamento	Comunicação de Dados	Processamento	Visualização/ Serviço	Segurança	Análise Semântica
GEERDINK, 2013		✓	✓					
ZHANG *et al.*, 2013		✓	✓					
HU *et al.*, 2013				✓				
XINHUA *et al.*, 2013			✓					
KHAN; ANJUM; KIANI, 2013	✓	✓	✓					
ROOPA; IYER; RANGASWAMY, 2013	✓	✓					✓	
MADKOUR e AREF; BASALAMAH, 2013	✓	✓	✓					✓
TOMIC e FENSEL, 2013		✓	✓					✓
KOWSAR e AFGAN, 2013					✓			
ZULKERNINE *et al.*, 2013		✓	✓			✓	✓	
ZIMMERMANN *et al.*, 2013								
MARCHAL *et al.*, 2014		✓	✓		✓		✓	
MAURO e SARNO, 2014		✓			✓			
AGRAWAL *et al.*, 2014			✓			✓	✓	
VIANA; SATO, 2014			✓				✓	

Artigo	Coleta	Análise de Dados	Armazenamento	Comunicação de Dados	Processamento	Visualização/ Serviço	Segurança	Análise Semântica
BALDOMINOS *et al.*, 2014		✓	✓		✓	✓		
CRICHTON *et al.*, 2014			✓					✓
KASHLEV; LU, 2014			✓					
KRUMEICH *et al.*, 2014		✓	✓			✓		
CECCHINEL *et al.*, 2014	✓		✓					
LIU *et al.*, 2014		✓	✓		✓			
AMATO; DI MARTINO; VENTICINQUE, 2015		✓	✓			✓		✓
RAYNAUD; HAQUE; AÏT-KACI, 2015			✓					
KOWSAR; DASHNOW; LONIE, 2015		✓	✓					
ELLEUCH; WALI; ALIMI, 2015	✓		✓					
TWARDOWSKI e RYZKO, 2014		✓	✓		✓			
HUGHES *et al.*, 2014			✓					✓
PENA; SARKAR; MAHESHWARI, 2015		✓	✓		✓		✓	
HUANG; YESHA; ZHOU, 2015			✓					
LIU *et al.*, 2015		✓	✓					

Artigo	Coleta	Análise de Dados	Armazenamento	Comunicação de Dados	Processamento	Visualização/ Serviço	Segurança	Análise Semântica
FERNANDEZ *et al.*, *2015*			✓					
MEZGHANI *et al.*, 2015		✓	✓			✓		✓
BALDOMINOS *et al.*, 2015		✓	✓			✓		
HUANG *et al.*, 2015		✓	✓					
GACHET *et al.*, 2015		✓	✓ Parcial					
IDRIS *et al.*, 2015		✓	✓		✓			
CHENG *et al.*, 2015		✓	✓		✓			
IMMONEN *et al.*, 2015		✓	✓		✓			
LI e NI, 2015		✓						
ABUSHAREKH *et al.*, 2015		✓	✓			✓		✓
RATHORE *et al.*, 2015		✓	✓		✓			
DUAN *et al.*, 2015		✓	✓					
SIRIWEERA *et al.*, 2015		✓	✓					
CAGLIERO *et al.*, 2015		✓	✓					
DANIEL *et al.*, 2015		✓	✓					
JUTLA e BODORIK, 2015		✓	✓					
RATHORE *et al.*, 2015		✓	✓		✓			
KELLY *et al.*, 2015		✓	✓			✓		

Artigo	Coleta	Análise de Dados	Armazenamento	Comunicação de Dados	Processamento	Visualização/ Serviço	Segurança	Análise Semântica
DAIF *et al.*, 2015		✓	✓					
RABELO *et al.*, 2015		✓	✓					✓
SINNOTT *et al.*, 2015		✓	✓					
NI *et al.*, 2015		✓	✓					
SINGH e LIU, 2015			✓					
ZHU *et al.*, 2016			✓					
LARKOU *et al.*, 2016		✓	✓					
WANG *et al.*, 2016			✓					

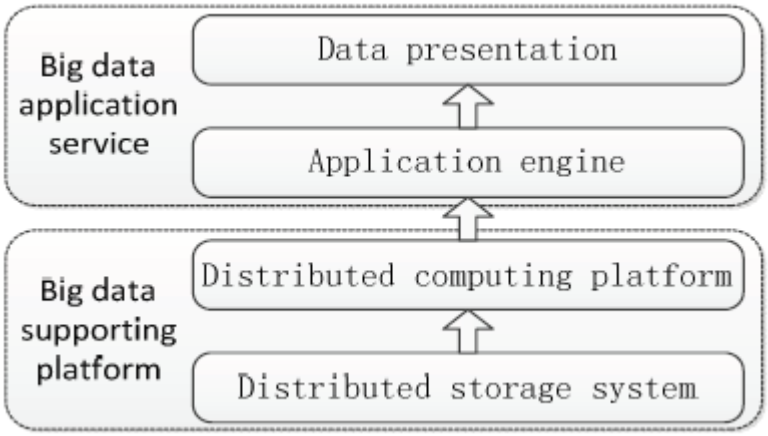
Como se pode perceber, as características mais discutidas nos trabalhos são o armazenamento e análise de dados, correspondendo a 50 e 39 trabalhos, respectivamente. Do total de trabalhos, apenas 10% abordam a característica segurança. E 14% dos 56 trabalhos, discorrem sobre análise semântica de dados.

A seguir serão detalhadas as arquiteturas de *big data* para a área da saúde que foram encontradas através da revisão sistemática. Dos 56 trabalhos selecionados, 5 são arquiteturas para a área da saúde.

3.4 Arquiteturas de *Big Data* para Saúde

A publicação de autoria de (LIU *et al.*, 2015), propõe um arquitetura de referência como solução de plataforma para superar grandes problemas nos processos de saúde em *big data* (ver Figura 4).

Figura 4: Proposta de arquitetura conceitual de *big data* em saúde.



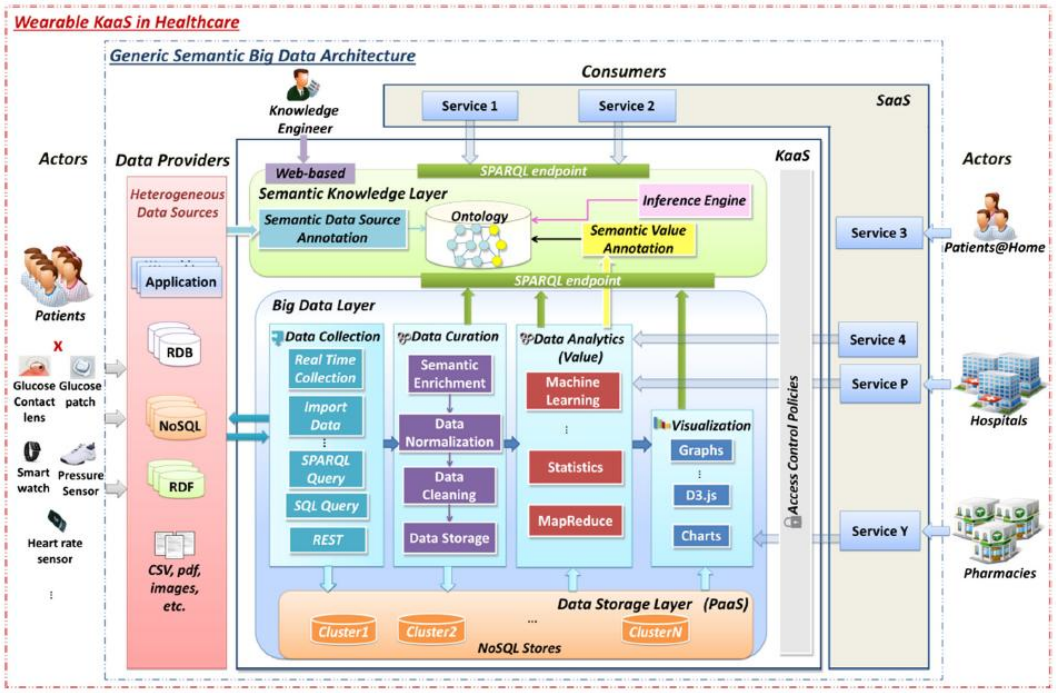
Fonte: Liu *et al.* (2015).

A arquitetura é formada por quatro módulos, respectivamente: sistema de armazenamento distribuído; plataforma de computação distribuída; *engine* de aplicação; e apresentação de dados. O sistema de armazenamento distribuído é projetado para tolerar falhas e que fornece acesso a aplicações de dados e é adequado a aplicações que possuem grandes conjuntos de dados. Já a plataforma de computação distribuída é um modelo de programação com a implementação associada para o processamento e geração de grandes conjuntos de dados. A *engine* de aplicação é uma interface entre a aplicação e os outros dois

módulos. E, por fim, a apresentação de dados, que é responsável por receber o resultado do processamento e disponibilizar para os usuários finais de uma forma intuitiva.

Em (MEZGHANI *et al.* 2015) é introduzida uma arquitetura de *big data* semântica genérica, que é uma extensão da arquitetura *NIST Cloud and Big Data*, baseada em ontologia para dar significado aos repositórios de dados heterogêneos. Os objetivos desta arquitetura são: proporcionar uma solução escalável para o armazenamento de grande volume de dados de saúde gerados por múltiplas fontes; apoiar o compartilhamento e a integração dos dados para uma melhor tomada de decisão; e extração de informação valiosa (conhecimento) a partir de dados heterogêneos. A arquitetura semântica genérica é formada pelas camadas: camada de armazenamento de dados; camada de *big data*; e a camada de conhecimento semântico (ver Figura 5).

Figura 5: Arquitetura extensão da NIST.

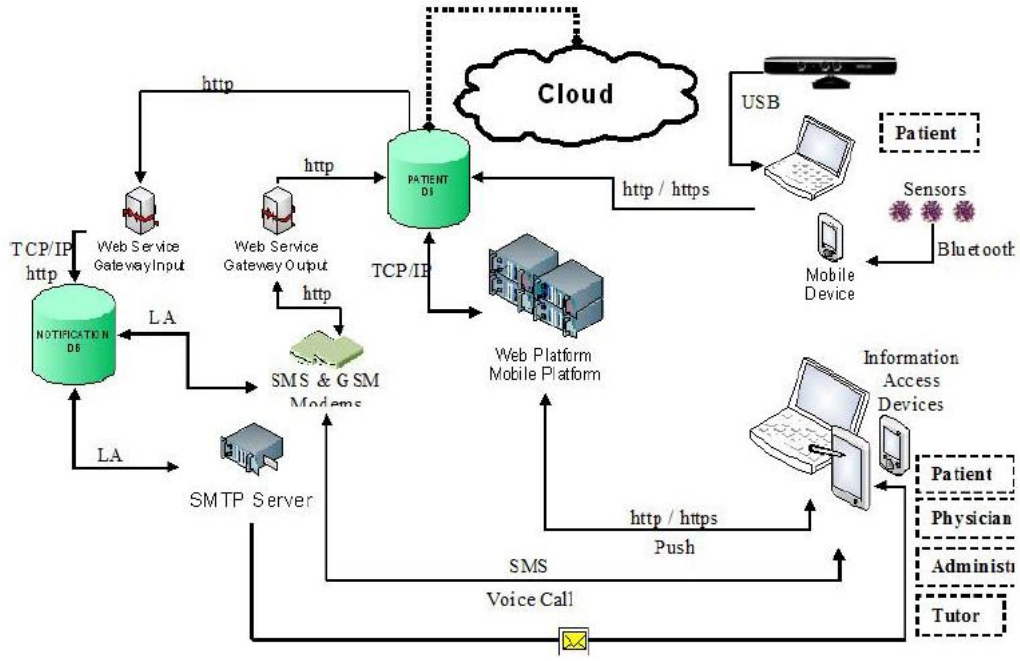


Fonte: Mezghani *et al.* (2015).

A camada de armazenamento de dados é responsável por armazenar grandes conjuntos de dados na nuvem, mais precisamente na *Platform as a Service* (PaaS). A camada de *big data* realiza o processamento dos dados através de tecnologias escaláveis. E a camada de conhecimento semântico oferece a compreensão comum sobre os dados.

Em (GACHET *et al.*, 2015), é proposta uma arquitetura para monitoramento de pacientes através de utilização de sensores (IoT) e *big data* (ver Figura 6).

Figura 6: Proposta de arquitetura para monitorar pacientes com *big data* e IoT.

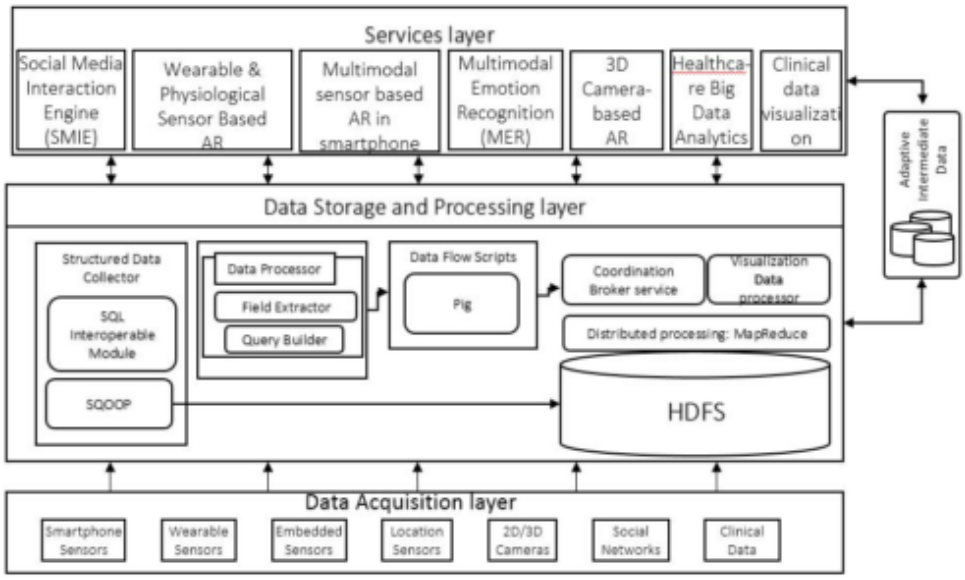


Fonte: Gachet *et al.* (2015).

A arquitetura permite o monitoramento de pacientes crônicos ou não, bem como pessoas saudáveis que precisam ser monitoradas por diferentes circunstâncias. Os principais elementos da arquitetura são: telefones móveis utilizados pelos pacientes que enviam dados obtidos por sensores (sensor que mede sinais vitais, por exemplo); o módulo de armazenamento de dados em nuvem e o módulo de análise de dados, que são o núcleo da arquitetura; a plataforma de interoperabilidade e mensagens para entrega de informação a todos os envolvidos no sistema; e, por fim, a plataforma *Web*, que permite tanto médicos como cuidadores e familiares consultarem informações sobre pacientes.

Em (IDRIS *et al.* 2015) é proposta o *Big Data Service Engine* (BISE) que é uma plataforma transforma os dados multiestruturados armazenados no HDFS (*Hadoop Distributed File System*) num banco dados relacional intermediário, e a partir desse banco de dados realizar reconhecimento de atividades, reconhecimento de emoções, recomendação e análise de sentimento, com base nos dados clínicos dos pacientes, tal como o *Electronic Health Record* (EHR). O BISE é composto por três camadas e um banco de dados intermediário adaptativo (Figura 7).

Figura 7: Arquitetura BISE.

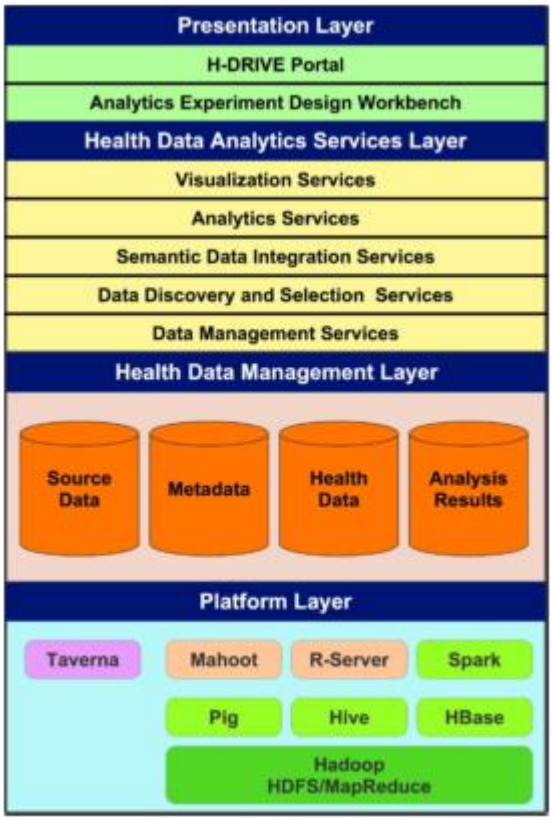


Fonte: Idris *et al.* (2015).

As três camadas são respectivamente: camada de aquisição de dados; camada de armazenamento e processamento de dados; e camada de serviço. O banco de dados intermediário adaptativo é utilizado pela camada de serviços para executar a computação e cálculos sobre os dados relacionais de saúde clínicos, físicos, mentais e sociais dos pacientes. A camada de aquisição de dados do BISE tem a função de obter os dados advindos de diversas fontes, como sensores de s*martphones*, sensores portáteis, sensores de localização, câmaras 2D/3D, registros de saúde, redes sociais, entre outros. Na camada de armazenamento e processamento de dados, os dados armazenados no HDFS são processados de acordo com a necessidade da camada de serviços. E, por fim, a camada de serviços que inclui subcomponentes que fornecem serviços individuais tais como interação em redes sociais, reconhecimento de atividades com base em registros de sensores, reconhecimento de emoções a partir de imagens, visualização de dados clínicos, entre outros.

Em (ABUSHAREKH *et al.* 2015) é apresentado uma arquitetura de análise de dados especializados em saúde chamada de *Health Data Reconcillation Inferencing and Visualization Environment* (H-DRIVE). O H-DRIVE é uma arquitetura de alto nível funcional e técnico, concentrando-se em delinear um subconjunto de funcionalidades em análise de dados para *big data* na área saúde (ver Figura 8). A arquitetura é disposta em quatro camadas, são elas: camada de apresentação, camada de serviços de *data analytics* em saúde, camada de gerenciamento de dados em saúde e a camada de plataforma. A camada de apresentação tem por objetivo expor as análises geradas por meio de *dashboards*, por exemplo.

Figura 8: Arquitetura H-DRIVE.



Fonte: Abusharekh *et al.* (2015).

A camada de serviços de *data analytics* em saúde oferta um conjunto de serviços *restful* que podem ser utilizados pela camada de interface do usuário, pelo *worflow* analítico e por qualquer outro cliente. A camada de gerenciamento de dados de saúde é responsável por armazenar e manter os repositórios. E, por fim, tem a camada de plataforma que é constituída pelas diferentes plataformas e *frameworks* que serão utilizados para executar e apoiar as atividades de análise de dados.

3.5 Análise das arquiteturas para a área da saúde

A análise das cinco arquiteturas propostas para âmbito da saúde baseou-se nas principais características para uma arquitetura de *big data*, adicionando as características estilo arquitetural e tipo de validação da proposta. O Quadro 4 apresenta um resumo das características encontradas nos trabalhos.

Quadro 4: Resumo das arquiteturas de *big data* em saúde selecionadas.

Autor(es)	Análise	Armazenamento	Coleta	Análise semântica	Serviço	Segurança	Validada	Estilo arquitetural
LIU *et al.*, 2015	✓	✓						Não mencionado
MEZGHANI *et al.*, 2015	✓	✓	✓ Parcial	✓	✓	✓ Parcial	✓ Parcial	Camadas
GACHET *et al.*, 2015	✓ Parcial	✓	✓ Parcial					Não mencionado
IDRIS *et al.*, 2015	✓	✓	✓	✓	✓		✓	Camadas
ABUSHAREKH *et al.*, 2015	✓	✓	✓ Parcial	✓	✓		✓	Camadas

A maioria das arquiteturas propostas foi validada por meio de estudos de caso, mesmo que parcialmente. Nenhuma das arquiteturas analisadas apresentou uma solução completa, ou seja, que contemple as características identificadas durante a revisão sistemática. A proposta de Mezghani *et al.* (2015) aborda todas as oito características, sendo que destas três de forma parcial. Foi atribuído parcial pelo fato de que, apesar da característica está presente na arquitetura, estas três não foram detalhadas.

Quanto ao estilo arquitetural, quando mencionado, as propostas adotam o estilo em camadas, refletindo a tendência atual em construção de soluções de *big data*, de acordo com a revisão de literatura realizada. Além disso, com esse estilo arquitetural, possíveis modificações nas camadas são mais fáceis de acontecer, além de ser portáveis para outras arquiteturas (SOMMERVILLE, 2010).

Shirudkar e Motwani (2015) colocam que a segurança e a privacidade dos dados em *big data,* quando se trabalha com registros médicos, são fundamentais e devem estar presentes em qualquer arquitetura. No entanto, nenhuma das propostas de arquiteturas analisadas apresentou a característica segurança e privacidade. Isto pode ser um problema quando se trabalha com dados médicos, clínicos e hospitalares.

Em se tratando de análise semântica, duas propostas abordam esta característica. Dinov (2016) reforça a importância de se manter uma mesma interpretação e semântica dos dados. No ambiente da saúde a análise semântica possibilita a interoperabilidade dos dados,

controle de vocabulário, conjunto mínimo de dados, entre outros recursos. Sendo assim, uma importante característica que deve ser contemplada.

As propostas encontradas, de maneira geral, não definem papéis importantes para orientar a implantação e utilização de *big data*.

Apesar da importância de *big data* para área da saúde, poucas propostas foram encontradas e destas observou-se a pouca atenção com segurança e privacidade, assim como, com a necessidade de interpretação dos dados pelos atores envolvidos.

3.6 Considerações Finais do Capítulo

Este capítulo apresentou os conceitos de arquitetura de sistemas e arquitetura de *big data,* assim como a definição de arquitetura segundo a norma ISO/IEC/IEE 42010. Também foi apresentado o protocolo de revisão sistemática que teve por objetivo identificar as propostas de arquitetura para *big data*, incluindo propostas para área da saúde. Os resultados foram apresentados, assim como, uma análise destes.

Apesar da quantidade expressiva de trabalhos selecionados (466 artigos), apenas 56 apresentaram propostas de arquitetura, sendo que destes, somente 5 propuseram arquiteturas para área da saúde. As propostas foram analisadas à luz das características encontradas nas propostas de arquitetura e os desafios de *big data*, como discutido no capítulo 2.

Segurança e análise semântica são estratégicos para a área da saúde, nesse sentido, no próximo capítulo será apresentada a arquitetura *Salutem,* que tem o intuito de ser uma arquitetura que compreenda as características para soluções de *big data* e seja de fácil aplicabilidade.

CAPÍTULO 4

ARQUITETURA SALUTEM

Para Anthony e Garland (2002), uma arquitetura bem definida produz um sistema que será mais fácil de projetar, desenvolver e manter, bem como sua boa representação geralmente facilita sua compreensão. Com relação a soluções de *big data* não é diferente, sendo importante definir uma arquitetura para facilitar seu entendimento e consequentemente sua aplicação. Tendo isso como premissa e também os desafios de *big data*, como já discutido anteriormente, nessa seção é apresentada a arquitetura *Salutem* e seus componentes.

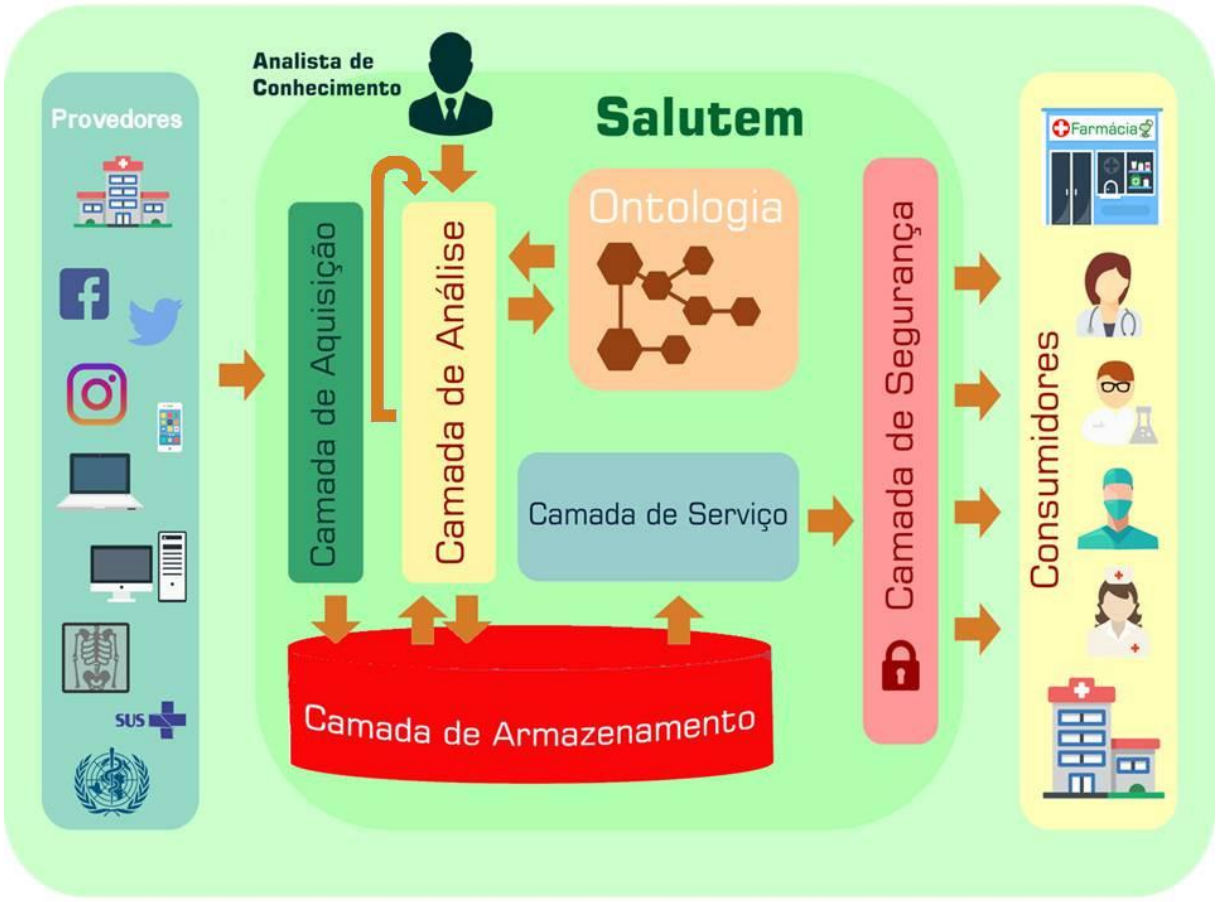
4.1 Estrutura da Arquitetura *Salutem*

Conforme coloca Sommerville (2008), o estilo arquitetural em camadas facilita a definição de responsabilidades, a coesão nas operações inerentes a cada camada, bem como, a independência entre elas. Dessa maneira, o estilo arquitetural em camadas foi adotado para a arquitetura *Salutem*, sendo que o número de camadas foi definido considerando as características essenciais no contexto de *big data*, como discutido no capítulo 3.

Assim, a arquitetura *Salutem*[1] está estruturada em camadas, a saber: camada de aquisição, camada de armazenamento, camada de análise, camada de serviço e camada de segurança (Figura 9). Além destas camadas, a arquitetura contempla também os componentes: ontologia, provedores e consumidores.

[1] *Salutem* significa saúde em latim (MICHAELIS,2017).

Figura 9: Arquitetura *Salutem*.



São responsabilidades de cada camada:

- Camada de aquisição: coletar os dados de diferentes fontes de dados com o objetivo de analisá-los. Também é nessa camada que é realizado o processo de ETL.
- Camada de armazenamento: armazenar os dados coletados e que serão manipulados pela arquitetura.
- Camada de análise: analisar os dados no viés de interesse com objetivo de responder perguntas, identificar *insigths*, entre outros.
- Camada de serviço: apresentar aos consumidores os resultados da análise por meio de serviços, a saber: *dashboard*, gráficos, *web service*, entre outros.
- Camada de segurança: proteger o acesso e a privacidade dos serviços disponibilizados contra acessos indevidos.

Além disso, a *Salutem* contempla também os componentes de ontologia, provedores de dados e consumidores de serviços. A ontologia corresponde ao vocabulário utilizado pelo problema em tela, assim manter a interpretação única pelos usuários. Os provedores de dados

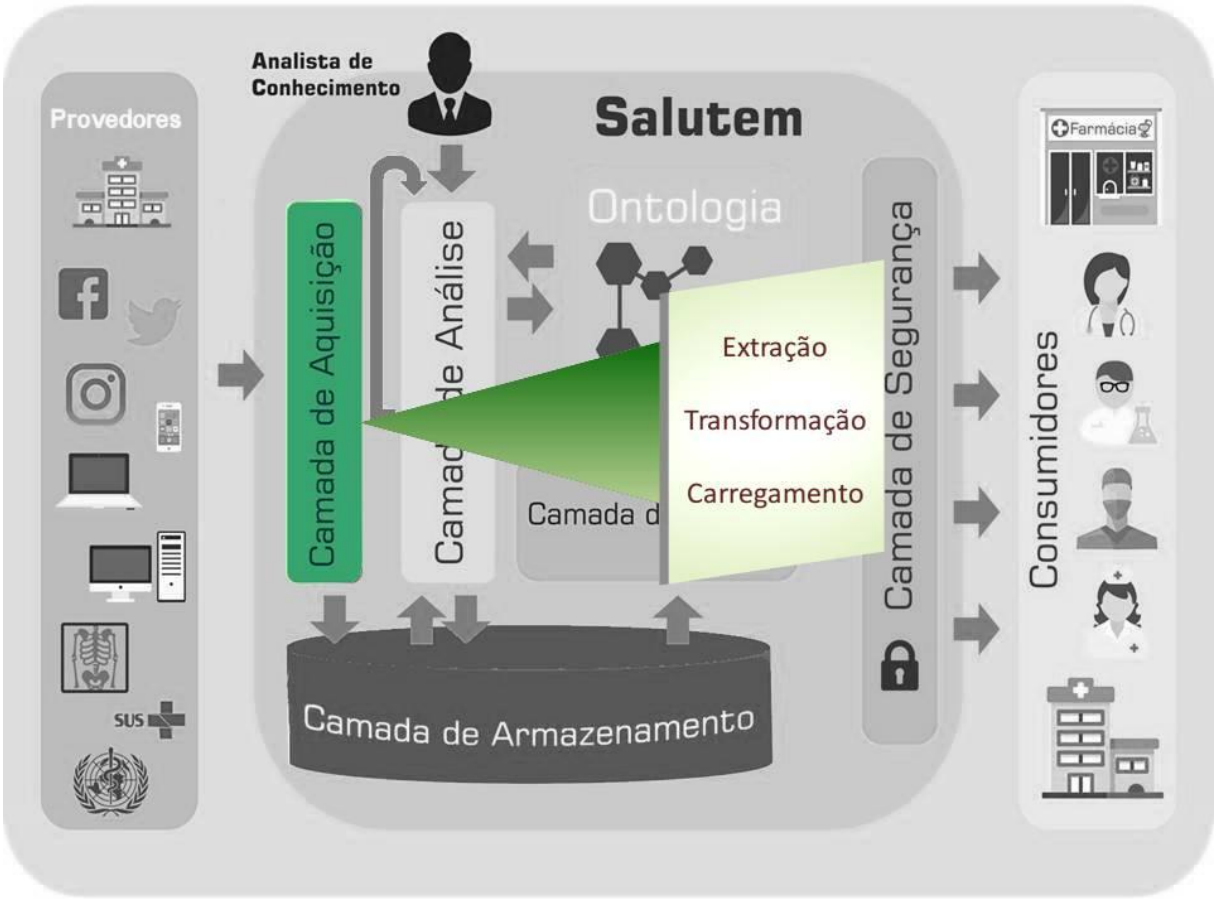
representam os dados em sua origem, que podem estar em diferentes formatos, dispositivos e locais. Já os consumidores de serviços representam os atores que demandam informações disponibilizadas por diferentes serviços. Já os consumidores de serviços vão ser determinados pelo escopo em análise, ou seja, quais papéis tem interesse na análise realizada pela arquitetura, sendo que o acesso ao serviço está condicionado para cada consumidor, atendendo aos parâmetros da camada de segurança.

Nas próximas seções serão detalhadas as camadas, assim como, os atores, os componentes e a taxonomia de referência da arquitetura *Salutem*.

4.2 Camada de Aquisição

A camada de aquisição de dados é responsável por obter, transformar e armazenar os dados de diferentes provedores, como aplicações em saúde, banco de dados, redes sociais, entre outros. A Figura 10 apresenta as principais operações da camada de aquisição.

Figura 10: Camada de Aquisição.



Os dados podem estar em formatos distintos e devem ser transformado para o formato padrão adotado e, dessa forma, poder ser manipulado mais facilmente. Os dados adquiridos são normalmente dados operacionais do dia a dia da organização, e que ajudam a responder/solucionar problemas de negócio. Os dados obtidos podem ser:

- Dados de sistemas: dados dos mais diversos sistemas que podem ser utilizados na área da saúde, como por exemplo, sistema ERP (*Enterprise Resource Planning*), sistema CRM (*Customer Relationship Management*); sistema de gerenciamento de conteúdo, entre outros;
- Arquivos: arquivos nos mais diversos formatos, como: documentos escaneados, formulários diversificados, planilhas eletrônicas, e-mails, entre outros;
- Sensores: sensores que registram dados clínicos dos pacientes, leitos, entre outros;
- Mídias sociais: como por exemplo: Whatsapp, Twitter, entre outros; e,
- Dados abertos: os dados abertos são dados disponibilizados normalmente por órgãos governamentais, com o objetivo de tornar transparente dados diversos, como dados de mortalidade, acesso à saúde, entre outros.

Na maioria das vezes a principal forma de adquirir os dados é por meio de APIs (*Application Programming Interface*). De acordo com Malcolm *et al.* (2014), a API é um conjunto de regras e instruções que uma aplicação pode seguir para acessar serviços e recursos fornecidos por outra aplicação. Uma das formas mais difundidas de acesso a esses APIs é pelo protocolo REST (*REpresentational State Transfer*), que fornece os recursos através das quatro principais operações HTTP (*Hypertext Transfer Protocol*), são elas: GET (retorno), POST (criação), PUT (atualizar) e DELETE (apagar) (PEREZ; CARRERA, 2015).

Depois de coletados, os dados são transformados para que os mesmos possam ser movidos, particionados, limpos, ordenados, separados, entre outros, e consequentemente serem mais facilmente manipulados. Por exemplo, os dados de uma fonte não muito confiável, são limpos de possíveis registros corrompidos ou incorretos.

Esse processo é chamado de ETL (*Extract-Transform-Load*). Para (BANSAL, 2014), o ETL é um processo computacional que é utilizado para integração de dados de múltiplas fontes ou aplicações, possivelmente de domínios diferentes, sendo composta pelas fases:

- Extração: é a primeira fase do processo que envolve a extração de dados das fontes;

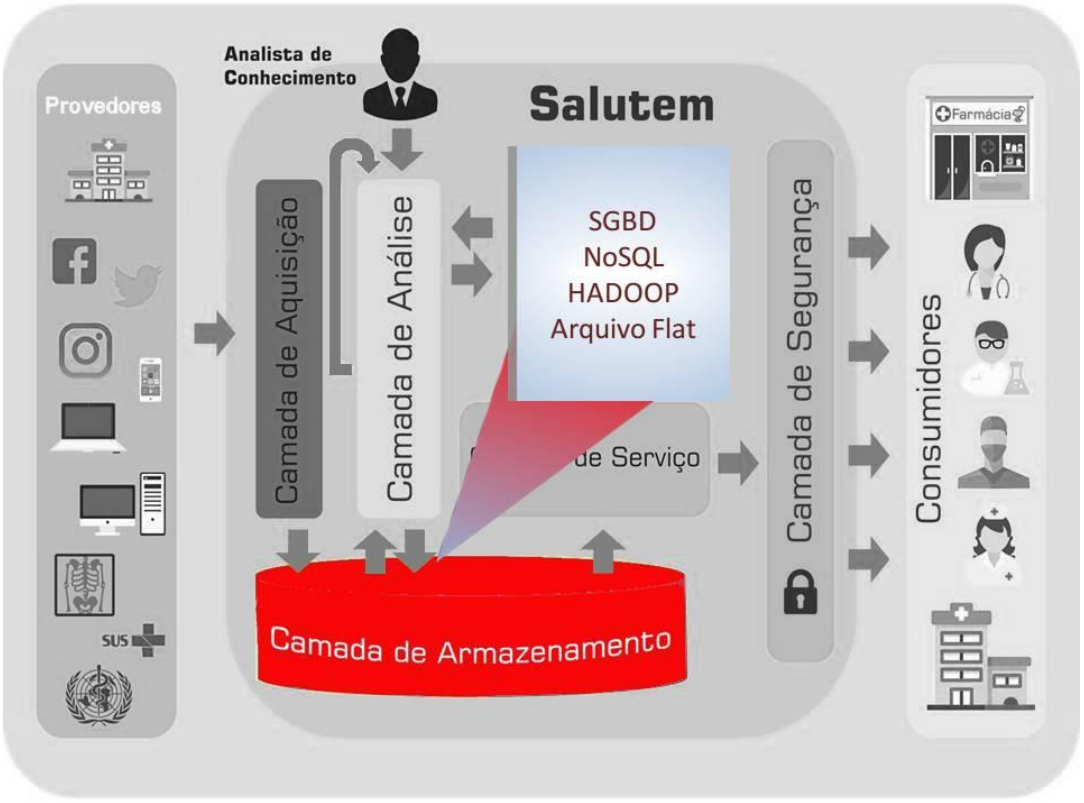
- Transformação: esta fase envolve a limpeza dos dados para torná-los aptos para serem manipulados. Entre as atividades típicas desta fase tem-se: a normalização dos dados, remoção de dados duplicados, verificação de restrições de integridade, filtragem de dados com base em expressões regulares, classificação e agrupamento de dados, entre outros; e,
- Carregamento: é a fase final que envolve armazenar os dados extraídos e transformados onde se deseja.

Segundo Mahmood e Afzal (2013), o ETL é um processo normalmente demorado que pode levar até 60% do esforço total do trabalho desejado. Sendo necessário nesse processo: a remoção ou transformação de valores inconsistentes, incompletos ou faltantes; normalizar, reduzir e tornar discretos os dados; garantir qualidade estatística dos dados através de *boxplots*, análise de agrupamento, testes de normalidade, entre outros; e compreensão dos dados através de estatística descritiva (correlações, testes de hipóteses, histogramas etc.).

4.3 Camada de Armazenamento

A camada de armazenamento é responsável por armazenar os dados coletados e analisados pela *Salutem* em bancos de dados apropriados (Figura 11).

Figura 11: Camada de Armazenamento.



Por poder existir dados de diferentes formatos, como estruturado, semiestruturado e não estruturado, a camada de armazenamento deve estar preparada para este desafio.

Os dados podem ser armazenados em Sistemas de Gerenciamento de Banco de Dados (SGBD) tradicionais, tais como Oracle, Mysql, Postgres, entre outros. As duas principais características dos SGBDs são o suporte a transações ACID (Atomicidade, Consistência, Integridade e Durabilidade), que garantem a integridade dos dados. E também o uso da *Structured Query Language* (SQL), que oferece a criação e a manipulação dos dados. No entanto, em casos em que se precisa de alta escalabilidade, disponibilidade e flexibilidade, os SGBDs não são a melhor alternativa a ser escolhida.

Nesses casos pode ser adotada a tecnologia NoSQL (*Not only* SQL). O NoSQL é a tecnologia de banco de dados não relacionais e que podem acomodar dados não estruturados. Diferentemente dos SGBDs, os bancos de dados NoSQL oferecem uma maior flexibilidade quanto ao esquema a ser adotado (BHOGAL; CHOKSI, 2015). Os bancos de dados NoSQL podem ser de quatro modelos diferentes, a saber:

- Modelo orientado a chave-valor: nesse modelo os dados são armazenados utilizando chaves como identificadores das informações gravadas no campo valor (GUDIVADA; RAO; RAGHAVAN, 2014).
- Modelo orientado a documentos: é uma extensão do modelo chave-valor, com a diferença que os dados armazenados no campo valor são documentos, semiestruturados, como XML e JSON (BHOGAL; CHOKSI, 2015).
- Modelo orientado a colunas: armazenam grupos de dados relacionados em famílias de colunas, sendo que cada registro pode ter um número diferente de colunas (GUDIVADA; RAO; RAGHAVAN, 2014).
- Modelo orientado a grafos: diferente dos demais, nesse modelo os dados são armazenados na forma de grafos, sendo os vértices utilizados para armazenar os dados e as arestas, os relacionamentos entre os mesmos (GROLINGER *et al.*, 2013).

Em caso de dados não estruturados também pode ser utilizado o *Hadoop Distributed File System* (HDFS). O HDFS é projetado para suportar conjuntos de dados volumosos e permite também *clusters* de armazenamento distribuídos, ou seja, vários nós em que os dados são armazenados e lidos, de forma transparente e rápida para o usuário. O HDFS é uma tecnologia *open source* e que permite também o armazenamento de dados estruturados e semiestruturados (MOHANTY; BHUYAN; CHENTHATI, 2015).

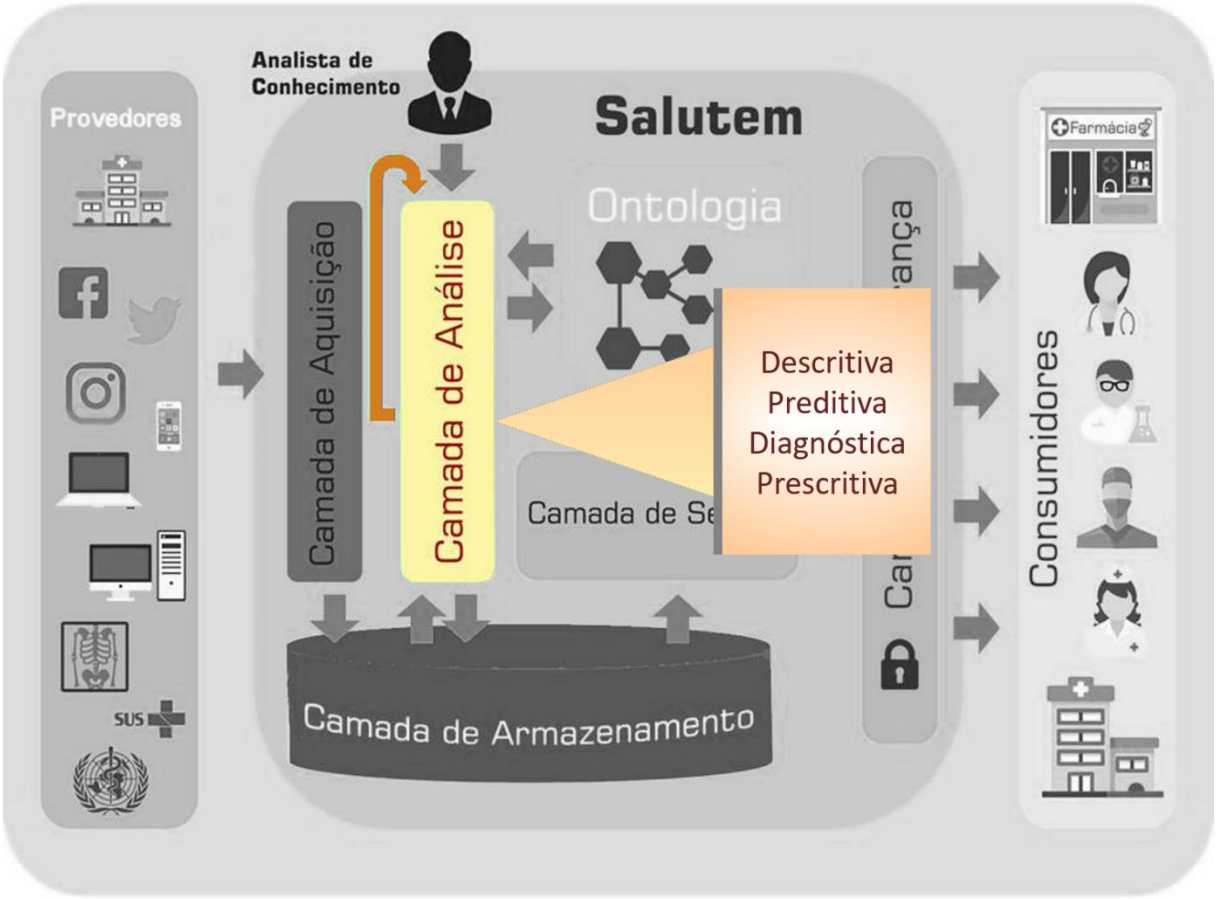
Existem também os casos em que o volume de dados adquiridos pela camada de aquisição não é extenso, e que dessa forma, seria mais dispendioso utilizar as formas de armazenamento já mencionadas. Nesse contexto, podem ser utilizados arquivos em formato csv, xls ou texto.

Um ponto que deve ser observado na camada de armazenamento é a velocidade. Como já visto, a dimensão velocidade é uma característica importante ao se trabalhar com *big data*. Dessa forma, esta camada deve ser rápida para o armazenamento, como também para a leitura dos dados nela contidos.

4.4 Camada de Análise

Na camada de análise de dados ou *data analytics*, é realizada a extração de padrões e informações anteriormente desconhecidos, úteis, válidos e ocultos de grandes conjuntos de dados (Figura 12).

Figura 12: Camada de Análise.



Como já discutido nesse trabalho, existem quatro modelos de análise de dados, a saber: preditivo, prescritivo, diagnóstico e descritivo. A depender da necessidade do que se deseja analisar, a camada de análise deve estar preparada para qualquer modelo.

Uma das formas de se realizar a análise dos dados é empregando algoritmos de aprendizado de máquina, como por exemplo, árvore de decisão, regressão linear, entre outros, ou mesmo mineração de dados.

Segundo Baldominos *et al.* (2014), o aprendizado de máquina (do inglês, *machine learning*) são técnicas que ajudam analisar grandes conjuntos de dados e extrair importantes tendências, *links*, regras, ou seja, conhecimento. O aprendizado de máquina é um ramo da inteligência artificial que permite tornar as aplicações inteligentes sem ser explicitamente

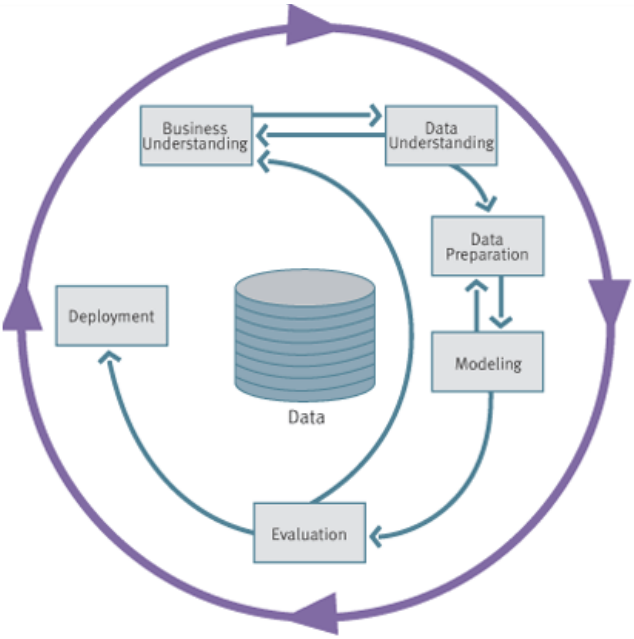
programada (PRAJAPATI, 2013). Existem basicamente três tipos de algoritmos de aprendizado de máquina, são eles:

- Supervisionado: utilizado quando possui a resposta ou rótulo, ou se é possível gerar um, a partir do conjunto de dados. Entre os algoritmos têm: regressão linear, regressão logística, *k-Nearest Neighbor* (kNN), entre outros (RAMASUBRAMANIAN; SINGH, 2017);
- Não supervisionado: usado para estruturas ocultas a partir de conjunto de dados não classificado ou rotulado. Entre os algoritmos têm: *clustering* (agrupamento), redes neurais e quantização vetorial (PRAJAPATI, 2013); e,
- Semi-supervisionado: é uma mistura do supervisionado com o não supervisionado, ou seja, utilizado quando se tem parte dos dados rotulados e outra parte não (RAMASUBRAMANIAN; SINGH, 2017).

De acordo com Al-Jaroodi e Mohamed (2016), para decidir o método apropriado a ser adotado na análise é necessário entender bem as características dos dados e o tipo de informação a ser gerada.

Para que a camada de análise obtenha sucesso é necessário seguir um processo. O processo adotado neste trabalho foi o *Cross Industry Standard Process for Data Mining* (CRISP-DM) (Figura 13).

Figura 13: Processo CRISP-DM.



Fonte: Siriweera *et al.*, 2015.

Este processo é composto por seis fases (SIRIWEERA *et al.*, 2015), a saber:

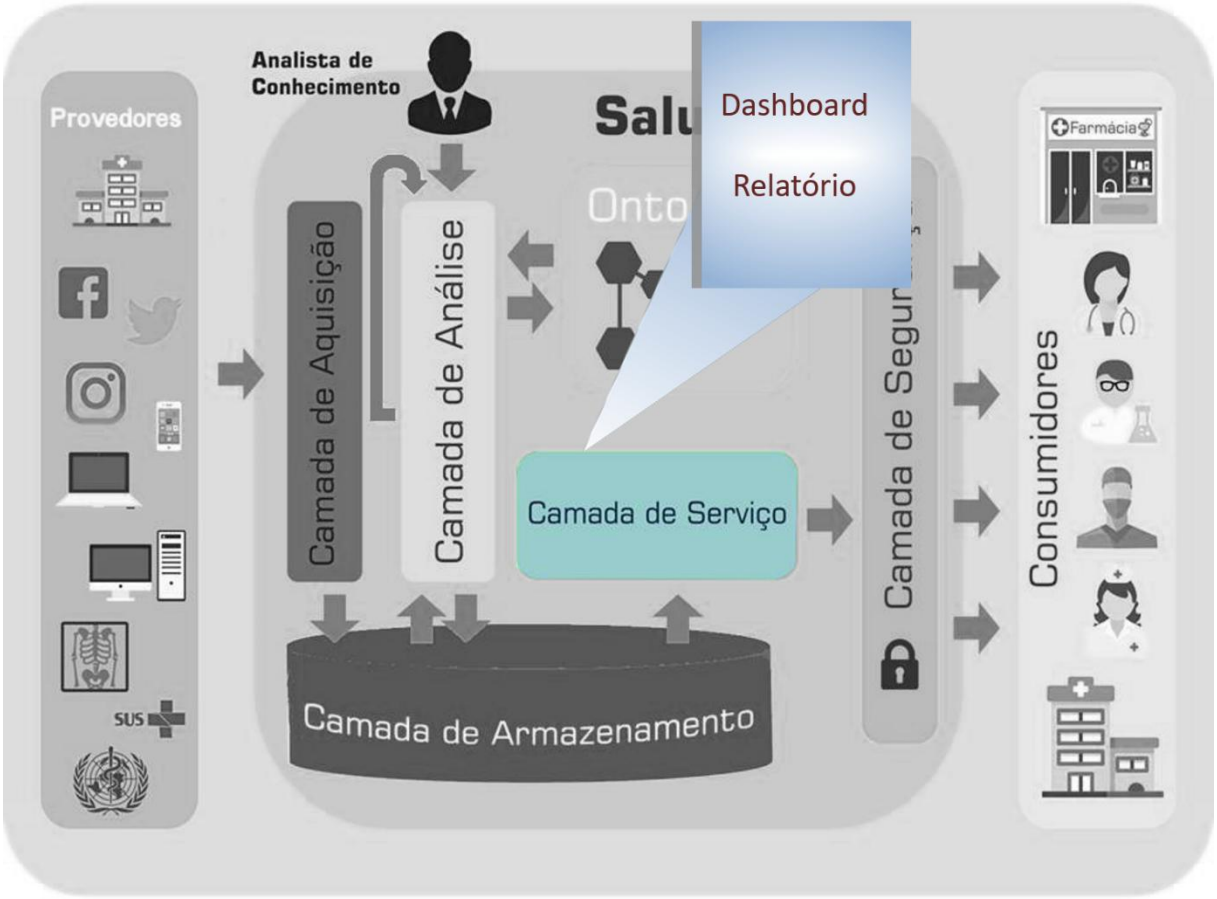
- Entendimento do negócio: é definido os objetivos e requisitos da análise de dados, bem como é estabelecido um plano a ser seguido;
- Compreensão dos dados: explora os dados até descobrir/compreender suas informações, como atributos, estrutura e contexto;
- Preparação dos dados: fase em que os dados são limpos, filtrados, integrados, caso ainda seja necessário;
- Modelagem dos dados: seleção dos dados para definição e construção do modelo;
- Avaliação do modelo: avalia o modelo com base nos objetivos definidos na primeira fase; e,
- Implantação: utilização do modelo, sendo monitorado e avaliado, em caso de manutenção.

As operações da camada de Análise, em alguns casos, podem ser necessárias algumas iterações, ou seja, a repetição de suas operações. Nesses casos, a repetição se faz necessária para avaliar outro modelo, ou mesmo realizar uma análise exploratória dos dados, e assim descobrir informações ao que se esteja buscando.

4.5 Camada de Serviço

Esta camada é responsável por disponibilizar aos consumidores, na forma de serviços, os dados analisados pela arquitetura (Figura 14).

Figura 14: Camada de Serviço.



A forma como os serviços serão disponibilizados irá depender de alguns fatores, como: análise histórica ou tempo real; forma de obtenção e visualização das informações, entre outros. De acordo com Iliinsky e Steele (2011), existem duas categorias de visualização de dados, a saber:

- Exploração: ideal quando se têm muitos dados e não tem certeza o que há neles. Normalmente os dados são traduzidos em meio visual para ajudar a identificar rapidamente suas características, incluindo curvas interessantes, linhas, tendências ou *outliers* (pontos fora da curva); e,
- Explicação: ideal quando já se sabe o que os dados têm a dizer e está tentando explicá-los a outras pessoas.

Nesse contexto, a camada de serviço tem por objetivo visualizar os dados analisados de forma exploratória ou explicativa, a depender do problema e da necessidade dos consumidores.

Dentre os tipos de serviços disponibilizados, podem ser:

- *Dashboards*: formas de visualização intuitiva e gráfica de visualizar algumas análises realizadas sobre os dados, revelando perspectivas potencialmente relevantes sobre os dados. Também pode fornecer resultados interessantes que podem ser resumidos ou exibidos para o cliente, revelando perspectivas potencialmente relevantes sobre os dados (BALDOMINOS *et al.*, 2014);
- Relatórios: é um documento que contém dados utilizados para leitura ou visualização. Podem ser simples como uma tabela de dados ou algo mais trabalhado, com totalizadores, agrupamentos, entre outros (ALEXANDER, 2016).

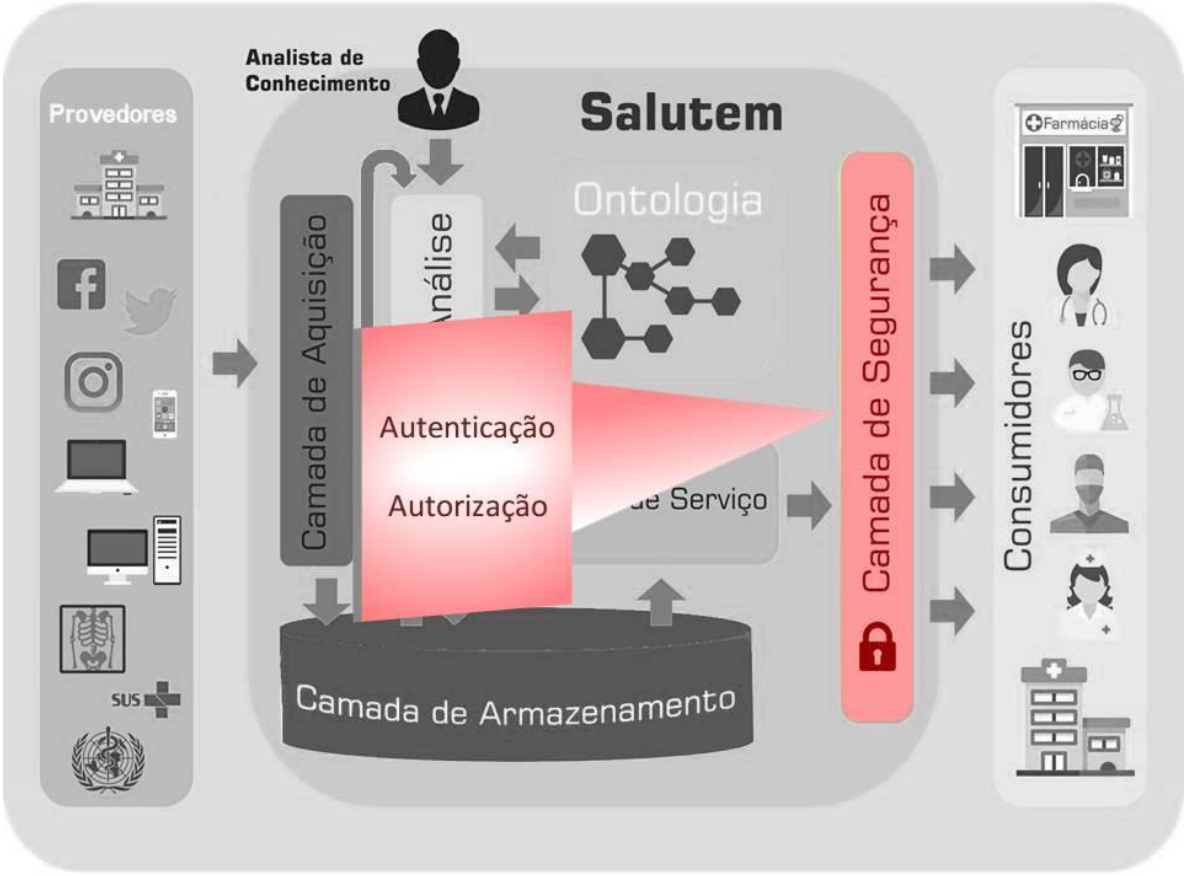
4.6 Camada de Segurança

De acordo com a Stanford (2017), existe uma tensão no que cerne aos cuidados de saúde com *big data*. Por um, os especialistas concordam que a partilha dos dados é essencial para maximizar os benefícios de *big data*. Por outro lado, as organizações de saúde tornam-se alvos de ataques de segurança e existem preocupações quando se trata de proteger a privacidade do paciente. Nesse sentido, a indústria deve resolver como aproveitar o potencial total de dados importantes, compartilhando-os para beneficiar a saúde pública, ao mesmo tempo em que deva assegurar que a informação do paciente permaneça a salvo e segura.

A ameaça à privacidade e à segurança dos dados, desde os ataques do *malware* "WannaCry" em maio de 2017, ganharam destaque e aumentou a preocupação com a segurança dos dados de paciente, visto o que ocorreu com o *British National HealthService* e empresas de dispositivos médicos dos Estados Unidos. Os especialistas concordam que conseguir um equilíbrio entre o compartilhamento e proteção dos dados médicos será um desafio. Encontrar soluções ótimas exigirá que os setores público e privado, pagadores e pacientes colaborem.

Nesse sentido, a camada de segurança tem por objetivo assegurar a garantia dos dados e privacidade das informações de profissionais e pacientes (Figura 15).

Figura 15: Camada de Segurança.



Esta camada possui dois componentes básicos, um para autenticação e outro para autorização. A autenticação serve para que o acesso aos serviços disponibilizados seja verificado, ou seja, se o usuário é quem diz ser. E a autorização permite elencar quais serviços e/ou dados determinados usuários podem ou não ter acesso. Por exemplo, um diagnóstico de AIDS de um paciente deve ser restrito apenas a médicos.

Segundo Rao, Suma e Sunitha (2015), a segurança e a privacidade dos dados para saúde relacionam-se não somente com as expectativas sobre a privacidade em geral, mas também com as normas que envolvem prática profissional, privilégio, autonomia, paternalismo, comunicação protegida e dever de confidencialidade, bem como requisitos para a coleta, disseminação e retenção de dados.

De acordo com Soceanu *et al.* (2015), a autenticação pode ser feita por diversas maneiras, a saber: usuário e senha (ou PIN), senha de *hash*, chaves compartilhadas, senhas

descartáveis (válidas somente para uma autenticação), certificados digitais, credenciais biométricas, entre outras.

Com relação à autorização, Gao *et al.* (2015) definem como um processo para verificar se um pedido satisfaz a uma regra, e assim determinar o efeito de autorização de tal regra. De forma simples, o efeito de autorização a uma regra é permitido ou negado. Assim funciona a autorização na camada de segurança. O usuário ao requisitar acesso a determinado serviço, pertencente à camada de serviço, será verificado se o consumidor tem a autorização para tal serviço, sendo assim liberada a visualização, ou negada, apresentando então uma mensagem informativa.

Nesse contexto, quem decide normalmente quais consumidores possuem acesso aos serviços é o dono do serviço. Ou seja, tem-se um papel de dono que concederá, a depender do nível de acesso, terá permissão a determinado serviço ou não. Além disso, existe o perfil administrador, que controla toda e qualquer permissão inicial a um serviço, depois ficando a cargo do dono do serviço a atualização das permissões (THURAISINGHAM, 2015).

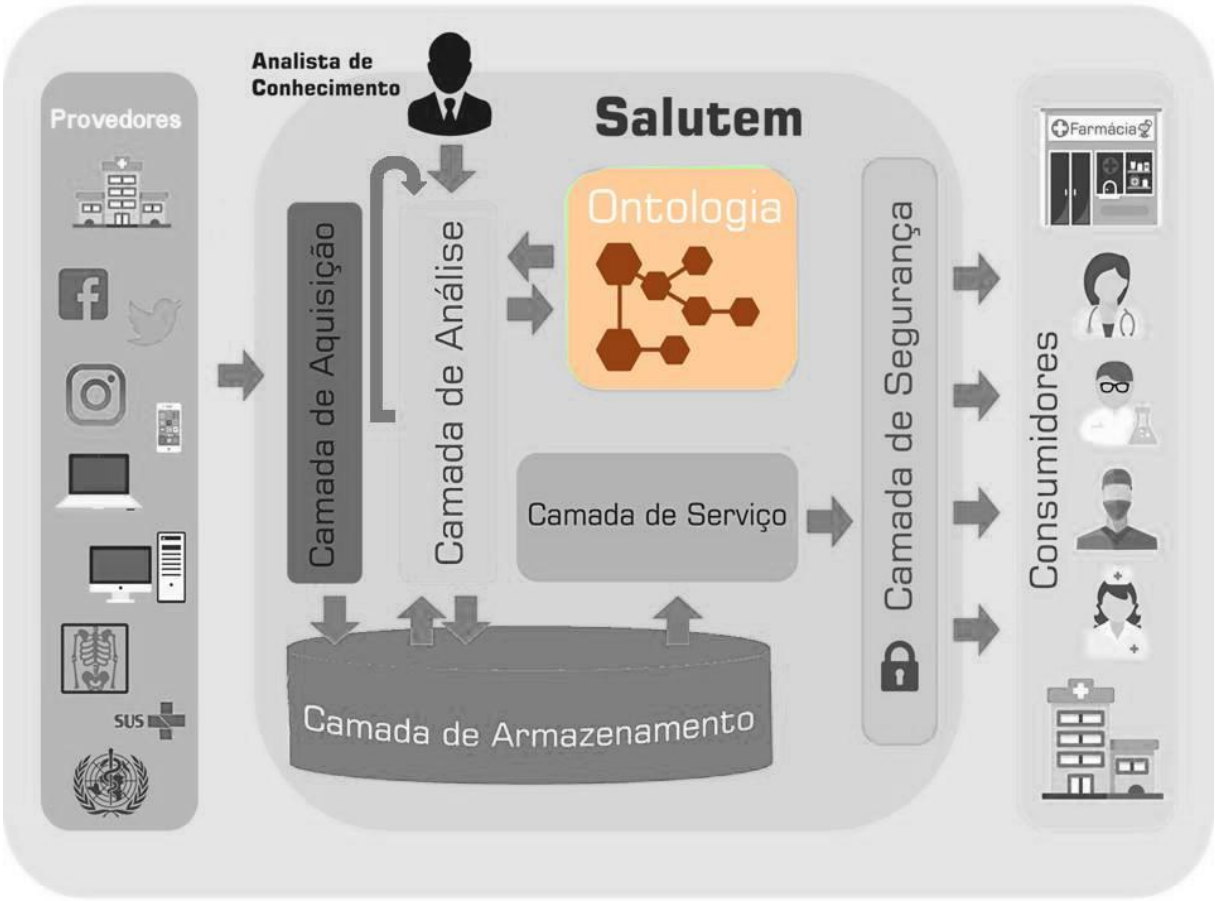
Essa é uma camada fundamental da arquitetura *Salutem*. De acordo com Andreu-Perez *et al.* (2015), no domínio da saúde, a privacidade dos dados é de extrema importância, conforme regulamentado pelas leis em diversos países.

4.7 Componente Ontologia

Segundo Blobel (2010), o desafio de qualquer arquitetura de sistema é sua conceituação, isto é, a descrição do sistema com seus componentes e todas as suas inter-relações lógicas. Isto tem de ser feito tanto a partir da perspectiva global do sistema como da perspectiva dos diferentes domínios, incluindo ainda a perspectiva do processo de desenvolvimento. Há uma ciência especializada que lida com esse desafio: a ontologia.

O uso da ontologia pela arquitetura *Salutem* tem por objetivo a representação do conhecimento global que é necessário para compartilhar um dicionário comum e descrever as relações entre as entidades e objetos que são parte do modelo (Figura 16).

Figura 16: Ontologia.



Conceitos da ontologia são usados no lado do cliente para descrever uma representação da realidade como ela é percebida pelo usuário. No *backend*, ou lado da tecnologia, a ontologia é usada para anotar recursos digitais como ponto de interesses, conteúdos, aplicações. Dessa forma é garantido que os dados obtidos estejam em harmonia com o contexto da saúde (AMATO; DI MARTINO; VENTICINQUE, 2015).

Qualquer ontologia pode ser utilizada pela SALUTEM, devendo apenas está de acordo com o domínio do problema que esteja sendo tratado. Por exemplo, no caso do diagnóstico de câncer por meio de imagens de ressonância magnética, a ontologia a ser utilizada deve conter o vocabulário comum desse contexto (ZIMMERMANN *et al.*, 2013).

4.8 Atores

Na definição da arquitetura *Salutem*, os atores representam os consumidores dos serviços. Alguns atores, a título de exemplo, são:

- Médicos: com objetivo de apoiar e facilitar a tomada decisão em relação aos cuidados dos pacientes, sendo suas decisões de forma eficaz e eficiente.
- Farmacêuticos: para compreender melhor doenças e direcionar os fármacos eficazes, ou seja, melhorando o tratamento dos pacientes.
- Enfermeiros: com intuito de melhorar a assistência dada aos pacientes e médicos.
- Pesquisadores: para ajudar em pesquisas e trabalhos no âmbito da saúde.
- Gestores: tomadores de decisão, como administradores, superintendentes, entre outros.
- Donos do serviço: são os atores que são os donos dos serviços providos por determinado problema tratado pela arquitetura.

Para uma melhor compreensão dos possíveis atores a serem envolvidos pela arquitetura, ver Figura 17.

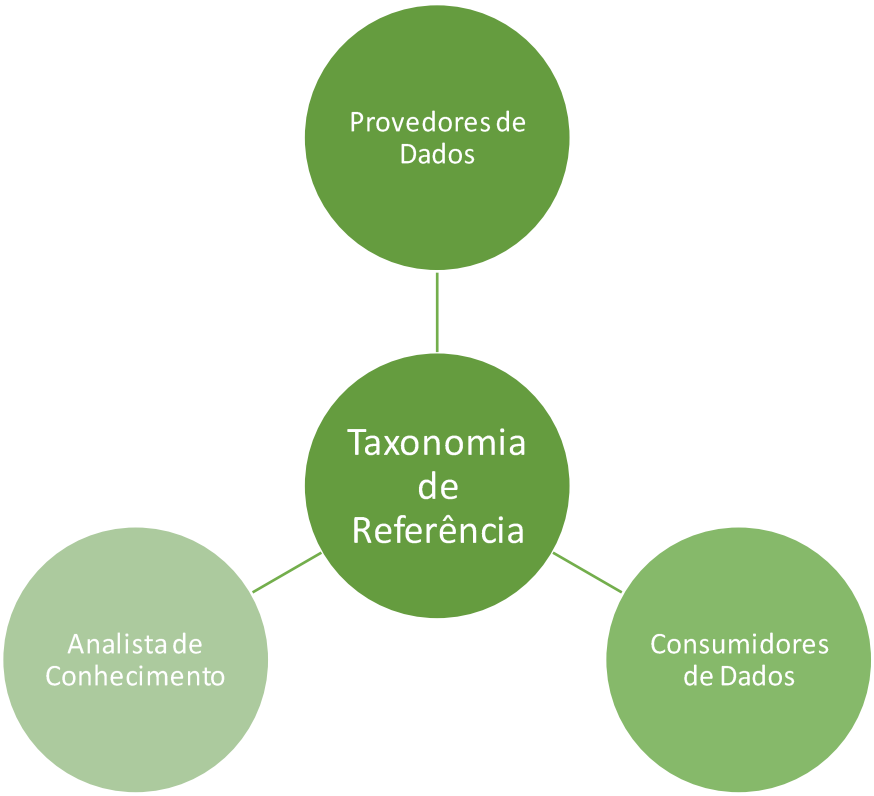
Figura 17: Exemplo de atores da *Salutem*.

Os atores vão ser determinados pelo escopo da análise em tela, ou seja, quais papéis são necessários e de interesse na análise realizada utilizando a arquitetura.

4.9 Taxonomia de referência da arquitetura *Salutem*

De acordo com Grady e Chang (2017), uma taxonomia de referência é importante para manter um consenso na definição de termos. Assim, foi desenvolvida uma taxonomia de referência da arquitetura *Salutem* (Figura 18).

Figura 18: Taxonomia de referência da arquitetura *Salutem*.



A taxonomia é composta por: provedores de dados, consumidores de dados e o analista de conhecimento.

Os provedores de dados disponibilizam os dados para si ou para outros, sendo que para isso é criado abstrações para os vários tipos de fontes de dados (dados brutos ou dados previamente transformados) e os concede através de diferentes interfaces funcionais. Os atores do provedor de dados podem ser qualquer um, desde um sensor, até um dado inserido manualmente (XINHUA *et al.*, 2013).

Os consumidores de dados recebem o valor de saída dispostas através de serviços. Em alguns casos, o consumidor pode utilizar o mesmo tipo de interfaces funcionais que os provedores de dados (GRADY e CHANG, 2017).

O papel analista de conhecimento define um profissional ou pessoa com sólidos conhecimentos da área da saúde, que possua expertise necessária para compreender e

oportunizar informações sobre a demanda em análise. Para a taxonomia um analista de conhecimento pode ser um médico, farmacêutico, biomédico, enfermeiro, estatístico, economista, entre outros.

4.10 Considerações Finais do Capítulo

Este capítulo apresentou a *Salutem,* uma arquitetura de *big data* para a saúde, que tem como premissa atender as características como coleta, armazenamento, análise, serviços, segurança e semântica de dados. Como visto, a arquitetura *Salutem* está estruturada em camadas, componentes e atores.

A arquitetura *Salutem* tenta cobrir a ausência de arquiteturas que incorporem as características minimamente necessárias para atender aos desafios no âmbito de *big data* na saúde. O intuito da *Salutem* é ser uma arquitetura viável e de simples implantação, de maneira a atender organizações de saúde que necessitam trabalhar com grandes volumes de dados e que, no entanto, encontram dificuldades seja por falta de profissionais especializados seja por falta de conhecimento.

No próximo capítulo, será apresentado um estudo de caso, no qual foi adotada a arquitetura *Salutem* na construção do Hemo *Data*, para validar a arquitetura proposta.

CAPÍTULO 6

ESTUDO DE CASO: HEMO *DATA*

Este capítulo descreve a utilização da arquitetura *Salutem* adotada para construção do Hemo *Data,* uma solução de serviços com informações sobre exames de hemograma realizados no Hospital Universitário da Universidade Federal de Sergipe (HU-UFS), no período de 2004 a 2016.

5.1 O Hospital Universitário e o Laboratório de Análises Clínicas

O HU-UFS é um hospital-escola vinculado à Universidade Federal de Sergipe desde 1984, o qual presta assistência médico-hospitalar de média e alta complexidade e serve de base para as atividades acadêmicas e científicas dos cursos de saúde. O HU-UFS também atua em programas de assistência e inclusão social por meio de parcerias com órgãos públicos, bem como desenvolve atividades de natureza preventiva e extensiva através de programas nacionais de saúde e educação oferecidos à população sergipana. (HU-UFS, 2012).

O hospital não realiza atendimentos particulares ou através de planos de saúde. Na verdade, o HU-UFS presta serviços para o Sistema Único de Saúde (SUS), sendo referência em atendimentos de média e alta complexidade no ambulatório e em relação a exames complementares, diagnóstico e internação. Atualmente o HU possui 105 leitos, 68 consultórios ambulatoriais, realiza mensalmente em torno de 13 mil consultas ambulatoriais, 53 mil exames e 280 cirurgias. Esse volume de atendimento faz com que o HU-UFS tenha hoje armazenado 21 terabytes aproximadamente, entre os exames de imagens, métodos gráficos, patologia, análises clínicas, entre outros, com perspectiva de aumento desse volume com a oferta de novos serviços nos próximos anos.

A Divisão de Apoio Diagnóstico e Terapêutico é responsável pela gestão da oferta de exames no HU, sendo no Laboratório de Análises Clínicas (LABCLIN) a realização dos exames de sangue (por exemplo, hemogramas), urina, coprológico, suor, entre outros. Entre 2004 e 2016 foram realizados no LABCLIN cerca de 3.000.000 de exames.

Destes exames o hemograma é um dos testes mais solicitados em hematologia laboratorial em todo mundo, inclusive no HU-UFS (Brereton *et al.*, 2016). Segundo Failace (2003) e Grotto (2009), o hemograma avalia os elementos celulares do sangue

qualitativamente e quantitativamente, sendo o exame complementar mais requerido e parte de todo *check-up* de saúde. O hemograma vem evoluindo desde os primeiros dias da medicina laboratorial, com a contagem manual das células através do microscópio, até os analisadores automatizados de última geração.

Para Maji, Mandal, Ganguly e Saha (2015), o hemograma é um exame requisitado por médicos por fornecer informações sobre as células do sangue do paciente. Basicamente as células são divididas em três tipos, a saber: glóbulos brancos ou leucócitos, glóbulos vermelhos (eritrócitos) e plaquetas ou trombócitos. O quantitativo anormalmente altos ou baixos de determinadas células podem indicam muitas formas de doenças.

No HU-UFS, não diferentemente, o hemograma é o exame mais processado. No entanto, a falta de soluções para manipular tal volume de dados que cresce diariamente, fez com que pouco se explore os dados armazenados, seja do ponto de vista clínico quanto do ponto de vista acadêmico/científico.

Neste contexto, uma demanda recorrente é a descoberta de informações e conhecimento quanto aos exames realizados no hospital. Profissionais de saúde, professores e pesquisadores solicitam acesso a dados para pesquisas, apoio a gestão, avaliação epidemiológica, entre outros. O HU-UFS necessita desenvolver e implantar soluções baseadas em grandes volumes de dados, apesar de ter os dados e alta demanda por serviços sobre os dados.

Sendo assim, este estudo de caso se propõe a desenvolver uma solução para tratar os dados dos hemogramas e ofertar serviços que poderão ser disponibilizados a partir do conhecimento gerado.

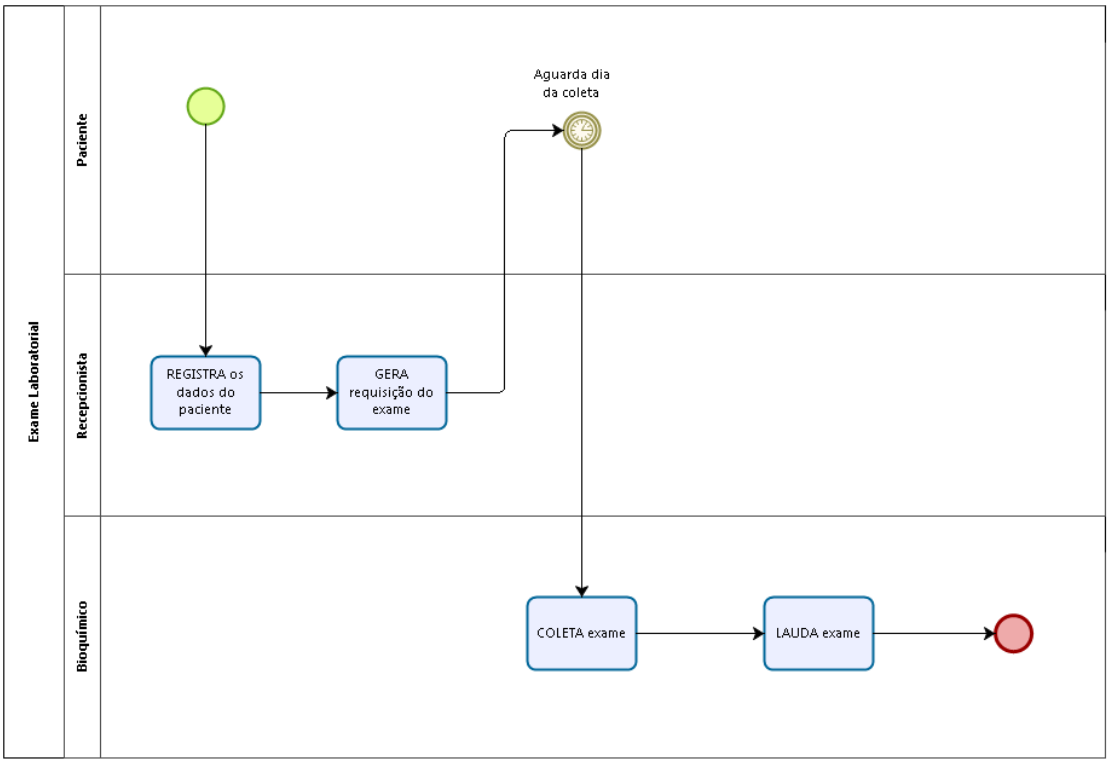
Para tanto, se propõe o desenvolvimento do protótipo de serviço Hemo *Data*. Na construção do protótipo foi adotada a arquitetura *Salutem*. Foram utilizados os resultados dos hemogramas realizados de 2004 a 2016 no HU-UFS, haja vista que possui uma quantidade expressiva em relação aos demais exames e por ser comumente requisitado pelos médicos, independente da especialidade.

Espera-se com esse estudo validar a *Salutem* como um guia para desenvolvimento de soluções que tratem grandes volumes de dados e ofertem serviços aos consumidores. Nas próximas seções, são descritas as atividades realizadas conforme propõe a arquitetura.

5.2 Processo de Realização de Hemograma

Para melhor conhecer a demanda foi mapeado o processo de realização de exames laboratóriais realizados no HU-UFS, conforme processo mapeado em *Business Process Model and Notation* (BPMN) apresentado na Figura 19.

Figura 19: Processo dos exames laboratoriais.



O processo é simples, os atendentes registram os dados do paciente, ou seja, prontuário médico com as informações pertinentes do paciente, como: nome, nome da mãe, data de nascimento, entre outros. Após é realizado cadastro da solicitação do exame a ser realizado pelo paciente, nesse caso hemograma. Então o paciente é agendado para realizar (coletar) o exame numa data posterior. Depois da coleta, os exames são laudados por profissionais habilitados para então ser disponibilizado o resultado ao paciente.

5.3 Serviços Requisitados

Os resultados dos exames hemogramas no HU-UFS são processados por equipamentos, no entanto, segundo Falaice (2003), apesar da generalização da tecnologia eletrônica, ainda há considerável variação interlaboratorial, pois o hemograma depende da qualidade do equipamento, do grau de especialização do pessoal técnico, da filosofia de trabalho do laboratório e das tradições locais. O autor também aponta que pode variar até por decisões político-econômicas de acordo com a procedência das requisições, e para interpretá-los há necessidade do conhecimento da tecnologia de todos os parâmetros fornecidos e da maneira de expressar os resultados.

Nesse cenário foram elencados alguns serviços que pudessem ser gerados a partir da análise dos dados de hemogramas, a saber: a proporção de casos de anemia pela população de cada município de Sergipe; e, descobrir para as células hemoglobina, eritrócitos, VCM e hematócritos, as seguintes informações:

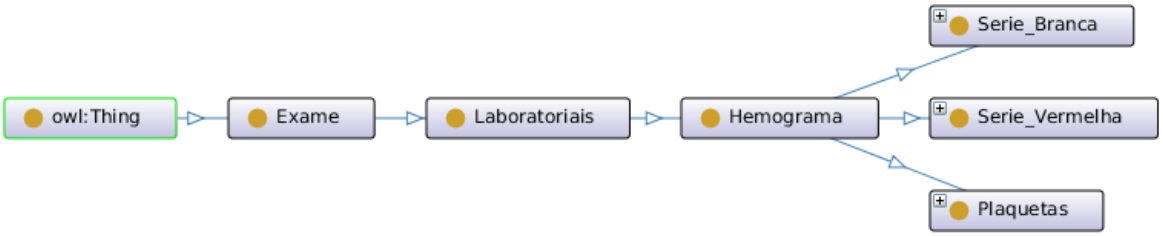
- Os valores de referência para os pacientes maiores de 18 anos tanto masculinos como femininos.
- O quantitativo de pacientes com valores normais e valores fora da referência, nas seguintes situações:
    - Por ano coletado;
    - Por sexo;
    - Por procedência do paciente, ou seja, paciente internado ou externo;
    - Por cor de pele.

Outros serviços podem ser disponibilizados a partir do mesmo conjunto de dados. Para este trabalho, foi limitado o número de serviços.

5.4 Ontologia ontoHemograma

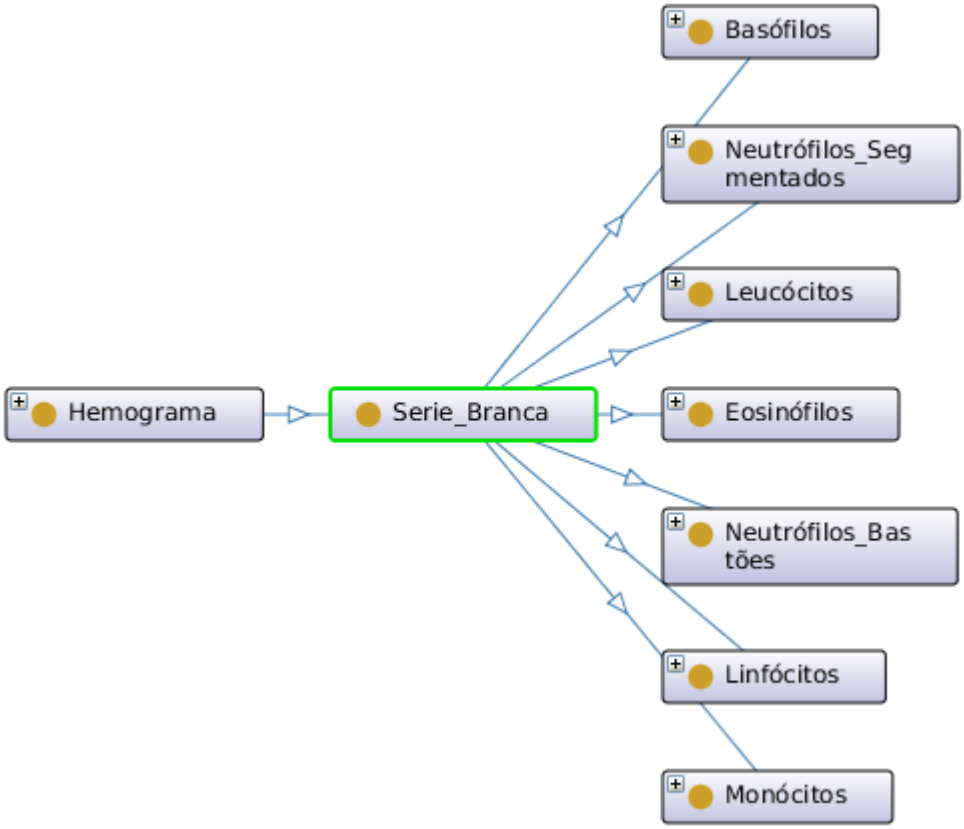
Uma das características da *Salutem* é a utilização de ontologias para manter uma unidade semântica com relação aos dados. Para tanto, nessa validação foi construída uma ontologia para vislumbrar o conhecimento relativo ao hemograma. A ontologia criada, denominada ontoHemograma, contém todos as características que contempla o exame hemograma (Figura 20).

Figura 20: Ontologia ontoHemograma.



Como visto na Figura 20, são encontradas no hemograma três tipos de células: a branca, a vermelha e plaquetas. A série branca é subdivida em: basófilos, neutrófilos segmentados, leucócitos, eosinófilos, neutrófilos bastões, linfócitos e monócitos (Figura 21).

Figura 21: Ontologia ontoHemograma série branca.



Segundo Bain (2008), os basófilos têm um papel importante nas reações de hipersensibilidade imediata, respostas alérgicas e inflamatórias, e no controle de infecções parasitárias. Tem como valor de referência 30 por microlitro de sangue (JUNQUEIRA; CARNEIRO, 2008).

Os leucócitos possuem a função de defesa celular do organismo. O número de leucócitos num adulto normal é de 6 a 10 mil mm³ (JUNQUEIRA; CARNEIRO, 2008) (VIVAS, 2008).

De acordo com Bain (2008), os neutrófilos têm a mesma função dos leucócitos e são divididos em neutrófilos bastões e neutrófilos segmentados. O primeiro é formado por células mais jovens, enquanto o segundo, por células maduras.

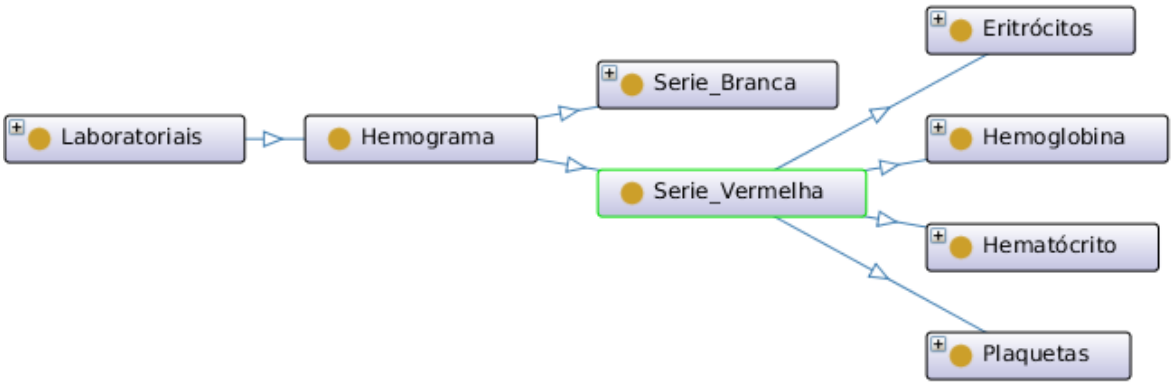
Para Junqueira e Carneiro (2008), os eosinófilos também têm a função igual aos neutrófilos, tendo como diferencial, que ajudam a controlar infecções parasitárias. Tem a quantidade aproximada de 150 por microlitro de sangue (JUNQUEIRA; CARNEIRO, 2008).

Já os linfócitos, de acordo com Vivas (2008), tem a função importante nas respostas imunes específicas do corpo, incluindo produção de anticorpos. A quantidade de linfócitos no sangue é de aproximadamente de 2400 por microlitro (JUNQUEIRA; CARNEIRO, 2008).

E os monócitos, segundo Bain (2008), são responsáveis pela defesa do organismo contra infecções bacterianas e fúngicas. Sua quantidade aproximada por microlitro de sangue é de 150 (JUNQUEIRA; CARNEIRO, 2008).

Já a série vermelha subdivide-se em: eritrócitos, hemoglobina e hematócrito (Figura 22). E finalmente as plaquetas.

Figura 22: Ontologia ontoHemograma série vermelha.



Para Junqueira e Carneiro (2008), os eritrócitos, ou também chamadas de hemácias, é responsável pelo transporte do oxigênio no sangue. Seus valores referências são 3,9 a 5,5 por mm³ nas mulheres e de 4,1 a 6 mm³ nos homens.

Ainda de acordo com Junqueira e Carneiro (2008), as plaquetas têm cerca de 200 a 400 mil por microlitro de sangue e é responsável pela coagulação do sangue.

A ontoHemograma além de apresentar a subdivisão dos tipos de células, descreve as características de cada uma, bem como indica os valores de referência. Os valores de referência são os valores considerados limites para cada tipo de célula. Ou seja, os valores obtidos a partir da análise do hemograma de cada tipo de célula não devem ser menores ou maiores, a depender da referência. Dessa forma, se tem um parâmetro a ser considerado ao analisar os resultados.

A ontoHemograma, foi importante para manter a unicidade de intepretação e semântica dos dados por parte do Analista de Conhecimento e também dos consumidores de serviços.

5.5 Provedores de dados

Para este estudo de caso, utilizou-se apenas a base de dados proveniente do sistema Medlynx, sistema *Web* desenvolvido em PHP e com banco de dados Postgres.

O sistema Medlynx, é um *Enterprise Resource Planning* (ERP), que contemplam módulos como: exames, internação, financeiro, entre outros. O Medlynx foi doado em 2004 ao HU-UFS, sendo desde então utilizado nas operações diárias do hospital. No entanto, em torno de 2008, o contrato de manutenção evolutiva e corretiva não foi renovado desde então, tornando o sistema estagnado.

Outras fontes de dados não foram consideradas, haja vista os serviços elencados não necessitavam.

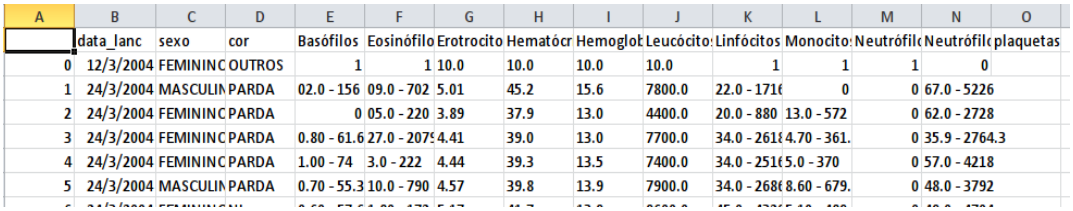
5.6 Camada de Aquisição

Nas operações desta camada foi utilizada a linguagem Python, que é uma linguagem desenvolvida sob a licença *open source* aprovada pela *Open Source Iniciative* (OSI), o que a torna livre para ser utilizada e distribuída, mesmo para uso comercial. O Python por padrão suporta a extensão de novas funcionalidades através de bibliotecas que podem ser instaladas de acordo com a necessidade.

Para atender aos preceitos da camada de aquisição, prevista na arquitetura *Salutem*, foi utilizada a biblioteca psycopg2, responsável por acessar a base de dados Postgres do sistema Medlynx e adquirir os dados dos hemogramas. Com os dados capturados, foi utilizada a biblioteca Pandas para manipular os dados no formato *Comma-Separated Values* (CSV) com os dados dos hemogramas do HU-UFS para serem trabalhados em memória. De acordo com

McKinney (2010), Pandas é uma biblioteca do Python para dados estruturados e ferramentas estatísticas. Uma das funções da biblioteca é a obtenção de dados no formato CSV para que estes possam ser manipulados em memória. A Figura 23 apresenta um extrato do arquivo CSV obtido dos hemogramas.

Figura 23: Extrato do arquivo CSV dos hemogramas.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
		data_lanc	sexo	cor	Basófilos	Eosinófilo	Eritrocito	Hematócr	Hemoglob	Leucócito	Linfócitos	Monocito	Neutrófil	Neutrófil	plaquetas
0		12/3/2004	FEMININ	OUTROS	1	1	10.0	10.0	10.0	10.0	1	1	1	0	
1		24/3/2004	MASCULIN	PARDA	02.0 - 156	09.0 - 702	5.01	45.2	15.6	7800.0	22.0 - 171	0	0	67.0 - 5226	
2		24/3/2004	FEMININ	PARDA	0	05.0 - 220	3.89	37.9	13.0	4400.0	20.0 - 880	13.0 - 572	0	62.0 - 2728	
3		24/3/2004	FEMININ	PARDA	0.80 - 61.6	27.0 - 207	4.41	39.0	13.0	7700.0	34.0 - 261	4.70 - 361.	0	35.9 - 2764.3	
4		24/3/2004	FEMININ	PARDA	1.00 - 74	3.0 - 222	4.44	39.3	13.5	7400.0	34.0 - 251	5.0 - 370	0	57.0 - 4218	
5		24/3/2004	MASCULIN	PARDA	0.70 - 55.3	10.0 - 790	4.57	39.8	13.9	7900.0	34.0 - 268	8.60 - 679.	0	48.0 - 3792	

Com os dados em memória realizou-se o processo de ETL. Primeiramente foram retirados todos os dados vazios (nulos) para que não influenciasse na análise dos mesmos. Além disso, os exames que tinham sido laudados erroneamente foram descartados, ou seja, erro humano. Este problema ocorreu frequentemente, haja vista que os resultados dos exames são digitados diretamente no Medlynx e não interfaciados com os equipamentos que realizam as análises.

Alguns resultados de exames foram cadastrados numa mesma variável, sendo tanto um valor relativo (medido em percentual) quanto um valor absoluto (normalmente metro cúbico), sendo estes resultados subdivididos.

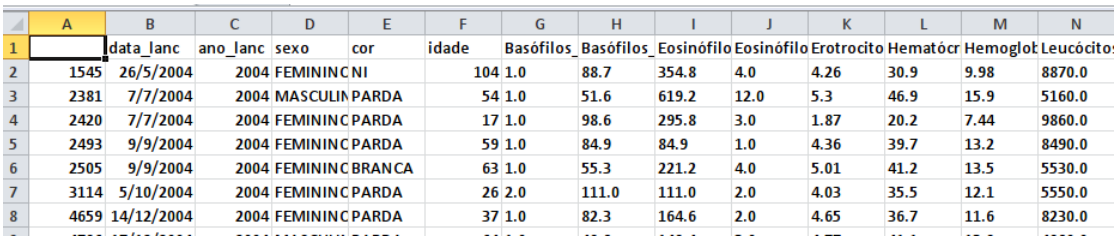
Outras variáveis também foram criadas a partir das existentes, como é o caso da idade, que armazena a idade do paciente no dia do exame e o ano da ocorrência do exame a partir da data do lançamento do exame. Essas variáveis foram criadas para obter mais detalhes dos exames realizados.

As operações executadas nesta camada demandaram tempo e esforço, tanto pela qualidade dos dados, quanto pela dificuldade na definição dos serviços a serem ofertados.

5.7 Camada de Armazenamento

Após a realização das operações previstas da camada de aquisição, o resultado foi armazenado em um novo arquivo, também no formato CSV, na camada de armazenamento. Um extrato do resultado é apresentado na Figura 24.

Figura 24: Extrato do arquivo CSV da camada de armazenamento.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1		data_lanc	ano_lanc	sexo	cor	idade	Basófilos_	Basófilos_	Eosinófilo	Eosinófilo	Erotrocito	Hematócr	Hemoglol	Leucócito
2	1545	26/5/2004	2004	FEMININO	NI	104	1.0	88.7	354.8	4.0	4.26	30.9	9.98	8870.0
3	2381	7/7/2004	2004	MASCULIN	PARDA	54	1.0	51.6	619.2	12.0	5.3	46.9	15.9	5160.0
4	2420	7/7/2004	2004	FEMININO	PARDA	17	1.0	98.6	295.8	3.0	1.87	20.2	7.44	9860.0
5	2493	9/9/2004	2004	FEMININO	PARDA	59	1.0	84.9	84.9	1.0	4.36	39.7	13.2	8490.0
6	2505	9/9/2004	2004	FEMININO	BRANCA	63	1.0	55.3	221.2	4.0	5.01	41.2	13.5	5530.0
7	3114	5/10/2004	2004	FEMININO	PARDA	26	2.0	111.0	111.0	2.0	4.03	35.5	12.1	5550.0
8	4659	14/12/2004	2004	FEMININO	PARDA	37	1.0	82.3	164.6	2.0	4.65	36.7	11.6	8230.0

O resultado não é extenso, o que não impactou na velocidade tanto no armazenamento quanto na leitura dos dados contidos no arquivo.

5.8 Camada de Análise

Na análise dos hemogramas foi utilizado o modelo descritivo inicialmente, ou seja, utiliza a estatística ou mineração de dados para verificar como estão os dados e que informações podem ser obtidas. Nesse sentido, foram analisados descritivamente os dados para extrair informações sob a perspectiva dos analistas de conhecimento, de maneira a atender os requisitos demandados. Bom frisar que a análise foi realizada com base no componente ontologia, e nesse estudo de caso a ontoHemograma.

Primeiramente foi necessário entender o negócio e o contexto, com a definição de objetivos e requisitos da análise almejada. Em seguida, partiu-se para a fase de compreensão dos dados obtidos. Nessa fase foram explorados os dados para descobrir e compreender através das variáveis existentes informações e conhecimentos que pudessem ser ofertados na forma de serviços.

Diversas operações foram realizadas pelo analista de conhecimento para descobrir informações, sendo executado diversas vezes o algoritmo de *clustering* K-means. No entanto, a exploração desse recurso, dado o nível de maturidade, não se apresentou resultados de interesse dos consumidores.

Na fase seguinte, houve uma preparação dos dados, retirando-se variáveis não necessárias para a análise em questão. Posteriormente, ocorreu a fase para seleção dos dados para construção do modelo. No caso do hemograma, outras variáveis foram descartadas, considerando a questão da privacidade dos pacientes, no caso o número do prontuário e o nome do paciente. Também foi necessária a criação de outras variáveis derivadas das existentes, para facilitar a análise, como por exemplo, a existência ou não de anemia.

Em seguida, realizou-se uma análise preditiva com o intuito de prever casos de anemia. Para tanto, foram utilizados três algoritmos diferentes para avaliar a acurácia, ou seja, a taxa de predições corretas ou incorretas para um determinado modelo. Foram utilizados os algoritmos *random forest*, *naive bayes* e regressão logística, com as acurácias 90,09%, 77,89% e 84,35%, respectivamente.

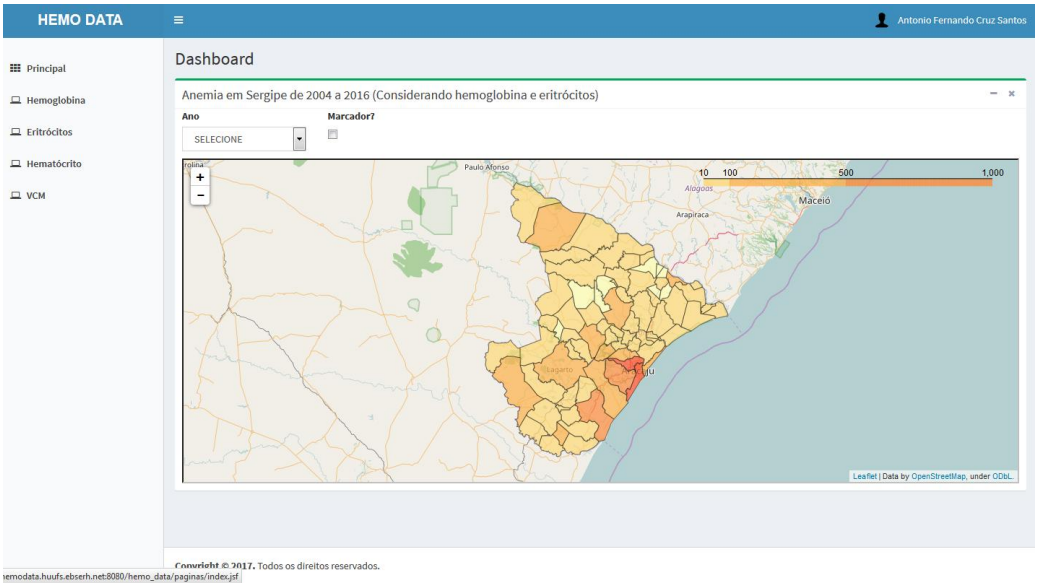
Nas operações da camada de análise foram utilizadas as bibliotecas Pandas, matplotlib e scikit-learn, que são bibliotecas *open source* da linguagem Python. O scikit-learn é uma biblioteca simples e eficiente que contém ferramentas para mineração e análise de dados.

Como já mencionado, a ontoHemograma foi utilizada pela camada de análise para manter a mesma interpretação dos dados e informações manipulados.

5.9 Camada de Serviços

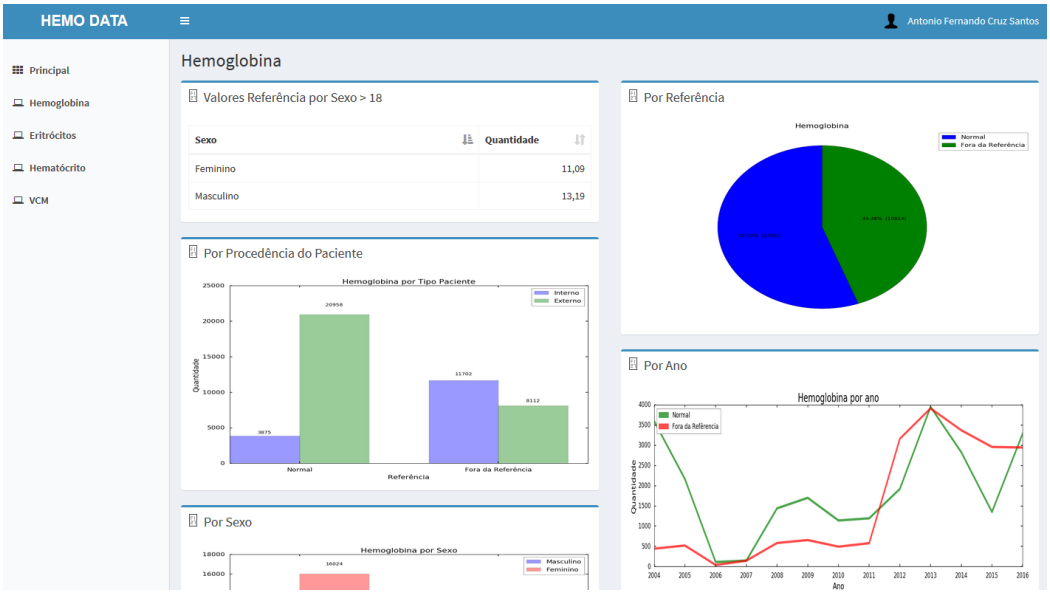
Os serviços demandados são disponibilizados na camada de serviços, conforme prevê a arquitetura adotada. A Figura 25 apresenta a página principal do Hemo *Data*.

Figura 25: Página principal do Hemo *Data*.



Na página principal do protótipo é apresentado o mapa de calor de Sergipe com a quantidade de casos de anemia identificados por município. Também foram disponibilizadas informações (dados estatísticos) sobre hemoglobina, eritrócitos, hematócrito e VCM, geradas a partir dos resultados das operações da camada de análise. A Figura 26 apresenta um exemplo com informações referente à hemoglobina.

Figura 26: Informações hemoglobina.



Entre as informações disponibilizadas, tem-se:

- Os valores de referência por sexo para os pacientes maiores que 18.
- O percentual de pacientes com valores de exame normais e fora da referência.
- O quantitativo de pacientes com valores normais e fora da referência por procedência, ou seja, interno ou externo.
- O quantitativo de pacientes com valores normais e fora da referência por procedência por ano.
- O quantitativo de pacientes com valores normais e fora da referência por procedência por sexo.
- O quantitativo de pacientes com valores normais e fora da referência por procedência por cor de pele.

5.10 Camada de Segurança e Consumidores

O Setor de Gestão de Processos e Tecnologia da Informação (SGPTI) do HU-UFS desenvolveu uma solução com o intuito de controlar o acesso aos diversos de sistemas utilizados no hospital. Nesse sentido foi utilizada tal solução, denominada Controle de Acesso (CA), cujo objetivo é autenticar e autorizar o acesso às aplicações (Figura 27).

Figura 27: Solução de Controle de Acesso (CA).

Com o CA foi possível cadastrar este protótipo como uma solução e também introduzir perfis de acesso. Dessa forma, o dono do serviço ou administrador concederá acesso a um consumidor do serviço através de perfis. Na solução CA, é possível cadastrar vários perfis e associar os mesmos aos usuários, ou mesmo, retirar permissão quando necessário.

Para este estudo de caso foi criado apenas o perfil consumidor e associado a um usuário.

O consumo dos serviços ofertados foi liberado inicialmente para que profissionais validassem a aplicabilidade dos resultados, bem como a identificação de novos serviços. Entre os profissionais estão: médicos, bioquímicos e pesquisadores.

5.11 Análise da aplicação da *Salutem* no estudo de caso

A demanda do HU-UFS em extrair informações e conhecimentos dos grandes volumes de dados é crescente. A proposição de uma solução para ofertar serviços com informações com alto valor agregado motivou o desenvolvimento deste estudo de caso.

A seleção dos hemogramas como foco deste estudo se mostrou acertada, uma vez que são os exames mais solicitados pelos especialistas e, portanto, toda e qualquer instituição de saúde possui milhões de resultados armazenados e pouco explorado seu valor.

Uma vez definida o provedor de dados, as operações da camada de aquisição são fundamentais. As operações de extração e limpeza dos dados foram trabalhosas e demoradas, visto que a base de dados utilizada é proveniente de um sistema legado que ao longo de 12

anos acumulou inconsistências e mudanças significativas em seus processos de execução. Isso reforça o papel da camada de aquisição onde estão contempladas as operações de ETL.

A posteriori, foi realizada a análise dos dados, sustentada pela ontoHemograma. A inclusão do componente ontologia se mostrou essencial, já que a interpretação unívoca dos dados, informações e conhecimentos, reflete diretamente nos serviços ofertados, assim como, no tempo de execução das operações da camada de análise.

Outro aspecto a ressaltar é o papel do analista de conhecimento, previsto na *Salutem*. Durante o desenvolvimento deste estudo de caso, a participação de médicos e bioquímicos foi determinante para a construção da solução, confirmando assim a indispensabilidade em soluções de *big data* para área da saúde.

Considerando que este estudo de caso não teve consumidores externos ao HU-UFS, a técnica de autenticação e autorização utilizada na camada de segurança foi suficiente. No entanto, é sabido que atualmente apenas esta técnica não satisfaz as exigências e conformidades definidas nas leis e regulamentações aplicáveis de segurança da informação, como a preservação da privacidade, aspecto importante no âmbito da saúde. A presença da camada de segurança na *Salutem* contempla positivamente essas demandas.

Ao analisar os resultados obtidos com o estudo de caso foi possível observar, de forma resumida, que:

- A importância do papel do analista de conhecimento.
- Os níveis de segurança e privacidade dos serviços a serem disponibilizados são aspectos que devem receber uma atenção especial, já que a depender dos consumidores, esses serviços não poderão ser disponibilizados. Neste estudo de caso não houve problema de privacidade uma vez que as informações apresentadas estavam consolidadas, sem a identificação dos pacientes.
- A necessidade da adoção de uma ontologia;
- A identificação dos serviços a serem disponibilizados, sendo que em alguns casos somente será possível após uma análise exploratória e identificação pelo analista de conhecimento;
- Os serviços, apesar de poucos, foram bem avaliados pelos consumidores, gerando expectativas para novos serviços, inclusive a inserção de novas fontes de dados; e,
- As operações de cada camada foram facilmente compreendidas e utilizadas pelos atores envolvidos.

Sendo assim, é possível afirmar dentro deste estudo de caso a aplicabilidade e viabilidade da arquitetura *Salutem*, em guiar o desenvolvimento de uma solução de *big data* para área da saúde. O estudo de caso também obteve êxito no cumprimento dos objetivos propostos e atendimento das expectativas dos consumidores com os serviços ofertados.

5.12 Considerações Finais do Capítulo

Este capítulo descreveu um estudo de caso para validar a aplicabilidade da arquitetura *Salutem*. O objeto de estudo foi a construção de uma solução de serviços baseada em grande volume de dados, no caso a base de dados dos resultados de exames de hemogramas realizados no Hospital Universitário da UFS de 2004 a 2016.

O estudo de caso escolhido permitiu avaliar a aplicabilidade da arquitetura em um projeto real. Os resultados apontaram que a arquitetura *Salutem* pode ser uma opção viável para construção de soluções de *big data* por instituições de saúde que ainda não adotam tal tecnologia. Os serviços entregues corroboraram para que chegasse ao objetivo do problema em questão, e por conseguinte satisfação dos clientes.

No próximo capítulo serão apresentadas as considerações finais, principais contribuições, limitações e trabalhos futuros.

CAPÍTULO 6

CONSIDERAÇÕES FINAIS E TRABALHOS FUTUROS

Na última década os avanços na produção e coleta de dados, bem como na capacidade de analisar eficazmente e o nível de compreensão da informação gerada, tornou *big data* uma poderosa ajuda no contexto da saúde. (SANFORD, 2017). *Big data* tende a potencializar a melhoria na qualidade da prestação de cuidados em saúde, através da vigilância de doenças, novos e mais adequados tratamentos de saúde para os pacientes, entre outros.

Entretanto, implantar e utilizar soluções de *big data* no contexto da saúde não é um processo trivial. Isto porque é um ambiente em que os processos são dinâmicos e, principalmente, trabalha com dados, em sua maioria, sensíveis no quesito segurança. Um exemplo são os registros eletrônicos do paciente, que devem ser assegurados contra acessos indevidos. Outro ponto importante ao se lidar com grandes volumes de dados na saúde diz respeito à unicidade do vocabulário e interpretação dos dados e informações. Nos ambientes de saúde, a utilização de uma semântica única permite que o mesmo dado seja entendível e gere interpretações não dúbias.

Com o objetivo de simplificar a implantação de *big data* no âmbito da saúde, além de garantir a segurança na coleta, análise e distribuição dos dados e serviços, bem como, assegurar a semântica dos dados e informações, este trabalho propôs a *Salutem*, uma arquitetura para *big data* para área da saúde. A arquitetura permite desenvolver soluções de serviços de alto valor agregado que podem auxiliar na tomada de decisões por profissionais de saúde, vislumbrando a melhoria na qualidade da prestação de serviços de saúde ao paciente. Também poderá contribuir em pesquisas com o intuito de facilitar e responder questões em aberto ou mesmo encontrar outras linhas de pesquisa.

A aplicabilidade da *Salutem* foi verificada em um estudo de caso real, com a construção de solução de serviços baseada em dados de hemogramas de um hospital/escola público federal, com vistas a explorar os dados respondendo questões problemas através de serviço. Portanto, a resposta para a pergunta que motivou esta pesquisa é:

- Sim, uma arquitetura de *big data* para área da saúde pode facilitar o uso e implantação em tais áreas, bem como tratar problemas inerentes a tais ambientes, como a semântica envolvida e a segurança das informações.

Finalmente, conclui-se que a *Salutem* pode ser empregada em soluções de *big data* na saúde como uma arquitetura simples e de fácil implantação, da qual podem ser utilizadas diversas tecnologias e técnicas nas operações em suas camadas. Outro ponto a destacar é o entendimento que a arquitetura favorece a identificação do que é necessário realizar para prover entregas (serviços) aos consumidores, de acordo com o problema tratado.

6.1 Principais Contribuições

Dentre as contribuições deste trabalho pode-se destacar a análise das arquiteturas para *big data*, no capítulo 3, através de uma Revisão Sistemática de Literatura. Foram identificadas propostas de trabalhos de arquitetura de *big data,* incluindo propostas para área da saúde, sendo estas analisadas à luz das características e desafios de *big data.* Sendo assim, destacaram-se as principais dificuldades e limitações existentes nas arquiteturas encontradas.

A principal contribuição deste trabalho é a proposta da arquitetura *Salutem*, a qual tem por objetivo facilitar o desenvolvimento de soluções de *big data* para área da saúde, tendo como premissa os desafios de *big data*. Na seção 4.9, como parte do desenvolvimento da *Salutem*, realizou-se um estudo para definição de uma taxonomia de referência para arquitetura.

No capítulo 5, descreve-se estudo de caso com a aplicabilidade da arquitetura *Salutem* e os resultados obtidos, como o protótipo da solução Hemo *Data*.

6.2 Limitações da Pesquisa e Trabalhos Futuros

Uma limitação deste trabalho refere-se à aplicação da *Salutem* em apenas um estudo de caso. A validação efetiva da arquitetura demanda sua aplicação em mais casos, com outros dados, como por exemplo, imagens médicas. Não obstante, a arquitetura foca a área da saúde, pode ser relevante e ter aplicabilidade em outras áreas como a jurídica, a financeira, a varejista, entre outras. Portanto, trabalhos futuros podem contribuir para a avaliação experimental e a melhoria contínua da *Salutem* em outros contextos.

Apesar dos benefícios proporcionados pelos aspectos singulares da *Salutem*, a arquitetura é altamente dependente do analista do conhecimento. No desenvolvimento de outros tipos de sistemas um bom levantamento de requisitos já possibilita o início do processo de desenvolvimento. Diferentemente, soluções e serviços baseados em *big data* somente serão efetivos com intensa participação do analista do conhecimento. Como trabalho futuro é

interessante verificar alternativas na literatura das atividades para este ator, bem como, identificar outros papéis que possam influenciar a aplicação da arquitetura.

Outro trabalho futuro previsto é a ampliação do Hemo *Data* tanto para a oferta de novos serviços quanto para novos provedores de dados, a exemplo dos dados oriundos de secretarias municipais e estaduais de saúde, possibilitando uma gama de serviços de interesse da saúde pública.

REFERÊNCIAS

AGRAWAL, R. *et al. A Layer Based Architecture for Provenance in Big Data. Big Data (Big Data), 2014 IEEE International Conference on*, p. 1–7, 2014.

AHMED, F. D. *et al. Agent-Based Big Data Analytics in Retailing : A Case Study*. p. 67–72, 2015.

ALEXANDER, M. Excel Dashboards & Reports For Dummies. 3ª edição ed. [s.l.] John Wiley & Sons, Inc., 2016.

AL-JAROODI, J.; MOHAMED, N. *Characteristics and Requirements of Big Data Analytics Applications. 2nd International Conference on Collaboration and Internet Computing Characteristics*, 2016.

AMATO, A.; DI MARTINO, B.; VENTICINQUE, S. *Big data processing for pervasive environment in cloud computing. Proceedings - 2014 International Conference on Intelligent Networking and Collaborative Systems*, IEEE INCoS 2014, p. 598–603, 2015.

ANDREU-PEREZ, J. *et al. Big Data for Health*. v. 19, n. 4, p. 1193–1208, 2015.

ANTHONY, R.; GARLAND, J. *Large-Scale Software Architecture: A Practical Guide Using UML*. 2002.

BABU, P. M. S. P.; SASTRY, S. H. *Big Data and Predictive Analytics in ERP Systems for Automating Decision Making Process. Software Engineering and Service Science* (ICSESS), 2014 5th IEEE *International Conference on*, p. 259–262, 2014.

BAIN, B. J. *A Beginner's Guide to Blood Cells*. [s.l.] Wiley-Blackwell, 2008.

BALDOMINOS, A. *et al. A scalable machine learning online service for big data real-time analysis*. 2014 IEEE *Symposium on Computational Intelligence in Big Data* (CIBD), p. 1–8, 2014.

BALDOMINOS, A. *et al. An efficient and scalable recommender system for the smart web. Proceedings - 2015 11th International Conference on Innovations in Information Technology*, IIT 2015, p. 296–301, 2015.

BANSAL, S. *Towards a Semantic Extract-Transform-Load (ETL) framework for big data integration. Proceedings - 2014 IEEE International Congress on Big Data, BigData Congress 2014.*

BARROS, A. J. S.; LEHFELD, N. A. S. Fundamentos de Metodologia: Um Guia para a Iniciação Científica. 2000.

BASS, L.; CLEMENTS, P.; KAZMAN, R. *Software Architecture in Practice. [s.l.] Addison-Wesley Professional; 3 edition*, 2012.

BENJELLOUN, F.-Z.; LAHCEN, A. A.; BELFKIH, S. *An overview of big data opportunities, applications and tools. Intelligent Systems and Computer Vision* (ISCV), 2015, p. 1–6, 2015.

BHOGAL, J.; CHOKSI, I. *Handling Big Data using NoSQL. 9th International Conference on Advanced Information Networking and Applications Workshops*. 2015.

BLOBEL, B. *Ontology driven health information systems architectures enable pHealth for empowered patients. International Journal of Medical Informatics*, v. 80, n. 2, p. e17–e25, 2010.

BRERETON, M., MCCAFFERTY, R., MARSDEN, K., KAWAI, Y., ETZELL, J., ERMENS, A. *Recommendation for standardization of haematology reporting units used in the extended blood count. International Journal of Laboratory Hematology*, 472–482. 2016

CAGLIERO, L. *et al. Monitoring the citizens' perception on urban security in Smart City environments. Proceedings - International Conference on Data Engineering, v. 2015–June*, p. 112–116, 2015.

CECCHINEL, C. *et al. An Architecture to Support the Collection of Big Data in the Internet of Things. 2014 IEEE World Congress on Services*, p. 442–449, 2014.

CHANDARANA, P.; VIJAYALAKSHMI, M. *Big Data analytics frameworks. 2014 International Conference on Circuits, Systems, Communication and Information Technology Applications* (CSCITA), p. 430–434, 2014.

CHEN, M. *et al. Big Data: Related Technologies, Challenges and Future Prospects. Cham: Springer International Publishing*, 2014.

CHEN, M.; MAO, S.; LIU, Y. *Big data: A survey. Mobile Networks and Applications*, v. 19, n. 2, p. 171–209, 2014a.

CHEN, M.; MAO, S.; LIU, Y. *Big Data : A Survey. n. January*, p. 171–209, 2014b.

CHENG, B. *et al. Building a Big Data Platform for Smart Cities: Experience and Lessons from Santander. Proceedings - 2015 IEEE International Congress on Big Data, BigData Congress* 2015, p. 592–599, 2015.

CRICHTON, D. *et al. A Scalable Planetary Science Information Architecture for Big Science Data. e-Science (e-Science), 2014 IEEE 10th International Conference on*, v. 1, p. 196–203, 2014.

DAIF, A. *et al. Review Current CRM Architectures and Introducing New Adapted Architecture to Big Data. IEEE International Professional Communication Conference*, 2015.

DANA KATHRIN TOMIC, S.; FENSEL, *A. OpenFridge: A platform for data economy for energy efficiency data. Proceedings - 2013 IEEE International Conference on Big Data, Big Data* 2013, p. 43–47, 2013.

DANIEL, A.; PAUL, A.; AHMAD, *A. Near real-time big data analysis on vehicular networks. Soft-Computing and Networks Security (ICSNS), 2015 International Conference on*, p. 1–7, 2015.

DAVENPORT, T. Big Data no Trabalho: Derrubando mitos e descobrindo oportunidades. Elsevier Brasil. 2014.

DEMCHENKO, Y. *et al. Addressing big data issues in Scientific Data Infrastructure. Proceedings of the 2013 International Conference on Collaboration Technologies and Systems*, CTS 2013, p. 48–55, 2013.

DI MAURO, M.; DI SARNO, C. *A framework for Internet data real-time processing: A machine-learning approach. 2014 International Carnahan Conference on Security Technology* (ICCST), p. 1–6, 2014.

DUAN, R. *et al. HiPerData: An autonomous large-scale model building and management platform for big data analytics. Industrial Electronics and Applications (ICIEA), 2015 IEEE 10th Conference on*, p. 449–454, 2015.

ELLEUCH, W.; WALI, A.; ALIMI, A. M. *Mining road map from big database of GPS data. 2014 14th International Conference on Hybrid Intelligent Systems*, HIS 2014, p. 193–198, 2015.

FAILACE, R. Hemograma: Manual de Interpretação. 4ª edição. Porto Alegre: Artmed, 2003.

FERNANDEZ, A.; CASADO, R.; USAMENTIAGA, R. *A Real-Time Big Data Architecture for Glasses Detection Using Computer Vision Techniques. Proceedings - 2015 International Conference on Future Internet of Things and Cloud, FiCloud 2015 and 2015 International Conference on Open and Big Data*, OBD 2015, p. 591–596, 2015.

GACHET, D. *et al. Big Data Processing of Bio-signal Sensors Information for Self-Management of Health and Diseases. Proceedings - 2015 9th International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing*, IMIS 2015, p. 330–335, 2015.

GAO, J. *et al. A Sticky Policy Framework for Big Data Security. n. October*, p. 130–137, 2015.

GARLAN, D.; SHAW, M. *An Introduction to Software Architecture. Knowledge Creation Diffusion Utilization, v. 1, n. January*, p. 1–40, 1994.

GARTNER GROUP. *Gartner Says Worldwide Enterprise IT Spending to Reach $2.7 Trillion in 2012*. Disponível em: <http://www.gartner.com/newsroom/id/1824919>. Último acesso em: 11 de dezembro de 2015.

GEERDINK, B. *A Reference Architecture for Big Data Solutions. The 8th International Conference for Internet Technology and Secured Transactions*, p. 71–76, 2013.

GODBOLE, N. S.; LAMB, J. *Using data science & big data analytics to make healthcare green. 2015 12th International Conference & Expo on Emerging Technologies for a Smarter World* (CEWIT), p. 1–6, 2015.

GRADY, N.; CHANG, W. *DRAFT NIST Big Data Interoperability Framework: Volume 2, Big Data Taxonomies*. 2017.

GROLINGER, K. *et al. Data management in cloud environments : NoSQL and NewSQL data stores. Journal of Cloud Computing: Advances, Systems and Applications 2013*.

GROTTO, H. Z. W. O hemograma: importância para a interpretação da biópsia *Blood cell analysis : the importance for biopsy interpretation*. Revista Brasileira De Hematologia E Hemoterapia. 2009.

GROVES, P. *et al. The "big data"revolution in healthcare. McKinsey Quarterly, n. January*, p. 22, 2013.

GUDIVADA, V.; RAO, D.; RAGHAVAN, V. *NoSQL Systems for Big Data Management. 10th World Congress on Services*. 2014.

HANSMANN, T.; NIEMEYER, P. *Big Data - Characterizing an Emerging Research Field using Topic Models*. 2014.

HARRISON, R. TOGAF TM 9 *Foundation Study Guide. [s.l.] Van Haren Publishing*, 2013.

HU, W. *et al. An Efficient Transportation Architecture for Big Data Movement*. 2013.

HUANG, M. J.; HUANG, C. F.; CHEN, W. S. E. *Architecting a Software-Defined Storage Platform for Cloud Storage Service. Proceedings - 2015 IEEE International Conference on Services Computing*, SCC 2015, p. 379–386, 2015.

HUANG, Y.; YESHA, Y.; ZHOU, S. *A database-based distributed computation architecture with Accumulo and D4M: an application of eigensolver for large sparse matrix*. 2015.

HUGHES, J. S. *et al. PDS4: A Model-Driven Planetary Science Data Architecture for Long-Term Preservation. n. June* 1990, p. 134–141, 2014.

HURWITZ, J. *et al. Big Data For Dummies. [s.l.] For Dummies; 1 edition*, 2013.

ILIINSKY, N.; STEELE, J. *Designing Data Visualizations*. [s.l.] O'Reilly Media, 2011.

IMMONEN, A.; OVASKA, E. *Evaluating the Quality of Social Media Data in Big Data Architecture*. v. 3, 2015.

JUNIOR, J. C. DA S. F. *et al.* Dimensões de Big Data e o Processo Decisório: Estudos de Casos Múltiplos no Varejo. p. 1–15, 2015.

JUNQUEIRA, L. C.; CARNEIRO, J. Histologia Básica. Guanabara Koogan SA. Rio de Janeiro. 2008.

JUTLA, D. N.; BODORIK, P. PAUSE: *A privacy architecture for heterogeneous big data environments. Proceedings - 2015 IEEE International Conference on Big Data, IEEE Big Data* 2015, p. 1919–1928, 2015.

KASHLEV, A.; LU, S. *A system architecture for running big data workflows in the cloud. Proceedings - 2014 IEEE International Conference on Services Computing*, SCC 2014, p. 51–58, 2014.

KATAL, A.; WAZID, M.; GOUDAR, R. H. *Big data: Issues, challenges, tools and Good practices. 2013 6th International Conference on Contemporary Computing*, IC3 2013, p. 404–409, 2013.

KELLY, C.; CRAIG, K.; MATTHEWS, M. *Real-time predictive analytics to estimate air traffic flow rates*. 2015.

KHAN, Z.; ANJUM, A.; KIANI, S. L. *Cloud based big data analytics for smart future cities. Proceedings - 2013 IEEE/ACM 6th International Conference on Utility and Cloud Computing*, UCC 2013, p. 381–386, 2013.

KIONG, Y. C.; PALANIAPPAN, S.; YAHAYA, N. A. *Health ontology system. 2011 7th International Conference on Information Technology in* Asia, p. 1–4, 2011.

KITCHENHAM, B. *Procedures for Performing Systematic Reviews*. Software Engineering Group, Keele University. 2004.

KORSTER, P.; SEIDER, C. *The world's 4 trillion dollar challenge. Executive Report of IBM Global Business Services*. 2010.

KOWSAR, Y.; AFGAN, E. *Support for data-intensive computing with CloudMan. 2013 36th International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2013 - Proceedings*, p. 243–248, 2013.

KOWSAR, Y.; DASHNOW, H.; LONIE, A. *Data interlocking: Coupling analytics to the data. Proceedings - 2014 IEEE/ACM 7th International Conference on Utility and Cloud Computing*, UCC 2014, p. 696–701, 2015.

KRISHNAN, K. *Data Warehousing in the Age of Big Data*. [s.l.] Morgan Kaufmann, 2013.

KRUMEICH, J. *et al. Advanced Planning and Control of Manufacturing Processes in Steel Industry through Big Data Analytics Case Study and Architecture Proposal. Big Data (Big Data), 2014 IEEE International Conference on*, p. 16–24, 2014.

KUPWADE PATIL, H.; SESHADRI, R. *Big Data Security and Privacy Issues in Healthcare. 2014 IEEE International Congress on Big Data*, p. 762–765, 2014.

LABRINIDIS, A.; JAGADISH, H. V. *Challenges and opportunities with big data. Proceedings of the VLDB Endowment*, v. 5, p. 2032–2033, 2012.

LARKOU, G. *et al. Managing big data experiments on smartphones. Distributed and Parallel Databases*, v. 34, n. 1, p. 33–64, 2016.

LAU, S.; YEW, B. *Building Use Cases With Activity Reference Framework for Big Data Analytics*. 2015.

LI, S.; NI, J. *Evolution of Big-Data-Enhanced Higher Education Systems. 2015 Eighth International Conference on Internet Computing for Science and Engineering* (ICICSE), p. 253–258, 2015.

LIU, R. *et al. Big Data architecture for IT incident management. Proceedings of 2014 IEEE International Conference on Service Operations and Logistics, and Informatics*, SOLI 2014, p. 424–429, 2014.

LIU, W. *et al. A prototype of healthcare big data processing system based on Spark*. p. 516–520, 2015.

LIU, Z.; YANG, P.; ZHANG, L. *A Sketch of Big Data Technologies. 2013 Seventh International Conference on Internet Computing for Engineering and Science*, p. 26–29, 2013.

MADKOUR, A.; AREF, W. G.; BASALAMAH, S. *Knowledge cubes - A proposal for scalable and semantically-guided management of Big Data. Proceedings - 2013 IEEE International Conference on Big Data, Big Data* 2013, 2013.

MAHMOOD, T; AFZAL, U. *Security Analytics: Big Data Analytics for Cybersecurit. 2nd National Conference on Information Assurance (NCIA). 2013.*

MAJI, P., MANDAL, A., GANGULY, M., & SAHA, S. *An Automated Method for Counting and Characterizing Red Blood Cells Using Mathematical Morphology*, 6–11. 2015.

MALCOLM, R. *Increasing the accessibility to Big Data systems via a common services API. Proceedings - 2014 IEEE International Conference on Big Data, IEEE Big Data 2014*.

MARCHAL, S. *et al. A big data architecture for large scale security monitoring. Proceedings - 2014 IEEE International Congress on Big Data, BigData Congress* 2014, p. 56–63, 2014.

MCKINNEY, W. *Data Structures for Statistical Computing in Python. Proceedings of the 9th Python in Science Conference*, 1697900(Scipy), 51–56. 2010.

MEZGHANI, E. *et al. A Semantic Big Data Platform for Integrating Heterogeneous Wearable Data in Healthcare. Journal of Medical Systems*, 2015.

MELLO, C. H. P. *et al* . Pesquisa-ação na engenharia de produção: proposta de estruturação para sua condução. Produção, São Paulo, v. 22, n. 1, 2012. Disponível em: < http://www.scielo.br/pdf/prod/v22n1/aop_t6_0010_0155.pdf>. Último acesso em 10 de março de 2017.

MICHAELIS. Dicionário Brasileiro da Língua Portuguesa, 2017. Disponível em: < http://michaelis.uol.com.br/moderno-portugues>. Último acesso em 10 de março de 2017.

MIGUEL, P. A. C. Estudo de caso na engenharia de produção: estruturação e recomendações para sua condução. Produção, São Paulo, v. 17, n. 1, Abril. 2007. Disponível em: < http://www.scielo.br/pdf/prod/v17n1/14.pdf>. Último acesso em 10 de março de 2017.

MOHAMED, N. *et al. Real-time big data analytics: Applications and challenges. 2014 International Conference on High Performance Computing & Simulation* (HPCS), p. 305–310, 2014.

MOHANTY, H.; BHUYAN, P.; CHENTHATI, D. *Big Data: A Primer*. [s.l.] Springer, 2015.

NI, J. *et al. SoSE Architecture of Data Intensive Computing for Healthcare Information System*. 2015.

PENA, P. A.; SARKAR, D.; MAHESHWARI, P. *A big-data centric framework for smart systems in the world of internet of everything. Proceedings - 2015 International Conference on Computational Science and Computational Intelligence*, CSCI 2015, p. 306–311, 2015.

PEREZ, J., CARRERA, D. *Performance characterization of the servioticy API: An IoT-as-a-service data management platform. Proceedings - 2015 IEEE 1st International Conference on Big Data Computing Service and Applications, BigDataService 2015.*

PRAJAPATI, V. *Big Data Analytics with R and Hadoop. Packt Publishing.* 238p. Novemvro, 2013.

RABELO, T. *et al. SmartLAK: A Big Data Architecture for Supporting Learning Analytics Services. Frontiers in Education Conference* (FIE), 2015. 32614 2015. IEEE, p. 1–5, 2015.

RAGHUPATHI, W.; RAGHUPATHI, V. *Big data analytics in healthcare: promise and potential. Health Information Science and Systems*, v. 2, n. 1, p. 3, 2014.

RAMASUBRAMANIAN, K.; SINGH, A. *Machine Learning Using R.* [s.l.] Apress; 1st ed. Edition. 2017.

RAMESH, B. *Big Data Architecture. In: MOHANTY, H.; BHUYAN, P.; CHENTHATI, D. (Eds.). Big Data: A Primer*. New Delhi: Springer India, p. 29–59. 2015.

RAO, S.; SUMA, S. N.; SUNITHA, M. *Security Solutions for Big Data Analytics in Healthcare. 2015 Second International Conference on Advances in Computing and Communication Engineering*, p. 510–514, 2015.

RATHORE, M. M. *et al. Hadoop based real-Time Big Data Architecture for remote sensing Earth Observatory System. 6th International Conference on Computing, Communications and Networking Technologies*, ICCCNT 2015, v. i, 2015a.

RATHORE, M. M. U. *et al. Real-Time Big Data Analytical Architecture for Remote Sensing Application. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, v. 8, n. 10, p. 4610–4621, 2015b.

RAYNAUD, T.; HAQUE, R.; AÏT-KACI, H. *CedCom: A high-performance architecture for Big Data applications. Proceedings of IEEE/ACS International Conference on Computer Systems and Applications*, AICCSA, v. 2014, p. 621–632, 2015.

ROOPA, T.; IYER, A. N.; RANGASWAMY, S. *CroTIS-Crowdsourcing Based Traffic Information System. 2013 IEEE International Congress on Big Data*, p. 271–277, 2013.

RYAN, *A. Towards semantic interoperability in healthcare: ontology mapping from SNOMED-CT to HL7 version 3. Reproduction*, v. 72, p. 69–74, 2006.

SANTOVENA, A. Z. *Big Data : Evolution , Components , Challenges and Opportunities*. p. 126, 2013.

SCHNEIDER, R. D. *Hadoop for Dummies. [s.l.] For Dummies; 1 edition*, 2012.

SHAO, G.; SHIN, S.-J.; JAIN, S. *Data Analytics using Simulation for smart Manufacturing. IEEE Winter Simulation Conference, n. Smlc* 2012, p. 2192–2203, 2014.

SILVA, E. L.; MENEZES, E. M. Metodologia da Pesquisa e Elaboração de Dissertação - 4a edição. Portal, p. 138p, 2005.

SINGH, S.; LIU, Y. *A cloud service architecture for analyzing big monitoring data. Tsinghua Science and Technology*, v. 21, n. 1, p. 55–70, 2016.

SINNOTT, R. O.; MORANDINI, L.; WU, S. SMASH: *A Cloud-Based Architecture for Big Data Processing and Visualization of Traffic Data*. p. 53–60, 2015.

SIRIWEERA, T. H. A. S. *et al. Intelligent Big Data Analysis Architecture Based on Automatic Service Composition. 2015 IEEE International Congress on Big Data*, p. 276–280, 2015.

SOCEANU, A. *et al. Managing the Privacy and Security of eHealth Data*. 2015 *20th International Conference on Control Systems and Computer Science*, p. 439–446, 2015.

SOMMERVILLE, I. *Software Engineering. 9. ed. Harlow, England: Addison-Wesley.* 2010.

STANFORD. *Stanford Medicine 2017 Health Trends Report. Harnessing the Power of Data in Health*.2017.

TEKIEH, M.; RAAHEMI, *B. Importance of Data Mining in Healthcare : A Survey*. 2015.

THURAISINGHAM, B. D. *Big Data Security and Privacy. Proceedings of the 5th ACM Conference on Data and Application Security and Privacy*, p. 279–280, 2015.

TWARDOWSKI, B.; RYZKO, D. *Multi-agent Architecture for Real-Time Big Data Processing. 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies* (IAT), p. 333–334, 2014.

VENTURA, M. M. O Estudo de Caso como Modalidade de Pesquisa. Pedagogia Médica. 2007.

VIANA, P.; SATO, L. *A proposal for a reference architecture for long-term archiving, preservation, and retrieval of big data. Proceedings - 2014 IEEE 13th International Conference on Trust, Security and Privacy in Computing and Communications, TrustCom 2014*, p. 622–629, 2015.

VIEIRA, M. R. *et al.* Bancos de Dados NoSQL: conceitos, ferramentas, linguagens e estudos de casos no contexto de Big Data. Simpósio Brasileiro de Bancos de Dados - SBBD 2012, n. 1, p. 1–30, 2012.

VIVAS, W. L. P. Manual Prático de Hematologia. Files.Blog-Da-Rosania.Webnode.Com, p. 1–33, 2008.

WANG, K. *et al. Mobile big data fault-tolerant processing for ehealth networks. IEEE Network*, v. 30, n. 1, p. 36–42, 2016.

WHO. *World Health Organization* 2017. Disponível em <http://www.who.int/gho/health_financing/per_capita_expenditure/en/>. Último acesso em 20 de junho de 2017.

XINHUA, E. *et al. Big Data-as-a-Service : Definition and architecture. 15th IEEE International Conference on Communication Technology*. 2013.

ZHANG, Q. *et al. A universal storage architecture for big data in cloud environment. Proceedings - 2013 IEEE International Conference on Green Computing and Communications and IEEE Internet of Things and IEEE Cyber, Physical and Social Computing, GreenCom-iThings-CPSCom 2013*, p. 476–480, 2013.

ZHU, S. *et al. A Distributed Architecture for Rule Engine to Deal with Big Data*. p. 602–606, 2016.

ZILLNER, S.; NEURURER, S. *Big Data in the Health Sector. New Horizons for a Data-Driven Economy: A Roadmap for Usage and Exploitation of Big Data in Europe. Cham: Springer International Publishing*, p. 179–194. 2016.

ZIKOPOULOS, P.; EATON, C. *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*. [s.l.] McGraw-Hill Education, 2011.

ZIMMERMANN, A. *et al. Towards service-oriented enterprise architectures for big data applications in the cloud. Proceedings - IEEE International Enterprise Distributed Object Computing Workshop*, EDOC, p. 130–135, 2013.

ZULKERNINE, F. *et al. Towards cloud-based analytics-as-a-service (CLAaaS) for big data analytics in the cloud. Proceedings - 2013 IEEE International Congress on Big Data, BigData* 2013, p. 62–69, 2013.