MINISTÉRIO DA CIÊNCIA, TECNOLOGIA, INOVAÇÕES E COMUNICAÇÕES
INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS

O USO DE CUBO DE DADOS COMO UMA SOLUÇÃO DE *BIG DATA* COMO UMA FERRAMENTA PARA TOMADA DE DECISÃO

Yuri Matheus Dias Pereira
Mauricio Vieira Ferreira Gonçalves
Rodrigo Rocha Silva

Relatório Técnico resultado do Exame de Proposta de Dissertação do Curso de Pós-Graduação em Engenharia e Gerenciamento de Sistemas Espaciais.

URL do documento original:
<http://urlib.net/>

INPE
São José dos Campos
2019

PUBLICADO POR:

Instituto Nacional de Pesquisas Espaciais - INPE
Gabinete do Diretor (GB)
Serviço de Informação e Documentação (SID)
Caixa Postal 515 - CEP 12.245-970
São José dos Campos - SP - Brasil
Tel.:(012) 3945-6923/6921
Fax: (012) 3945-6919
E-mail: pubtc@sid.inpe.br

MINISTÉRIO DA CIÊNCIA, TECNOLOGIA, INOVAÇÕES E COMUNICAÇÕES
INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS

O USO DE CUBO DE DADOS COMO UMA SOLUÇÃO DE *BIG DATA* COMO UMA FERRAMENTA PARA TOMADA DE DECISÃO

Yuri Matheus Dias Pereira
Mauricio Vieira Ferreira Gonçalves
Rodrigo Rocha Silva

Relatório Técnico resultado do Exame de Proposta de Dissertação do Curso de Pós-Graduação em Engenharia e Gerenciamento de Sistemas Espaciais.

URL do documento original:
<http://urlib.net/>

INPE
São José dos Campos
2019

Informar aqui sobre marca registrada (a modificação desta linha deve ser feita no arquivo publicacao.tex).

"But I try not to think with my gut. If I'm serious about understanding the world, thinking with anything besides my brain, as tempting as that might be, is likely to get me into trouble. It's OK to reserve judgment until the evidence is in.".

Carl Sagan e Ann Druyan
em *"O Mundo Assombrado pelos Demônios: A Ciência Vista Como Uma Vela no Escuro"*, 1995

RESUMO

Satélites são monitorados pelas equipes de solo via pacotes de telemetria, que informam o estado atual dos equipamentos e permitem avaliar a capacidade do satélite de continuar a sua missão. Esses pacotes de telemetria constituem um corpo de dados de tamanho e complexidade significativa, sendo que satélites que funcionam por vários anos geram dados históricos de grande volume, ainda úteis para a operação. Neste artigo apresentamos uma arquitetura baseada em conceitos de Big Data e Business Intelligence para criar uma representação de dados de telemetria pronta para a análise por operadores e engenheiros de satélite no Instituto Nacional de Pesquisas Espaciais (INPE), bem como apresentamos o fluxo de dados utilizado pelos dados históricos de telemetria de um dos satélites operados pelo INPE.

Palavras-chave: Cubo de Dados. Big Data. Operação. Satélite. Data Warehouse.

LISTA DE FIGURAS

Pág.

LISTA DE TABELAS

Pág.

LISTA DE ABREVIATURAS E SIGLAS

WETAMC	–	Campanha de Mesoescala Atmosférica na Estação Úmida
IBGE	–	Instituto Brasileiro de Geografia e Estatística
MC	–	Método das Covariâncias
EDO	–	Equações Diferenciais Ordinárias
EDP	–	Equações Diferenciais Parciais
ECT	–	Energia Cinética Turbulenta
FDP	–	Função de Distribuição de Probabilidade
PR	–	Plot de Recorrência
FFT	–	Fast Fourier Transform
tS1200	–	Temperatura medida no nível superior às 12 horas
tS2300	–	Temperatura medida no nível superior às 23 horas
tM1200	–	Temperatura medida no nível médio às 12 horas
tM2300	–	Temperatura medida no nível médio às 23 horas
tI1200	–	Temperatura medida no nível inferior às 12 horas
tI2300	–	Temperatura medida no nível inferior às 23 horas
wS1200	–	Velocidade vertical do vento medida no nível superior às 12 horas

SUMÁRIO

Pág.

1 INTRODUÇÃO

O Centro de Controle de Satélites (CCS) localizado no Instituto Nacional de Pesquisas Espaciais (INPE) atualmente monitora e controla alguns satélites: a família do Satélite de Coleta de Dados (SCD), composta de dois satélites SCD-1 e SCD-2, e o Satélite Sino-Brasileiro de Recursos Terrestres (CBERS), com o quinto satélite em operação atualmente, o CBERS-4. Estes satélites realizam passagens sobre as estações terrenas do INPE, durante o qual o CCS recebe dados de telemetria e envia telecomando, bem como realiza atividades de manutenção e estimativa, como medidas de velocidade e posição (AZEVEDO; AMBRÓSIO, 2010).

Dados de telemetria geralmente carregam medidas de sensores e verificações de saúde dos instrumentos, como temperatura das baterias, corrente de algum subsistema, se um dado equipamento está ativo ou não, bem como dados que os operadores e engenheiros acham necessários para a operação, entre outros (AZEVEDO; AMBRÓSIO, 2010). Estes dados precisam ser guardados por toda a vida do satélite, sendo que para satélites que estão funcionando por vários anos, eles adquirem um volume considerável, que não pode ser descartado, como é o caso dos satélites da família SCD, com o SCD-1 já estando operacional por mais de 25 anos, e continuando a gerar dados.

Para satélites complexos como os da família CBERS, que possuem mais de 4 mil telemetrias sendo rastreadas múltiplas vezes por dia, temos um grande volume de dados cuja análise não é trivial, e só pode ser propriamente feita pela engenharia do satélite qualificados e com experiência para isso. Com os lançamentos futuros do CBERS-4A e do Amazônia-1, o volume de dados e a complexidade da análise dos mesmos deve aumentar, criando novas necessidades de operação. Não tenho citação pra isso!

1.1 Problemas

These data are used by the satellite operators to check the operational capacity of the satellites, see the health of the subsystems and if they're working properly, and that the satellite will continue to perform its duties properly in the near future. In the case of an emergency they might need to check old telemetry data to see if a situation has occured before, and check whether that might prove a danger to the satellite or not [@AzevedoAmbrVieiEsSoTe].

The historical analysis is very important for satellite operations, as it might unearth

rare phenomena and can serve as an early warning that some issue might appear in the future. One example is in the case of CBERS-2, which had the phenomenon of thermal breakdown happening to one of its batteries [@Magalhaes:2012:EsAvTe], and having the historical telemetries was fundamental in the analysis of the phenomenon, and to the lessons learned with it by the operations team.

In this work, we present a data structure called data cubes to make the analysis on satellite telemetry data easier to be performed. Since that a Data Warehouse hasn't been implemented yet for the telemetry data, doing analysis is quite a slow process that involves a lot of manual steps and the creation of complex, not easy to generalize, queries and custom code. The aim of this structure is to make the analysis an action that is easy to perform for the operation of current and future satellites, with good average response times, thus aiming to improve INPE's satellite operations capabilities.

1.2 Objetivos

?

1.3 Organização da proposta

Os capítulos restantes deste trabalho estão organizados da seguinte maneira:

- Capítulo 2: Este capítulo apresenta os conceitos e fundamentos teóricos, como apresentando os conceitos do Cubo de Dados, *Big Data* e operação de satélites.
- Capítulo 3: Neste capítulo os trabalhos correlatos de Cubo de Dados são apresentados, bem como outros operadores de satélite estão resolvendo os problemas identificados.
- Capítulo 4: Neste capítulo a arquitetura proposta é apresentada e seus conceitos principais explicados, bem como o fluxo de dados atual do CCS e como a nova arquitetura vai melhorá-lo.
- Capítulo 5: Esse capítulo apresenta os resultados alcançados até o momento.
- Capítulo 6: Com base nos resultados intermediários alcançados, esse capítulo apresentará as conclusões obtidas, bem como as direções de implementação para o resto do trabalho.

2 FUNDAMENTAÇÃO

Este capítulo apresenta os conceitos fundamentais relacionados a essa proposta, começando pelo básico de operação dos satélites, apresentando os conceito de *Data Warehouse*, *OLAP* e Cubo de Dados, e por fim a definição de *Big Data* utilizada no trabalho.

2.1 Operação de Satélites

Dataflow? Explicar a operação de satélites bem por cima

2.2 Big Data

O termo *Big Data* vem evoluindo ao longo dos anos, e para este trabalho vamos utilizar a definição dos *5 Vs* (BIMONTE, 2016): Volume, Variedade, Velocidade, Valor e Veracidade. Em detalhes:

- **Volume**: esse termo geralmente especifica uma quantidade de dados em que um sistema tradicional de gerenciamento de banco de dados é ineficaz. É importante ressaltar que isso não se trata apenas do armazenamento dos dados, mas também do seu processamento (BOUSSOUF et al., 2018). Usar um grande volume de dados geralmente implica em modelos melhores, que então produzem análises melhores, justificando a coleta de uma grande quantidade de dados.

- **Variedade**: dados são provenientes de fontes diferentes, com formatos diferentes, sem um esquema de modelagem padronizado, como dados advindos de *logs* de computadores, dados de sensores, dados multimídia, etc. Como consequência, esses dados devem ser utilizados da forma mais transparente o possível na análise.

- **Velocidade**: dados são disponibilizados de uma forma muito rápida, e devem ser analisados da forma mais rápida o possível. Isso implica que os dados podem ser guardados e analisados até em tempo real.

- **Valor**: os dados devem ser armazenados para criar algum valor para os seus usuários, seja ele econômico, científico, social, organizacional, etc.

- **Veracidade**: os dados não possuem garantias quanto a sua qualidade, como inconsistências e falta de acurácia, porém a análise deve ser de alta qualidade de qualquer forma.

Estes V's estão relacionados com a construção de um *Data Warehouse*, sendo que também podem ser vistos como requisitos para a crição de um para um conjunto de dados caracterizado como *Big Data* (ZHANG et al., 2017). Em especial, existe um certo relacionamento com a ideia de "*NoSQL*"("Não apenas SQL", em inglês), em que não apenas sistemas de banco de dados relacionais são utilizados, mas também outros paradigmas são utilizados, como orientados a documentos, chave e valor, etc (BIMONTE, 2016).

2.3 Data Warehouse

Um Armazém de Dados ou Data Warehouse (DW) é um repositório de dados orientado por assunto, integrado, variado ou particionado em função do tempo e não volátil, que auxilia no gerenciamento do processo de tomada decisões (INMON; HACKATHORN, 1994). Essa definição pode ser dividida em:

- **Orientado por assunto**: o DW é utilizado para a análise de uma área em específico. Por exemplo, é de interesse analisar especialmente os dados da carga útil de uma forma específica.
- **Integrado**: o DW deve integrar dados vindos de múltiplas fontes de uma forma estrutura. Por exemplo, mesmo que existam duas representações diferentes para um mesmo produto, o DW deve possuir apenas uma representação. Isso requer o uso de técnicas de limpeza e integração dos dados, de modo a garantir a consistência dos dados.
- **Variado em função do tempo**: o DW deve conter, explícita ou implicitamente a perspectiva de tempo. Isso quer dizer que o DW possui dados históricos e eles podem ser consultados durante a análise. Por exemplo, pode se querer saber de dados de dias, meses ou anos atrás.
- **Não volátil**: uma vez dentro do DW, os dados não são removidos ou atualizados, sendo um requisito para a consulta de dados históricos.

Essas características diferem o *Data Warehouse* de outros sistemas de repositório, como sistemas de banco de dados, sistemas de processamento de transações e sistemas de arquivos (HAN et al., 2011).

Um DW é geralmente representado por um modelo dimensional que permite eficiência na organização dos dados e na recuperação de informações gerenciais (KIMBALL;

ROSS, 2013). Neste modelo são definidos fatos, dimensões e medidas. Um fato corresponde ao assunto de negócio a ser analisado, cada dimensão é uma perspectiva de visualização do assunto de negócio e medidas são valores numéricos que quantificam o assunto de negócio. Uma das dimensões é sempre temporal para permitir a análise do assunto ao longo do tempo (SILVA, 2015).

2.4 OLAP

On-line Analytical Processing (OLAP) é um termo que se refere a um conjunto de ferramentas que são utilizadas para resumir, consolidar, visualizar, aplicar formulações e sintetizar dados de acordo com múltiplas dimensões (CODD et al., 1998).

Um sistema OLAP permite a resposta de consultas multidimensionais usando dados armazenados no *Data Warehouse* (KIMBALL; ROSS, 2013), sendo que as características principais são (BIMONTE, 2016):

- **Consultas Online**: as consultas devem ser feitas *Online*, isto é, em tempo real para o usuário.
- **Consultas Multidimensionais**: Consultas são definidas utilizando as dimensões e medidas providas pelo *Data Warehouse*, que esperam dados de alta qualidade.
- **Representação simples**: os resultados das consultas devem ser representados utilizando tabelas e gráficos, pois os usuários finais geralmente são tomadores de decisão que precisam de visualizações relevantes.
- **Exploratórias**: as consultas são utilizadas em carácter exploratório, pois geralmente os usuários não conhecem de antemão todos os dados disponíveis para consultas.

Cada ferramenta OLAP deve manipular um novo tipo abstrato de dados (TAD), chamado de cubo de dados, utilizando estratégias específicas devido ao modo de como os dados são armazenados, sendo classificadas em (MOREIRA; LIMA, 2012):

- ***Relational OLAP* (ROLAP)**: utilizam Sistemas de Gerenciamento de Banco de Dados (*Data base Management System* - DBMS) relacionais para o gerenciamento e armazenamento dos cubos de dados. Ferramentas ROLAP incluem otimizações para cada DBMS, implementação da lógica de navegação em agregações, serviços e ferramentas adicionais;

- ***Multidimensional* OLAP (MOLAP)**: implementam estruturas de dados multidimensionais para armazenar cubo de dados em memória principal ou em memória externa. Não há utilização de repositórios relacionais para armazenar dados multidimensionais e a lógica de navegação já é integrada a estrutura proposta;
- ***Hybrid* OLAP (HOLAP)**: combinam técnicas ROLAP e MOLAP, onde normalmente os dados detalhados são armazenados em base de dados relacionais (ROLAP), e as agregações são armazenadas em estruturas de dados multidimensionais (MOLAP).

Além desses, existem sistemas OLAP voltados para um domínio ou estilo de dados específico, como é o caso do *Spatial OLAP* (SOLAP), voltado para consultas espaciais (VISWANATHAN; SCHNEIDER, 2014).

É importante ressaltar a diferença entre OLAP e *Online Transaction Processing* (OLPT), visto que sistemas comuns de banco de dados utilizam apenas OLTP, que tem o objetivo de realizar transações e processar consultas online. Isso cobre a grande maioria das operações do dia a dia, como controle de estoque, operações bancárias, etc, servindo a diversos usuários de uma organização. Já o OLAP é utilizado por tomadores de decisão e analistas de dados, sendo voltado para decisões de mais alto nível na organização (HAN et al., 2011).

2.5 Cubo de Dados

O Cubo de Dados originalmente foi criado como um operador relacional que gera todas as combinações possíveis de seus atributos de acordo com uma medida (GRAY et al., 1996).

A estrutura do cubo de dados permite que os dados sejam modelados e visualizados em múltiplas dimensões, e ele é caracterizado por dimensões e medidas. Uma medida é um atributo cujos valores são calculados pelo relacionamento entre as dimensões, sendo que esse é calculado utilizando funções de agragação como soma, quantidade, média, moda, mediana, etc. Uma dimensão é feita pelas entidades que compõe os nossos dados, determinando o contexto do assunto em questão (HAN et al., 2011). Uma dimensão pode ainda ser dividida em membros, que podem ter uma hierarquia, como uma divisão da dimensão tempo em dia, mês e ano.

A organização de um cubo de dados possibilita ao usuário a flexibilidade de visualização dos dados a partir de diferentes perspectivas, já que o operador gera combina-

ções através do conceito do valor *ALL*, onde este conceito representa a agregação de todas as combinações possíveis de um conjunto de valores de atributos. Operações em cubos de dados existem a fim de materializar estas diferentes visões, permitindo busca e análise interativa dos dados armazenados (SILVA, 2015).

Um cubo de dados é composto por células e cada célula possui valores para cada dimensão, incluindo *ALL*, e valores para as medidas. O valor de uma medida é computado para uma determinada célula utilizando níveis de agregação inferiores para gerar os valores dos níveis de agregação superiores na estratégia *Top-down*, com a ordem inversa sendo a *Bottom-up* (SILVA, 2015).

Traduzir exemplo do rodrigo do curso p/ satélites?!

2.5.1 Células do Cubo de Dados

Cubóides?! Imagem to lattice tem que vir aqui também

Diversos subcubos compõe um cubo de dados e cada subcubo é composto por diversas células base e células agregadas. Deste modo uma célula em um subcubo base é uma célula base. Da mesma maneira uma célula em um subcubo não base é uma célula agregada. Uma célula agregada agrega sobre uma ou mais dimensões, onde cada dimensão agregada é indicada pelo valor especial ALL ("*") na notação da célula (LIMA, 2009).

Caso exista um cubo de dados n-dimensional. Seja a = (a 1 , a 2 , a 3 , ..., a n , medidas) uma célula de um dos subcubos que constituem um cubo de dados qualquer. A célula a é uma célula m-dimensional, se exatamente m (m n) valores entre (a 1 , a 2 , a 3 , ..., a n) não são "*". Se m = n, então a é uma célula base, caso contrário, ela é uma célula agregada.

Considere o cubo de dados da Figura 2.1, com as dimensões tempo, departamento e disciplina, e a medida nota. As células (T1, *, *, 78.9) e (*, Ciência da Comp., *, 81.3) são células de 1 dimensão, (T1, *, Calc1, 76.3) é uma célula de 2 dimensões, e (T1, Ciência da Comp., Calc1, 78.8) é uma célula de 3 dimensões. Aqui todas as células base possuem 3 dimensões, enquanto que as células com 1 e 2 dimensões são células agregadas. Um relacionamento de descendente-ancestral pode existir entre células. Em um cubo de dados n-dimensional, uma célula a = (a 1 , a 2 , a 3 , ..., a n , medidas) de nível i é um ancestral de uma célula b = (b 1 , b 2 , b 3 , ..., b n , medidas) de nível j, e b é um descendente de a, se e somente se i $<$ j e 1 m n, onde a m = b m sempre que a m *. Em particular, uma célula a é chamada de pai

de uma célula b, e b de filho de a, se e somente se j = i+1 e b for um descendente de a (HAN; KAMBER, 2006). Com base no mesmo exemplo, uma célula a = (T1, *, *, 78.9) com um membro e uma célula b (T1, *, Calc1, 76.3) com dois membros são ancestrais da célula c = (T1, Ciência da Comp., Calc1, 78.8) que possui três membros, e c é descendente de a e b, onde b é pai de c.

2.5.2 Modelagem dimensional

Existem três esquemas principais para a modelagem dimensional de um cubo de dados: Esquema Estrela (*Star Schema*), Esquema Floco de Neve (*Snowflake Schema*) e Constelação de Fatos (*Fact Constellation Schema*).

2.5.2.1 Esquema estrela

Imagem

Idealizado e criado por Ralph Kimball, o Esquema Estrela é uma forma de dispor as tabelas do modelo relacional para o modelo dimensional (KIMBALL; ROSS, 2002).

Conforme ilustra a Figura 2.5, o Esquema Estrela é uma estrutura com tabelas e ligações bem definidas, baseado no formato de uma estrela. É formado por uma tabela central, denominada tabela de fatos, a qual possui os dados principais da visão da análise, ou seja, o assunto que está sendo analisado, por exemplo, as vendas em dólar, as unidades vendidas, o custo do dólar, etc. Nela ficam ligadas as tabelas de dimensão, que possuem os aspectos pelos quais se deseja observar as medidas relativas ao processo que se está analisando.

2.5.2.2 Esquema Floco de Neve

Imagem

2.5.2.3 Esquema Constelação de Fatos

Imagem

2.5.3 Hierarquias de conceito

Uma hierarquia de conceitos é utilizada para definir uma sequência de mapeamento entre um conjunto de conceitos de baixo nível para um conjunto de conceitos de alto nível, mais gerais. É um estilo de agrupamento e discretização, pois agrupa os valores de modo a reduzir a cardinalidade de uma dimensão (HAN et al., 2011). Elas

ajudam a tornar a análise mais fácil de ser entendida, pois as operações traduzem os dados de baixo nível em uma representação que é mais fácil para o usuário final, assim facilitando a execução das consultas e o seu subsequente uso.

Han, figura 4.10

2.5.4 Medidas

Cada célula de um cubo é definida como um par $\langle (d_1, d_2, ..., d_n), medidas \rangle$, onde $(d_1, d_2, ..., d_n)$ representam as combinações possíveis de valores de atributos sobre as dimensões. Uma medida é calculada para uma certa célula agregando os dados correspondentes a combinação de dimensões e valores (HAN et al., 2011). Medidas podem ser classificadas em três tipos: distributiva, algébrica e holística.

Uma medida distributiva é uma medida cujo cálculo pode ser particionado e depois combinado, e o resultado seria o mesmo se o cálculo fosse executado em todo o conjunto de dados. Por exemplo, a função de soma é distributiva: dividindo os dados N em conjuntos n, e fazendo a soma de cada conjunto n, teremos o mesmo resultado que se a fosse feita diretamente sobre N.

Uma medida algébrica é uma medida cujo cálculo pode ser feito sobre duas ou mais medidas distributivas. Por exemplo, uma medida de média pode ser calculada com a divisão da medida *soma* pela a medida *contagem*, que são ambas distributivas.

Uma medida é holística se não existe uma medida algébrica com M argumentos que caracterize a computação. Isso quer dizer que a computação não pode ser particionada, com valores exatos obtidos apenas se a medida for aplicada em todos os dados. Alguns exemplos são as medidas de moda, desvio padrão e mediana (HAN et al., 2011).

2.5.5 Operações OLAP

Para realizar consultas no *Data Warehouse*, é necessário utilizar de algumas operações sobre o cubo de dados para obter os resultados adequados. Essas consultas também devem conseguir passar na hierarquia de conceitos de cada dimensão, bem como seguir o modelo dimensional do cubo definido, para conseguir oferecer uma interface amigável com o usuário para análise interativa (HAN et al., 2011). Algumas operações comuns são:

- *Roll-up*: realiza agregação no cubo de dados, seja navegando na hierarquia

de conceitos de nível específico para um mais genérico, ou reduzindo uma dimensão.

- *Drill-down*: o inverso da operação *roll-up*, navega na hierarquia de conceitos do nível mais genérico para o nível mais específico, ou adiciona dimensões ao cubo atual. Essa operação visa aumentar o nível de detalhes dos dados.
- *Slice*: ou "fatiamento", realiza uma seleção em uma dimensão do cubo, resultando em um subcubo.
- *Dice*: define um subcubo realizando uma seleção (*slice*) em duas ou mais dimensões.
- *Pivot*: também chamada de rotação, permite mudar a posição das dimensões na visualização, portanto alterando linhas por colunas e vice-versa.

Rodrigo, figura 2.1

Dependendo do sistema OLAP, é possível que outras operações sejam possíveis, como *drill-across* que passa por mais do que uma tabela de fatos, e *drill-through* que permite executar consultas direto na representação em baixo nível do cubo (HAN et al., 2011).

2.5.6 Computação do cubo de dados

A computação do cubo de dados é um problema exponencial em relação ao tempo de execução e ao consumo de memória, portando dada uma relação de entrada **R** com tuplas de tamanho n, a saída é 2^n, onde n é o número de dimensões de um cubo.

To properly calculate a data cube for some measures and dimensions, you have to count the cardinality of each dimension against the cardinality of all other dimensions. While manageable for a few dimensions, this computation becomes almost impossible for cubes with high dimensions as the number of combinations becomes too much for a single computer to handle. This lead to the development of data cube algorithms that optimize for the most relevant measures in the data cube [@silva:2015:abordagensParaCubo].

A *cuboid* is a part of a data cube. For example, if you have three dimensions: temperature, tension and time, a 2-D cuboid could be made from the dimensions

temperature and tension, and a 3-D cuboid would be the same as the full data cube, like figure [ref(fig:datacube)]. The data cube algorithms focus on the computation of cuboids, as the every cube is composed of these smaller cuboids. This leads to the existence of the *curse of dimensionality*, as for n of dimensions there will be 2^n possible cuboid computations, making full materialization very difficult after a few dimensions [@hanDataMiningConcepts2011;@silva:2015:abordagensParaCubo].

For the algorithms, they can be in three different categories: Computing all the cuboids for the data cube leads to a fully materialized data cube; not computing any cuboid beforehand leads to a non-materialized data cube, and partially computing some cuboids leads to partial materialization.

The non-materialized cube has the lowest amount of required memory, but the highest query response time. The fully materialized cube leads to the lowest query response times, as all combinations are already computed, but it needs the highest amount of memory and is thus very hard to compute. As for the partially materialized cube, it is the main issue for most of the algorithms: how to materialize only the most relevant cuboids, and thus achieve a good compromise between memory usage and query performance?

There's some different types of partially materialized cubes, like the *iceberg*, which is a cube with only cells that have passed a certain condition; *shell fragments* compute only cubes with a few dimensions (from 3 to 5) and aggregate those cubes when a bigger number of dimensions is required and *closed cubes* are cubes whose cells with identical measures are grouped into a single abstraction, also called *closed cells*.

To choose which cuboids to materialize, there's a plethora of different algorithms. Two of the classical ones are *Bottom-up* or *Top-down* strategies. Bottom-up starts from the most specific cuboid, called the base cuboid, and goes to the less specific cuboid. Top-down is the inverse: it starts from the least specific cuboid, called the apex cuboid, and goes to the base cuboid. Most of these are tested and overviewed in [@silva:2015:abordagensParaCubo], and won't be repeated here for brevity.

[BUC, Top-down, Bottom-up]

3 TRABALHOS CORRELATOS

Nessa seção, vamos apresentar os trabalhos correlatos a essa proposta. Eles podem ser divididos em duas seções: uma focada nos algoritmos de construção do cubo de dados e outra em como outros operadores estão realizando tarefas semelhantes.

3.1 Computação do Cubo de Dados

O trabalho de (SILVA, 2015) mostra algumas das abordagens de construção do cubo de dados que obtiveram sucessos na litetura, bem como apresenta uma comparação de resultados de algumas das que vamos apresentar aqui. Dentre este trabalho, podemos citar:

- O *FragCubing* (LI et al., 2004) apresenta o conceito de *cube shells*, onde cubóides com poucas dimensões (de 3 a 5 neste exemplo) são calculados utilizando de índices invertidos, que funcionam apenas utilizando memória principal.
- *qCube* (SILVA et al., 2013) extende a abordagem *Frag-Cubing* para permitir consultas sobre intervalos de valor, extendendo os operadores de consultas clássicas em cubo de dados além do operador de igualdade.
- *HFrag* (SILVA et al., 2015) apresenta o uso de memória externa na computação dos índices invertidos, utilizando de um sistema híbrido de memória para armazenar as partições do cubo tanto na memória principal quanto na secundária, com os valores mais frequentes sendo armazenados na memória principal e os valores menos frequentes na memória secundária.
- A abordagem *Hybrid Inverted Cubing* (HIC) (SILVA et al., 2016) extende a abordagem *HFrag* com o parâmetro de frequência acumulada crítica, obtendo resultados melhores do que este nas mesmas consultas.

Utilizandos abordagens diferentes temos:

Precursor para o computação distribuída do cubo, (DOKA et al., 2011) apresenta o *Brown Dwarf*, um sistema *Peer-to-Peer* que permite atualização das células, desenhado para diminuir a redundância em cubos distribuídos.

(HEINE; ROHDE, 2017) apresentam o *PopUp-Cubing*, que utiliza de icebergs para lidar com dados em formato de *stream*, obtendo resultados superiores ao FTL e

Star-Cubing. Este trabalho é de interesse especial por utilizar de dados de *stream*, que permitiriam resultados parecidos com tempo-real, de especial interesse para a operação.

Com foco em *Big Data* e utilizando como base o esquema *MapReduce*, (WANG et al., 2013) apresenta o algoritmo *HaCube* para computação do cubo em paralelo. Este trabalho apresenta um balanço entre computação do cubo em paralelo por vários nós de *MapReduce*, que permite algumas atualizações e computação incremental de medidas. Devido a própria natureza distribuída, ele precisa de mecanismos de tolerância a falha, e também os testes foram executados com no máximo apenas 5 dimensões, porém com até 2,4 bilhões de tuplas. Ainda na linha do *MapReduce*, (YANG; HAN, 2017) demonstra a computação de medidas holísticas apresentando o *Multi-RegionCube*, porém realizando menos testes que o *HaCube*.

Em (ZHAO et al., 2018) é apresentado o *Closed Frag-Shells Cubing*, que utiliza de uma combinação da abordagem de cubos fechados com a abordagem *Shell fragments*, obtendo resultados melhores que a aplicação de cada uma delas separadamente. Essa abordagem é interessante por utilizar de índices *bitmap* e índices invertidos, sendo que lidam com dados altamente dimensionais e sem uma hierarquia de forma similar ao necessário neste trabalho.

3.2 Outros Operadores

A tabela 3.1 mostra uma revisão feita em artigos recentes sobre os operadores de satélite e quais tecnologias eles estão utilizando para atingir objetivos semelhantes, principalmente com o uso de *Big Data*, como demonstrado pelos artigos publicados.

Os objetivos em comum desses trabalhos são facilitar as atividades dos operadores por meio de algoritmos de detecção de anomalias e de verificação dos limites nos valores das telemetrias. Alguns dos operadores dessa lista estão responsáveis pela operação de constelações de satélites complexos, como constelações de sensoriamento remoto, que faz necessário um certo nível de automação ou a operação contínua teria um custo inviável.

Nesses trabalhos, o uso dessas tecnologias é apenas para os operadores de satélite, pois em nenhum desses trabalhos eles estão na mesma estrutura de ingestão dos dados da carga útil, mesmo utilizando as mesmas tecnologias, como demonstrado em (MATEIK et al., 2017) e (ADAMSKI, 2016).

Alguns desses trabalhos não utilizam de estruturas completas que seguem um fluxo

Tabela 3.1 - Operadores e Arquiteturas de Big Data

Referência	Operador	Ferramenta	Tecnologias
(ADAMSKI, 2016)	L3 (EUA)	InControl	Hadoop, Spark, HBase, MongoDB, Cassandra, Amazon AWS
(BOUSSOUF et al., 2018)	Airbus	Dynaworks	Hadoop, Spark, HDFS, HBase, PARQUET, HIVE
(SCHULSTER et al., 2018)	EUMETSAT	CHART	MATLAB, MySQL, Oracle
(ZHANG et al., 2017)	SISET (China)	-	Hadoop, HDFS, PostgreSQL, MongoDB, Logstash, Kibana, ElasticSearch, Kafka, MapReduce
(YVERNES, 2018)	Telespazio France	PDGS	OLAP (DataCube), Saiku, Pentaho, Jaspersoft OLAP
(DISCHNER et al., 2016)	SwRI + NOAA	CYGNSS MOC	SFTP, -
(EDWARDS, 2018)	EUMETSAT	MASIF	FTP, RESTful service, JMS Messague Queue, PostgreSQL
(EVANS et al., 2016)	S.A.T.E + ESA/ESOC	-	Java, CSV
(FEN et al., 2016)	CSMT (China)	-	não menciona as tecnologias
(TROLLOPE et al., 2018)	EUMETSAT	CHART	algoritmos ad-hoc, estudo de caso
(GILLES, 2016)	L-3	InControl	Amazon EC2, LXC, Nagios
(HENNION, 2018)	Thales Alenia	AGYR	Logstash, Kafka, InfluxDB, ElasticSearch, Kibana, Grafana
(MATEIK et al., 2017)	Stinger, NASA	-	Logstash, ElasticSearch, Kibana, HDF5, CSV, R, Python, AWS, Excel
(FERNÁNDEZ et al., 2017)	NASA	MARTE	R, CSV, ad-hoc

Fonte: Produção do autor

de dados, como é o caso de (FERNÁNDEZ et al., 2017) e (TROLLOPE et al., 2018) que utilizam de scripts feitos conforme foram necessários, não mostrando uma visão da arquitetura completa do fluxo de dados e apenas na ferramenta utilizada para análise pontual, ao contrário de (YVERNES, 2018).

O trabalho de (YVERNES, 2018) utiliza de estratégias OLAP e do Cubo de Dados, tendo utilizado uma modelagem dimensional para a operação de uma constelação de satélites, porém esse trabalho menciona apenas em alto nível a modelagem utilizada, e menciona que o trabalho foi somente na parte da modelagem dimensional e integração dos dados utilizando ferramentas já existentes.

3.3 Análise de Dados no INPE

O INPE já realiza análise de dados em outros setores, inclusive sobre as telemetrias de satélite. Os operadores devem monitorar os valores das telemetrias e informar a engenharia caso algum problema que não pôde ser corrigido aparece (TOMINAGA et al., 2017). Um exemplo está no trabalho (AES, 2012) feito sobre uma falha no satélite CBERS-2, onde o modelo proposto visa melhorar o conhecimento sobre avalanche térmica nas baterias para impedir que isso aconteça novamente em outros satélites. A motivação principal dos trabalhos da tabela 3.1 era a detecção de anomalias, que teve alguns algoritmos estudados em (AZEVEDO et al., 2011).

Para os outros setores, isso comumente se dá na análise de dados vindos da carga útil do satélite ou de agentes externos ao INPE, como dados de sensoriamento remoto, cuja análise não é trivial e estão classificados como Big Data. (MONTEIRO, 2017) utilizam de conceitos de Big Data para análise de trajetórias de objetos; (RAMOS et al., 2016) demonstram o uso de softwares como o Hadoop para a análise de dados do clima espacial, com uma arquitetura relacionada as arquiteturas revisadas na seção anterior; e (OES et al., 2018) mostra uma arquitetura que utiliza de Cubo de Dados para a análise de séries temporais.

4 PROPOSTA

Nesta seção apresentamos os dados que serão utilizados e são relevantes para a operação, bem como apresentamos o fluxo de dados utilizado nos trabalhos correlatos e a arquitetura proposta, por fim explicando o cubo de dados que será implementado.

4.1 Dados

A tabela 4.1 mostra os tipos de dados relevantes para a operação, a sua origem e o seu formato esperado, ignorando os dados provenientes da carga útil.

Tabela 4.1 - Dados de Operação

Tipo de Dado	**Origem**	**Formato**
Sensores de bordo	Equipamentos no satélite	Tabelas, CSV
Registros do Computador	Computador de Bordo	Texto (*Logs*)
Multimídia	Câmeras	MP4, JPG, RAW
Parâmetros orbitais	Operação, Rastreio	TLE, texto, tabelas
Documentação associada	Operadores, engenharia	Texto (Word, Excel)
Clima Espacial	Sensores no solo ou espaço	Texto, tabelas, avisos
Situational Awareness	Radares, US-STRACOM, etc	Texto, tabelas, avisos

Fonte: Adaptado de (ZHANG et al., 2017)

Para este trabalho, apenas os dados vindos de sensores de bordo serão considerados. Os outros dados nesta tabela poderiam ser considerados para uma *Data Warehouse* mais completa, pois um cubo de dados pode ser formado sobre quaisquer um desses dados.

Por exemplo, um cubo de dados textual poderia ser feito sobre os documentos associados a operação, como o CONOPS, tabelas de telecomandos e documentação de engenharia de sistemas para facilitar a análise da documentação sendo gerada pelo satélite. Um cubo multimídia poderia ser gerado sobre os dados multimídia tirados pelas câmeras do satélite para correlacionar com os dados gerados pelos sensores, e assim em diante. Alguns exemplos de cubos possíveis de serem feitos estão em (SILVA, 2015).

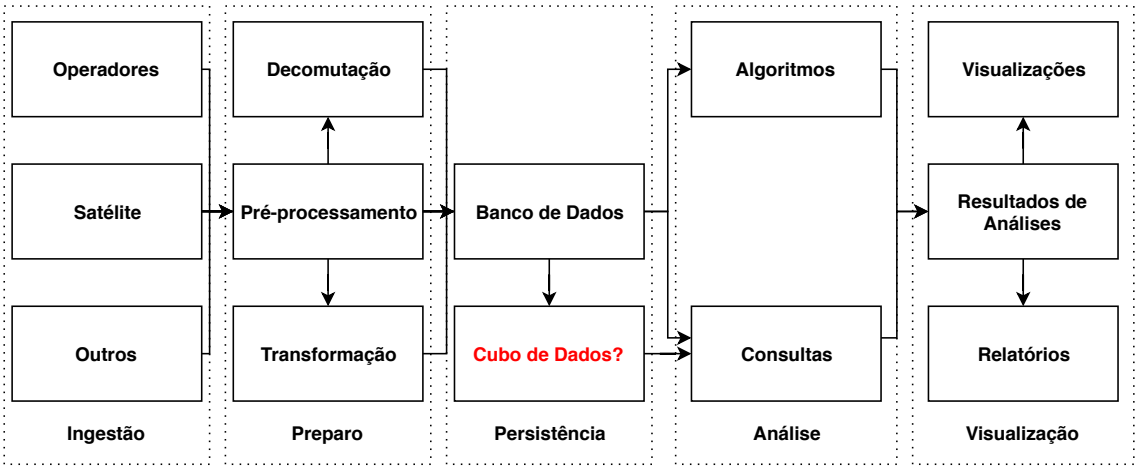
Esta lista não é exaustiva, e pode incluir dados da carga útil caso sejam relevantes para a análise em questão, como ajudar na georeferênciação de imagens tiradas pelo

satélite, bem como outros dados que os operadores acharem necessários e tiverem acesso a eles. Isso vai depender diretamente do fluxo de dados que é adotado, pois esses dados precisam ser coletados e preparados para serem utilizados.

4.2 Fluxo dos Dados de Operação

Baseado nos trabalhos correlatos e nos dados levantados na seção 4.1, a figura 4.1 demonstra o fluxo de dados esperado de uma arquitetura de *Big Data* para a operação de satélites.

Figura 4.1 - Fluxo de dados em uma arquitetura de Big Data



Fonte: Adaptado de (ZHANG et al., 2017)

.

Este fluxo está separado em cinco etapas que vão desde a origem dos dados até o seu resultado de análise, e este trabalho visa apenas mapear qual seria esse fluxo baseado nos trabalhos correlatos. Cada uma das etapas está detalhada a seguir:

- **Ingestão**: onde os dados serão coletados na sua fonte (satélites, sensores no solo, outras fontes, etc). Essa etapa se importa em **onde** estão os dados e **como** coletá-los, bem como **quais** são os dados importantes de serem coletados. A "fonte"aqui pode ser um serviço de terceiros, dentro da própria instituição ou disponível de outra forma.
- **Preparo**: os dados relevantes são selecionados, e transformações são reali-

zadas para inserir os mesmos na base de dados. Essa etapa se importa no formato específico dos dados, realizando operações de limpeza, verificação da qualidade e da relevância para a análise, entre outras. O seu objetivo é garantir que os dados tem qualidade, relevância, e estão no formato adequado para a base de dados.

- **Persistência**: após o devido processamento, os dados de alta qualidade são guardados em uma base de dados, de onde ficarão disponíveis para a análise. Nessa etapa um banco de dados é utilizado, se importando apenas em como esses dados estão guardados e como eles serão disponibilizados para as consultas e execução de algoritmos.

- **Análise**: nesta etapa são executadas as consultas e os algoritmos de interesse para a análise. Podem ser desde consultas simples ("qual era o valor da telemetria X durante a passagem Y?"), a execução de algoritmos complexos ("preveja os valores da telemetria X para a próxima passagem").

- **Visualização**: os resultados das consultas e algoritmos são visualizados. Podem conter desde gráficos simples, como um histograma de uma telemetria, a relatórios complexos de um subsistema/satélite, bem como resultados de algoritmos.

Esse fluxo é geralmente seguido pelos trabalhos correlatos na seção 3.2, sendo que os trabalhos de (ZHANG et al., 2017), (MATEIK et al., 2017) e (BOUSSOUF et al., 2018) definem esse processo mais claramente.

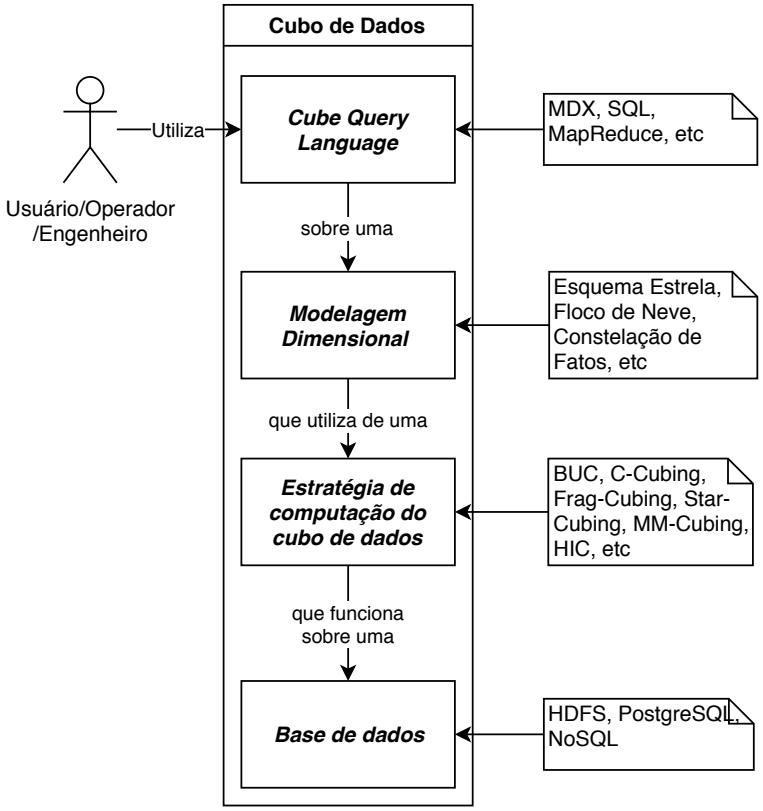
4.3 Arquitetura de um Cubo de Dados

A figura 4.2 demonstra a divisão em 4 camadas de uma estrutura de Cubo de Dados. Essas camadas demonstram tudo o que é necessário para a implementação de um Cubo de Dados, não sendo necessário que uma camada esteja fortemente atrelada a outra.

Para esta proposta, vamos nos concentrar apenas na proposição de um algoritmo de computação do cubo de dados mais apropriado, utilizando das outras seções quando elas vão se tornando necessárias. Os detalhes, algoritmos e conceitos listados na figura estão majoritariamente descritos na seção 2.5.

Uma informação interessante é que esta estrutura mostra o uso de pelo menos duas linguagens de computação sobre os dados: uma é a Cube Query Language que será

Figura 4.2 - Estruturado Cubo de dados



Fonte: Produção dos autores

.

utilizada pelo usuário para realizar as operações sobre o cubo (Drill-Down, Roll-up, etc), e outra é a linguagem que será utilizada pelo cubo para realizar essas operações, e elas podem ser independentes, por exemplo, pode-se utilizar SQL extendida com vocabulário de OLAP, porém o algoritmo de cubo de dados internamente pode consultar uma estrutura feita com MapReduce para o cálculo das medidas e das agregações.

Porém, utilizar duas linguagens muito diferentes nesse ponto pode não ser uma boa ideia, pois adicionaria um nível de diferença entre o usuário e os dados. Caso seja necessário realizar uma consulta OLTP normal, sem o uso do cubo de dados, por exemplo, essa diferença ficaria mais óbvia, por exemplo traduzir uma consulta de SQL para MapReduce não seria muito fácil simplesmente por ter que entender de ambas as linguagens bem para conseguir fazer isso. Deste modo, é interessante manter a mesma linguagem ao longo da estrutura, apenas alterando nas operações

relevantes para o cubo de dados.

Com isso se torna necessário ressaltar o último nível, a Base de Dados: a escolha de banco de dados vai impactar como o algoritmo funciona, visto que existem diferentes sistemas de arquivos e como eles são atingidos, bem como o estilo do banco vai mudar como o algoritmo deve gerar o cubo, pois a base pode utilizar diferentes paradigmas de banco de dados (CUZZOCREA et al., 2013).

4.3.1 Algoritmos de construção do cubo

Uma das necessidades de usar algoritmos diferentes de cubo de dados está no número de dimensões que um certo cubo consegue realizar pesquisas: consultas com mais que 15 dimensões não são comummente(ou praticamente) executadas em alguns algoritmos, como o trabalho de (SILVA, 2015) demonstra.

Como estabelecido na seção 4.1, os dados de telemetria de interesse possuem muito mais do que o limite de consultas em até 60 dimensões: com mais de 130 telemetrias para os satélites da família da SCD, e milhares para satélites maiores, a execução de consultas complexas seria normalmente inviável nos algortimos de construção do cubo. Esse problema é geralmente resolvido pela modelagem dimensional, como em (AZEVEDO; AMBRÓSIO, 2010), porém isso transforma os dados de um formato "largo"para um formato "longo", aumentando o número de tuplas.

Deste modo é necessário investigar as abordagens de construção de cubo que funcionem com muitas dimensões, e que permitam o cálculo das medidas necessárias para a operação. Neste trabalho, iremos investigar algumas dessas abordagens e como elas se comparam para a análise.

4.4 Discussão

TODO: Passos do que seria feito? Estou achando melhor deixar isso no final, nas conclusões...

5 RESULTADOS PRELIMINARES

Para a implementação da arquitetura, é necessário conhecimento sobre o domínio dos dados e como eles estão organizados. Para esse estágio da pesquisa, foram utilizados dados do SCD-2 fornecidos pelo CCS, porém alguns softwares foram implementados como parte da pesquisa sobre o cubo de dados e como resultados da análise dos dados.

5.1 SCD-Dashboard

Os dados vindos do SCD2 possuem 135 dimensões e estão distribuídos em um período de 4 anos, tornando a sua análise não trivial. Durante atividades de *Data Science* era necessário criar muitos gráficos e visualizações com períodos e dimensões variadas, e com tantas dimensões isso levava vários pedaços de código copiados para criar relatórios sobre os dados e dimensões.

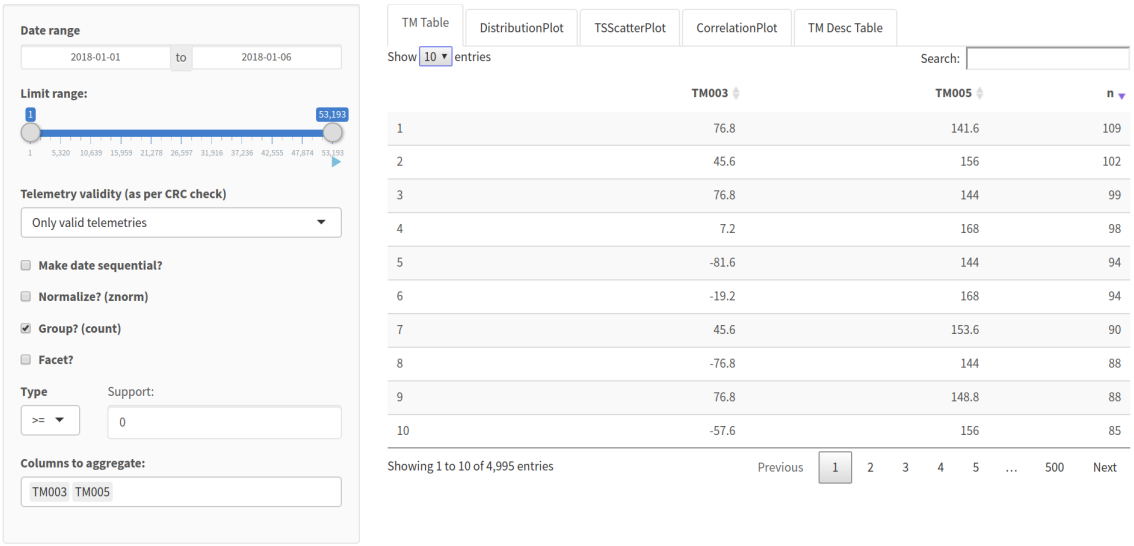
Para facilitar essa análise e automatizar esse problema, foi criado um software chamado **SCD-dashboard** para facilitar na visualização dos dados de telemetria. Ele foi implementado utilizando o pacote Shiny (CHANG et al., 2019) da linguagem R (R Core Team, 2018), que já estava sendo utilizada para criar as visualizações, a interface permite apenas criar as visualizações e aplicar os outros algoritmos relacionados de uma interface amigável e sem ter a necessidade de escrever código.

Ele funciona sobre um banco de dados PostgreSQL que possui o histórico das telemetrias já importadas e propriamente transformadas por outro script feito só para os dados do SCD2, porém devem funcionar para qualquer dado exportado via CSV pelo SatCS.

Essa ferramenta também pode ser vista como um piloto para implementar algo semelhante as *dashboards* criadas por outras agências, sendo que tem um precedente na ferramenta MARTE utilizada pela NASA (FERNÁNDEZ et al., 2017), que utiliza das mesmas tecnologias e conceitos, porém focada no algoritmo de detecção de anomalias. Ferramentas mais parecidas, e mais completas, estão no CHART e nas dashboard criadas usando o Kibana em (MATEIK et al., 2017) e (ZHANG et al., 2017).

Essa ferramenta foi utilizada extensamente para visualização dos dados e para análise exploratória, sendo que foi melhorada durante as disciplinas de *Data Science* e Algoritmos de *Data Mining*, com apresentações da mesma feitas utilizando ela.

Figura 5.1 - Aparência do SCD-Dashboard



Fonte: Produção dos autores

5.2 RFragCubing

O algoritmo FragCubing, apresentado na seção 3.1, foi disponibilizado em forma compilada via código de C++, porém, ele foi feito com uma interface complicada de ser utilizada e automatizada, e não permitia a importação dos dados de telemetria sem um trabalho de pré-processamento não trivial antes. Como essa implementação foi a utilizada no trabalho de (SILVA, 2015), os resultados possuem histórico para comparação, portanto o seu uso contínuo seria interessante.

Para isso, foi criado um pacote na linguagem R chamado de **RFragCubing**, que permite a integração com o algoritmo implementado em C++. Esse pacote faz a interface com o código do FragCubing, permitindo importar os dados, executar as queries e retornar os resultados das mesmas, bem como algumas adições de medidação de memória e tempo de processamento. Uma das vantagens do pacote está na distribuição, com uma interface que permite a execução da mesma consulta em outros algoritmos de construção do cubo, uma das razões para fazer a implementação, para facilitar a comparação entre os algoritmos.

Figura 5.2 - Carregando dados de telemetria no RFragCubing

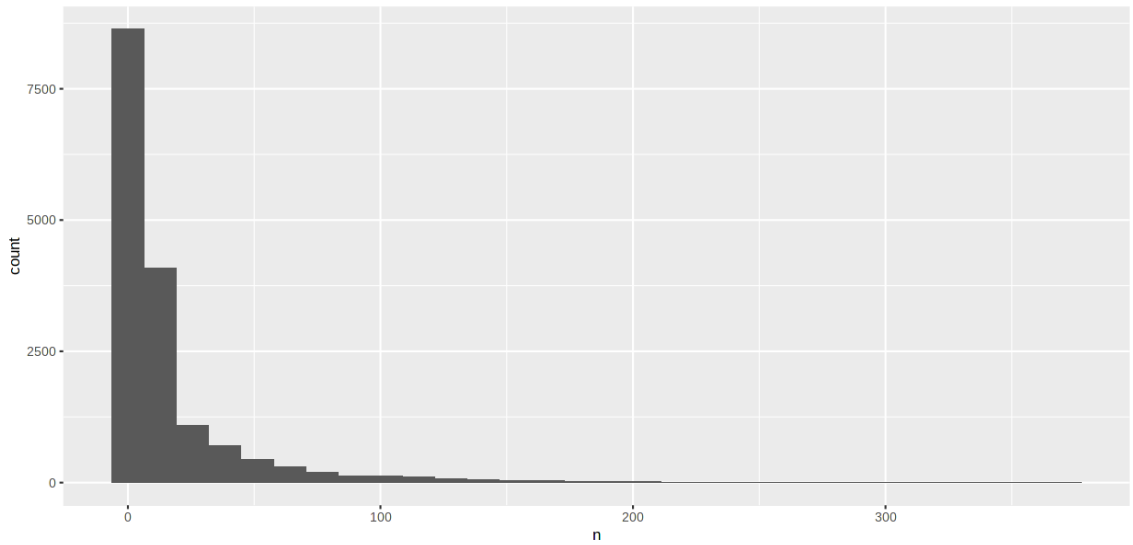
```
Initializing ... used time: 26 ms.
Reading input file ... used time: 1654 ms.
Computing shell fragments ... used time: 1681 ms.

135 dimensional data loaded.
296909 tuples read.
135 shell fragments of size 1 constructed.
```

Fonte: Produção dos autores

5.3 Medida de Similaridade

Utilizando a ferramenta criada em 5.1 para uma atividade de exploração de dados que envolvia realizar consultas multidimensionais baseadas no conceito do cubo de dados, um padrão foi notado entre as telemetrias: quando uma medida de agregação por contagem era executada sobre telemetrias que se sabia ter algum tipo de relacionamento entre elas, elas possuiam uma curva característica, como da figura 5.3.

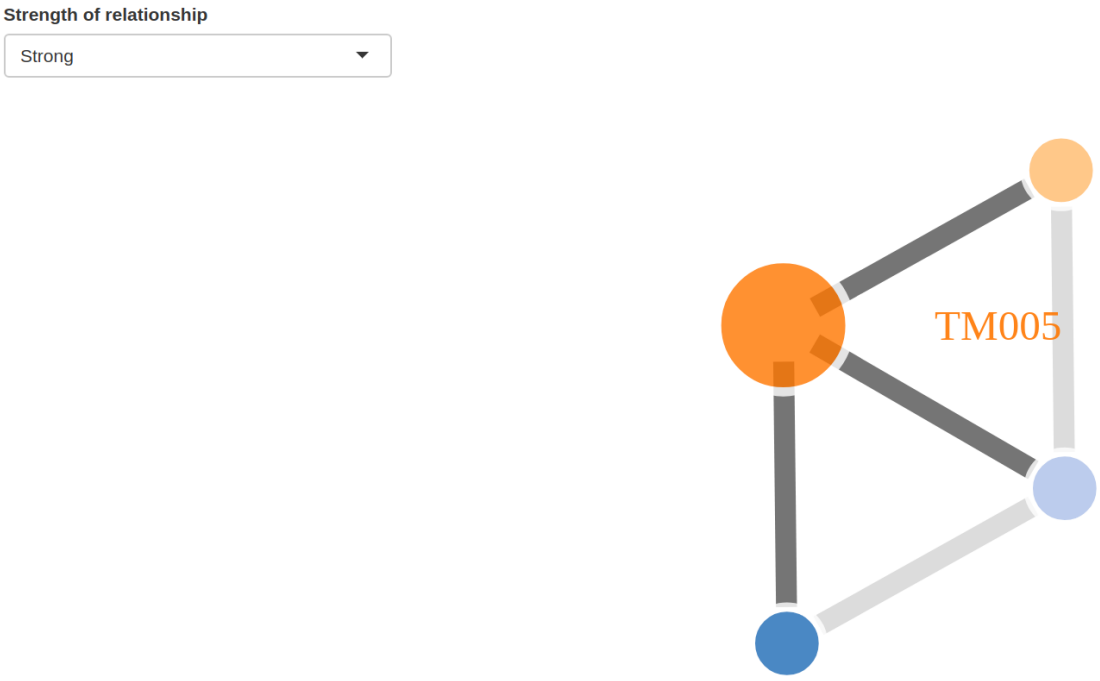
Figura 5.3 - Curva de de agregação gerada pela SCD-Dashboard



Agregação das telemetrias TM003 e TM004 ao longo de 01/2018
Fonte: Produção dos autores

Essa curva foi utilizada para desenvolver um algoritmo de classificação do relacionamento entre as telemetrias, pois possuia algumas vantagens: a medida de contagem dos valores que se repetem é independente do tipo dos valores em si, permitindo comparar um valor contínuo com um discreto bem como discretos com discretos e contínuos com contínuos, de equipamentos diferentes e com características diferentes; e é uma medida que funciona com qualquer número de dimensões, sendo uma operação $\Theta(D)$, com D sendo o número de dimensões, tornando a sua integração, e subsequente otimização, com um algoritmo de construção de cubo de dados simplificada.

Figura 5.4 - Resultados preliminares da Medição de Relacionamento



Um grupo de telemetrias com relacionamento forte
Fonte: Produção dos autores

O algortimo está sujeito a Maldição de Dimensionalidade, pois ele verifica a relação entre um número n de telemetrias definida pelo usuário, tendo 2^D possíveis combinações para D dimensões, assim ele atualmente só foi executado para todas as combinações de 2 dimensões dos dados do SCD. Mesmo assim, e por ser uma operação demorada mesmo com apenas um ano de telemetrias, ainda são 8911 com-

binações. Esse número só cresce com o número de combinações, sendo que a execução de todas as combinações com mais de 4 dimensões é inviável.

O algoritmo está sendo revisado atualmente, mas o objetivo é implementar o resultado no algoritmo de construção do cubo, sendo que as dimensões que tem um relacionamento associado como forte ou média seriam as combinações com maior interesse pelos operadores, pois provavelmente seriam as operações mais comumemente executadas por eles.

Isto também abre espaço para um algoritmo de detecção de anomalias: caso durante uma passagem um relacionamento que antes era tido como forte muda para um relacionamento fraco ou perde o relacionamento, isso pode ser um sinal de que alguma coisa está de errado com esse grupo de telemetrias. Isso pode ajudar no conhecimento dos operadores, pois os relacionamentos entre grupos de telemetrias são difíceis de serem visualizados e apenas os operadores com mais experiência em um dado satélite conseguem visualizar alguns desses relacionamentos. Citação?

Existem alguns trabalhos de visualização dos relacionamentos, porém estes geralmente utilizam algoritmos de indução de regras, cuja saída é de difícil interpretação, como demonstrado em (KANNAN; DEVI, 2016). A abordagem aqui proposta tem potencial de ter respostas mais relevantes para os operadores, porém precisa ser avaliada por um operador ainda, um dos próximos passos deste trabalho.

6 CONCLUSÃO E TRABALHOS FUTUROS

Este trabalho apresenta uma abordagem de cubo de dados para executar operações de análise nos dados de telemetrias de satélites. Essa abordagem utiliza de conceitos de *Big Data* para orientar a execução de consultas em dados com muitas dimensões e de alta complexidade. Uma revisão da literatura de arquiteturas de *Big Data* é apresentada, demonstrando que tipos de tecnologias e abordagens estão em uso por outros operadores de satélite, bem como uma revisão sobre os conceitos e abordagens de cubo de dados.

São apresentados resultados intermediários de análises dos dados de telemetria, bem como os produtos de software utilizados para realizar as análises e algumas descobertas desse processo. Esses resultados mostram que a aplicação do cubo é interessante para os dados disponíveis, bem como que é possível implementar algumas etapas do fluxo de dados, propriamente adequadas para lidar com *Big Data*.

Como essa arquitetura é melhor(diferente?) da utilizada por outros operadores? O que o Cubo de Dados traz de diferente?

6.1 Planejamento

Para o trabalho da dissertação, os passos seguintes são:

a) Formalizar quais são as consultas relevantes para os operadores de satélite, e quais são as atividades de análise que podem ser expressas como consultas;

b) Criar uma representação dimensional do cubo de dados apropriada para as consultas identificadas, mapeando as medidas que são necessárias e quais os seus tipos;

c) Implementar a representação com as medidas em vários algoritmos da literatura recente, mais notadamente os revisados no capítulo 3.1 e coletar os resultados da execução das consultas relevantes para os operadores;

d) Avaliar os resultados da implementação dos algoritmos e mostrar qual das abordagens é mais apropriada para o cenário da operação.

O passo *c* está parcialmente implementado no pacote mostrado na seção 5.2, porém precisa de trabalho significativo de implementação para executar outros algoritmos e realizar os testes das consultas relevantes.

Como resultado esperados da dissertação, teríamos o mapeamento das consultas que são relevantes para os operadores de satélite, e a sua representação em um cubo de dados, que teria resultados de implementações diferentes para conseguir avaliar qual dos algortimos disponíveis é o mais adequado para o cenário de operação.

6.1.1 Trabalhos futuros

Como trabalhos futuros, a arquitetura do fluxo de dados mostrada na seção 4.2 pode ser implementado nos moldes das outras agências a exemplo de 3.2, numa arquitetura que permita a inclusão de todos os tipos de dados elecandos neste trabalho na seção 4.1. Também seria interessante expandir os tipos de algoritmos que serão testados para esse trabalho.

Expandir o uso dos dados para a abordagem de cubo seria relevante, pois existem desafios diferentes quando se lida com um satélite de tamanho grande (ex. GEO) e uma constelação de satélites menores, porém com um volume de dados comparável. Alguma abordagem para lidar com dados de CubeSats seria relevante para o momento, principalmente se forem de diferentes cubesats e/ou de constelações.

REFERÊNCIAS BIBLIOGRÁFICAS

ADAMSKI, G. Data Analytics for Large Constellations. In: **SpaceOps 2016 Conference**. [S.l.]: American Institute of Aeronautics and Astronautics, 2016. (SpaceOps Conferences). 00000. 14, 15

AES, R. O. de M. **Estudo de avalanche térmica em um sistema de carga e descarga de bateria em satélites artificiais**. 00000. Tese (Doutorado) — Instituto Nacional de Pesquisas Espaciais, São José dos Campos, fev. 2012. Acesso em: 01 ago. 2018. 16

AZEVEDO, D. N. R.; AMBRÓSIO, A. M. Dependability in Satellite Systems: An Architecture for Satellite Telemetry Analysis. In: Workshop em Engenharia e Tecnologia Espaciais, 1. (WETE)., 30 mar. - 1 abr. 2010, São José dos Campos. **Anais...** São José dos Campos: INPE, 2010. IWETE2010-1065, p. 6. ISSN 2177-3114. 00000. Acesso em: 30 jul. 2018. 1, 21

AZEVEDO, D. N. R.; AMBRÓSIO, A. M.; VIEIRA, M. **Estudo sobre técnicas de detecção automática de anomalias em satélites**. Tese (Doutorado) — Instituto Nacional de Pesquisas Espaciais, São José dos Campos, 2011. Acesso em: 10 ago. 2018. 16

BIMONTE, S. Open issues in Big Data Warehouse design. **Revue des Nouvelles Technologies de l'Information**, p. 10, 2016. 3, 4, 5

BOUSSOUF, L.; BERGELIN, B.; SCUDELER, D.; GRAYDON, H.; STAMMINGER, J.; ROSNET, P.; TAILLEFER, E.; BARREYRE, C. Big Data Based Operations for Space Systems. In: **2018 SpaceOps Conference**. [S.l.]: American Institute of Aeronautics and Astronautics, 2018. 00000. 3, 15, 19

CHANG, W.; CHENG, J.; ALLAIRE, J. J.; XIE, Y.; MCPHERSON, J. **Shiny: Web Application Framework for R**. [S.l.: s.n.], 2019. 23

CODD, E. F.; CODD, S.; SALLEY, C. Providing olap to user-analysts: An it mandate. In: . [S.l.: s.n.], 1998. 5

CUZZOCREA, A.; BELLATRECHE, L.; SONG, I.-Y. Data Warehousing and OLAP over Big Data: Current Challenges and Future Research Directions. In: **Proceedings of the Sixteenth International Workshop on Data Warehousing and OLAP**. New York, NY, USA: ACM, 2013. (DOLAP '13), p. 67–70. ISBN 978-1-4503-2412-0. 00000. 21

DISCHNER, Z.; REDFERN, J.; ROSE, D.; ROSE, R.; RUF, C.; VINCENT, M. CYGNSS MOC; Meeting the challenge of constellation operations in a cost-constrained world. In: **2016 IEEE Aerospace Conference**. [S.l.: s.n.], 2016. p. 1–8. 00000. 15

DOKA, K.; TSOUMAKOS, D.; KOZIRIS, N. Brown Dwarf: A fully-distributed, fault-tolerant data warehousing system. **Journal of Parallel and Distributed Computing**, v. 71, n. 11, p. 1434–1446, nov. 2011. ISSN 0743-7315. 00000. 13

EDWARDS, T. Dealing with the Big Data - The Challenges for Modern Mission Monitoring and Reporting. In: **15th International Conference on Space Operations**. Marseille, France: American Institute of Aeronautics and Astronautics, 2018. ISBN 978-1-62410-562-3. 00000. 15

EVANS, D. J.; MARTINEZ, J.; Korte-Stapff, M.; VANDENBUSSCHE, B.; ROYER, P.; RIDDER, J. D. Data Mining to Drastically Improve Spacecraft Telemetry Checking: A Scientist?s Approach. In: **SpaceOps 2016 Conference**. [S.l.]: American Institute of Aeronautics and Astronautics, 2016, (SpaceOps Conferences). 00000. 15

FEN, Z.; YANQIN, Z.; CHONG, C.; LING, S. Management and Operation of Communication Equipment Based on Big Data. In: **2016 International Conference on Robots Intelligent System (ICRIS)**. [S.l.: s.n.], 2016. p. 246–248. 00000. 15

FERNÁNDEZ, M. M.; YUE, Y.; WEBER, R. Telemetry Anomaly Detection System Using Machine Learning to Streamline Mission Operations. In: **2017 6th International Conference on Space Mission Challenges for Information Technology (SMC-IT)**. [S.l.: s.n.], 2017. p. 70–75. 00003. 15, 16, 23

GILLES, K. Flying Large Constellations Using Automation and Big Data. In: **SpaceOps 2016 Conference**. [S.l.]: American Institute of Aeronautics and Astronautics, 2016, (SpaceOps Conferences). 00000. 15

GRAY, J.; BOSWORTH, A.; LYAMAN, A.; PIRAHESH, H. Data cube: A relational aggregation operator generalizing GROUP-BY, CROSS-TAB, and SUB-TOTALS. In: . [S.l.]: IEEE Comput. Soc. Press, 1996. p. 152–159. ISBN 978-0-8186-7240-8. 03145. 6

HAN, J.; KAMBER, M.; PEI, J. **Data Mining: Concepts and Techniques, Third Edition**. 3 edition. ed. Haryana, India; Burlington, MA: Morgan Kaufmann, 2011. 00006. ISBN 978-93-80931-91-3. 4, 6, 8, 9, 10

HEINE, F.; ROHDE, M. PopUp-Cubing: An Algorithm to Efficiently Use Iceberg Cubes in Data Streams. In: **Proceedings of the Fourth IEEE/ACM International Conference on Big Data Computing, Applications and Technologies**. New York, NY, USA: ACM, 2017. (BDCAT '17), p. 11–20. ISBN 978-1-4503-5549-0. 00000. 13

HENNION, N. Big-data for satellite yearly reports generation. In: **2018 SpaceOps Conference**. [S.l.]: American Institute of Aeronautics and Astronautics, 2018. 00000. 15

INMON, W. H.; HACKATHORN, R. D. **Using the Data Warehouse**. Somerset, NJ, USA: Wiley-QED Publishing, 1994. ISBN 978-0-471-05966-0. 4

KANNAN, S. A.; DEVI, T. Mining satellite telemetry data: Comparison of rule-induction and association mining techniques. In: **2016 IEEE International Conference on Advances in Computer Applications (ICACA)**. [S.l.: s.n.], 2016. p. 259–264. 00000. 27

KIMBALL, R.; ROSS, M. **The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling**. Edição: 3rd. Indianapolis, IN: John Wiley & Sons, 2013. ISBN 978-1-118-53080-1. 5

LI, X.; HAN, J.; GONZALEZ, H. High-dimensional OLAP: A Minimal Cubing Approach. In: **Proceedings of the Thirtieth International Conference on Very Large Data Bases - Volume 30**. [S.l.]: VLDB Endowment, 2004. (VLDB '04), p. 528–539. ISBN 978-0-12-088469-8. 13

MATEIK, D.; MITAL, R.; BUONAIUTO, N. L.; LOUIE, M.; KIEF, C.; AARESTAD, J. Using Big Data Technologies for Satellite Data Analytics. In: . [S.l.]: American Institute of Aeronautics and Astronautics, 2017. ISBN 978-1-62410-483-1. 00001. 14, 15, 19, 23

MONTEIRO, D. V. **A FRAMEWORK FOR TRAJECTORY DATA MINING**. Tese (Doutorado), 2017. 16

MOREIRA, A. A.; LIMA, J. d. C. Full and partial data cube computation and representation over commodity PCs. In: **2012 IEEE 13th International Conference on Information Reuse Integration (IRI)**. [S.l.: s.n.], 2012. p. 672–679. 5

OES, R. E. d. O. S.; CAMARA, G.; QUEIROZ, G. R. de. Sits: Data analysis and machine learning using satellite image time series. In: Workshop de Computação

Aplicada, 18. (WORCAP), 21-23 ago., São José dos Campos, SP. **Resumos...** [S.l.], 2018. p. 18. Acesso em: 02 maio 2019. 16

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2018. 00000. 23

RAMOS, M. P.; TASINAFFO, P. M.; de Almeida, E. S.; ACHITE, L. M.; da Cunha, A. M.; DIAS, L. A. V. Distributed Systems Performance for Big Data. In: LATIFI, S. (Ed.). **Information Technology: New Generations**. [S.l.]: Springer International Publishing, 2016, (Advances in Intelligent Systems and Computing). p. 733–744. ISBN 978-3-319-32467-8. 16

SCHULSTER, J.; EVILL, R.; PHILLIPS, S.; FELDMANN, N.; ROGISSART, J.; DYER, R.; ARGEMANDY, A. CHARTing the Future – An offline data analysis and reporting toolkit to support automated decision-making in flight-operations. In: **15th International Conference on Space Operations**. Marseille, France: American Institute of Aeronautics and Astronautics, 2018. ISBN 978-1-62410-562-3. 00001. 15

SILVA, R. R. **Abordagens para Cubo de Dados Massivos com Alta Dimensionalidade Baseadas em Memória Principal e Memória Externa: HIC e BCubing**. 00000. Tese (Doutorado) — Instituto Tecnológico de Aeronáutica, São José dos Campos, 2015. Acesso em: 01 ago. 2018. 5, 7, 13, 17, 21, 24

SILVA, R. R.; HIRATA, C. M.; LIMA, J. d. C. A Hybrid Memory Data Cube Approach for High Dimension Relations. In: HAMMOUDI, S.; MACIASZEK, L. A.; TENIENTE, E. (Ed.). **ICEIS 2015 - Proceedings of the 17th International Conference on Enterprise Information Systems, Volume 1, Barcelona, Spain, 27-30 April, 2015**. [S.l.]: SciTePress, 2015. p. 139–149. ISBN 978-989-758-096-3. 13

_____. Computing BIG data cubes with hybrid memory. p. 18, 2016. 13

SILVA, R. R.; LIMA, J. d. C.; HIRATA, C. M. qCube: Efficient integration of range query operators over a high dimension data cube. **JIDM**, v. 4, n. 3, p. 469–482, 2013. 13

TOMINAGA, J.; FERREIRA, M. G. V.; AMBRÓSIO, A. M. Comparing satellite telemetry against simulation parameters in a simulator model reconfiguration tool. In: CERQUEIRA, C. S.; BÜRGER, E. E.; YASSUDA, I. d. S.; RODRIGUES,

I. P.; LIMA, J. S. d. S.; OLIVEIRA, M. E. R. de; TENÓRIO, P. I. G. (Ed.). **Anais...** São José dos Campos: Instituto Nacional de Pesquisas Espaciais (INPE), 2017. ISSN 2177-3114. 00000. Acesso em: 30 jul. 2018. 16

TROLLOPE, E.; DYER, R.; FRANCISCO, T.; MILLER, J.; GRISO, M. P.; ARGEMANDY, A. Analysis of automated techniques for routine monitoring and contingency detection of in-flight LEO operations at EUMETSAT. In: **2018 SpaceOps Conference**. Marseille, France: American Institute of Aeronautics and Astronautics, 2018. ISBN 978-1-62410-562-3. 00001. 15, 16

VISWANATHAN, G.; SCHNEIDER, M. User-centric spatial data warehousing: A survey of requirements and approaches. **International Journal of Data Mining, Modelling and Management**, v. 6, n. 4, p. 369, 2014. ISSN 1759-1163, 1759-1171. 00004. 6

WANG, Z.; CHU, Y.; TAN, K.-L.; AGRAWAL, D.; ABBADI, A. E.; XU, X. Scalable Data Cube Analysis over Big Data. **arXiv:1311.5663 [cs]**, nov. 2013. 00021. 14

YANG, H.; HAN, C. Holistic and Algebraic Data Cube Computation Using MapReduce. In: **2017 9th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)**. [S.l.: s.n.], 2017. v. 2, p. 47–50. 00000. 14

YVERNES, A. Copernicus Ground Segment as a Service: From Data Monitoring to Performance Analysis. In: **15th International Conference on Space Operations**. Marseille, France: American Institute of Aeronautics and Astronautics, 2018. ISBN 978-1-62410-562-3. 00000. 15, 16

ZHANG, X.; WU, P.; TAN, C. A big data framework for spacecraft prognostics and health monitoring. In: **2017 Prognostics and System Health Management Conference (PHM-Harbin)**. [S.l.: s.n.], 2017. p. 1–7. 00000. 4, 15, 17, 18, 19, 23

ZHAO, Q.; ZHU, Y.; WAN, D.; TANG, S. A Closed Frag-Shells Cubing Algorithm on High Dimensional and Non-Hierarchical Data Sets. In: **Proceedings of the 12th International Conference on Ubiquitous Information Management and Communication**. New York, NY, USA: ACM, 2018. (IMCOM '18), p. 6:1–6:8. ISBN 978-1-4503-6385-3. 00000. 14

ANEXO A - CRONOGRAMA E PUBLICAÇÕES

A tabela mostra o cronograma esperado para as próximas atividades do mestrado.

Tabela A.1 - Cronograma de atividades

Atividade	maio	jun.	jul.	ago.	set.	out.	nov.	dec.	jan.	fev.
Exame de Proposta	X									
Submissão Artigo Periódico					X					
Apresentação Conferência						X				
Defesa final							X			X

A tabela mostra os veículos de publicação planejados e os já publicados/em processo de publicação.

Tabela A.2 - Publicações planejadas

Nome	**Qualis**	**Prazo**	**Notas**
WETE 2018	Conferência		Publicado
IAC 2019	Conferência		**Artigo aceito, porém feito pela matéria do Prof. Geilson (não é nessa área), porém sou o 1 autor**
BDCAT	Conferência	29/08/2019	Nova Zelândia...
WETE 2019	Conferência		?
IEEE América Latina	B2		?
International Journal of Data Warehousing and Mining	B1 (CC)		Não tem qualis para ENG-III, caro, JCR 0,66

PUBLICAÇÕES TÉCNICO-CIENTÍFICAS EDITADAS PELO INPE

Teses e Dissertações (TDI)

Teses e Dissertações apresentadas nos Cursos de Pós-Graduação do INPE.

Manuais Técnicos (MAN)

São publicações de caráter técnico que incluem normas, procedimentos, instruções e orientações.

Notas Técnico-Científicas (NTC)

Incluem resultados preliminares de pesquisa, descrição de equipamentos, descrição e ou documentação de programas de computador, descrição de sistemas e experimentos, apresentação de testes, dados, atlas, e documentação de projetos de engenharia.

Relatórios de Pesquisa (RPQ)

Reportam resultados ou progressos de pesquisas tanto de natureza técnica quanto científica, cujo nível seja compatível com o de uma publicação em periódico nacional ou internacional.

Propostas e Relatórios de Projetos (PRP)

São propostas de projetos técnico-científicos e relatórios de acompanhamento de projetos, atividades e convênios.

Publicações Didáticas (PUD)

Incluem apostilas, notas de aula e manuais didáticos.

Publicações Seriadas

São os seriados técnico-científicos: boletins, periódicos, anuários e anais de eventos (simpósios e congressos). Constam destas publicações o Internacional Standard Serial Number (ISSN), que é um código único e definitivo para identificação de títulos de seriados.

Programas de Computador (PDC)

São a seqüência de instruções ou códigos, expressos em uma linguagem de programação compilada ou interpretada, a ser executada por um computador para alcançar um determinado objetivo. Aceitam-se tanto programas fonte quanto os executáveis.

Pré-publicações (PRE)

Todos os artigos publicados em periódicos, anais e como capítulos de livros.