DataHub e Apache Atlas: Uma Análise Comparativa de Ferramentas de Catalogação de Dados

DataHub and Apache Atlas: A Comparative Analysis of Data Catalog Tools

Diogo Rodrigues, Centro de Computação Gráfica, Centro de Investigação ALGORITMI, Universidade do Minho, Portugal, diogo.rodrigues@ccg.pt

Mariana Almeida, Centro de Computação Gráfica, Centro de Investigação ALGORITMI, Universidade do Minho, Portugal, mariana.almeida@ccg.pt

Pedro Guimarães, Centro de Computação Gráfica, Centro de Investigação ALGORITMI, Universidade do Minho, Portugal, pedro.guimaraes@ccg.pt

Maribel Yasmina Santos, Centro de Investigação ALGORITMI, Universidade do Minho, Portugal, maribel@dsi.uminho.pt

Resumo

Big Data introduz um acréscimo significativo de complexidade aos projetos, nos quais, a utilização de dados inadequados irá produzir, inevitavelmente, análises inadequadas e incorretas. Os catálogos de dados centralizam todos os metadados de um sistema de *Big Data* num único local, fornecendo uma visão global dos dados armazenados, pelo que é fundamental a utilização de ferramentas de catalogação de dados adequadas aos projetos em que inserem. A escolha da ferramenta que melhor se adequa às necessidades dos projetos deve ser o mais fundamentada possível. Neste trabalho foi aplicada a metodologia OSSpal, de comparação de tecnologias *open-source*, para a análise comparativa de duas ferramentas: DataHub e Apache Atlas.

Palavras-chave: Catalogação de Dados; Análise Comparativa; Metodologia OSSpal; DataHub; Apache Atlas

Abstract

Big Data introduces a significant increase of complexity to projects, in which, the use of inadequate data will inevitably produce inadequate and incorrect analysis. Data Catalogs centralize the system's metadata into one place, providing a global view of the stored data, so it is essential to use appropriate data catalog tools. The choice of the tool that best suits the needs of the projects must be well-founded. This paper uses the OSSpal methodology, usually used for comparing open-source technologies, to do a comparative analysis of two tools: DataHub and Apache Atlas.

Keywords: *Data Catalog; Comparative Analysis; OSSpal Methodology; DataHub; Apache Atlas*

1. INTRODUÇÃO

Na era *Big Data*, os catálogos de dados surgem como um *standard* para a gestão de metadados (Dibowski & Schmid, 2021). Estes são serviços de gestão de metadados que mantêm um inventário de dados para auxiliar a descoberta, descrição e organização de conjuntos de dados. O catálogo

fornece contexto a todos os utilizadores dos dados, de forma a que consigam encontrar e compreender determinado conjunto de dados com a finalidade de extrair valor para o seu negócio (Zaidi et al., 2017).

O componente mais importante de um catálogo de dados são os metadados e por esse facto torna-se importante perceber o que são e para que servem. Os metadados são dados sobre dados ou informação sobre dados (Preziuso et al., 2021). Descrevem de forma concisa e consistente os dados, ajudando a interpretar o seu significado (Costa, 2019). Os metadados são a informação estrutural que qualquer coleção de dados tem associada a si. São utilizados pelos analistas para decifrar o conteúdo dos dados à sua disposição (Inmon, 2016).

Existem diversas ferramentas de catalogação de dados, tais como Data Hub, Apache Atlas, Collibra Catalog, Cloudera Navigator, Oracle Cloud Infrasctruture Data Catalog, SAP Data Intelligence, entre outras. Todavia, o presente artigo foca-se na comparação entre o Apache Atlas e o DataHub, uma vez que o primeiro se trata de uma ferramenta criada há mais tempo, estabelecida e amplamente utilizada, com atualizações menos frequentes, enquanto a segunda é bastante mais recente, ainda em crescimento exponencial, com atualizações constantes, tal como poderá ser comprovado no decorrer do presente artigo. Por estes factos considerou-se necessário perceber o que muda entre as duas tecnologias, quais os pontes fortes e fracos de cada uma e em que contextos cada uma pode ser utilizada, sendo isso uma das motivações para este artigo. O presente artigo procura concluir qual destas ferramentas melhor responde às necessidades dos seus utilizadores, como engenheiros e cientistas de dados, detalhando as suas características, funcionalidades e arquiteturas. Este artigo é motivado pela reduzida literatura científica na comparação de ferramentas de *Data Catalog* e utiliza a metodologia OSSpal (Wasserman et al., 2017) como base para a definição de critérios e métricas a avaliar.

Este artigo encontra-se dividido da seguinte forma. A secção 2 aborda trabalhos relacionados que utilizam a metodologia OSSpal. A secção 3 introduz as ferramentas em análise. A secção 4 explica a metodologia de comparação, nomeadamente, o procedimento de avaliação e a definição de critérios. A secção 5 compara as ferramentas de catalogação seguindo a metodologia definida. A secção 6 diz respeito a conclusões e trabalho futuro.

2. TRABALHOS RELACIONADOS

Os autores (Calcada & Bernardino, 2019), (Metelo et al., 2021), (Marques & Bernardino, 2019) e (Leite et al., 2018) aplicam a metodologia OSSpal para a avaliação de ferramentas *open-source*. Entre estes, a principal diferença está associada ao tipo de tecnologias que são alvo de análise.

(Carvalho et al., 2022) procede a uma análise comparativa de ferramentas de *design* de modelação de dados. Para efetuar essa análise comparativa, os autores utilizam como base o modelo *Business Readiness Rating* (BRR) e a metodologia de avaliação OSSpal.

A maior distinção para os artigos é que este último, considera que a melhor abordagem para a avaliação de ferramentas *open-source* passa por uma metodologia de avaliação híbrida entre a BRR e a OSSpal, por forma a colmatar os problemas apontados a estas metodologias, como é o caso do viés e da falta de detalhe do BRR e a penalização por altas pontuações em medidas menos críticas da OSSpal.

De notar que em nenhum dos artigos mencionados nesta secção é utilizada a ferramenta de *Quick Assessment* para auxílio nas avaliações.

3. DESCRIÇÃO DAS FERRAMENTAS

A presente secção apresenta as ferramentas de catalogação de dados em análise, com a descrição de funcionalidades e da arquitetura das mesmas.

3.1. DataHub

O DataHub é uma plataforma de metadados *open-source* e extensível que procura ajudar o utilizador a lidar melhor com a complexidade da sua *data stack*, permitindo a observabilidade e a identificação dos dados, bem como a governança federada dos dados (About DataHub, 2021).

Esta ferramenta foi originalmente criada pelo LinkedIn e foi, subsequentemente, tornada *open-source* sob a versão 2.0 da licença Apache (Apache License, Version 2.0, 2004) e é agora uma comunidade com centenas de contribuidores e usada numa variedade de empresas (Open Sourcing DataHub | LinkedIn, 2020).

Esta tecnologia procura ajudar a lidar com a crescente complexidade dos ecossistemas de dados, bem como auxiliar na extração de todo o valor que advém dos dados em determinada organização. Por forma a realizar tais objetivos, esta plataforma necessita de integrar uma variedade de funcionalidades, que passam a ser explanadas de seguida (*Features* | *DataHub*, 2021).

Pesquisa e Identificação de Dados

- Permite a pesquisa de recursos numa variedade de fontes, como bases de dados, *Data Lakes*, plataformas de *Business Intelligence* (BI), entre outros;
- Permite compreender o percurso dos dados rastreando a linhagem dos mesmos;

- Permite navegar pelo grafo da linhagem para obter informação sobre o contexto de determinada entidade com que o utilizador se depara, melhorando assim a compreensão do fluxo dos dados;
- Disponibiliza uma variedade de estatísticas relativas ao *profiling* e à utilização dos dados.

Documentação e Etiquetagem

- Permite introduzir e atualizar a documentação referente a qualquer conjunto de dados via API ou através da *User Interface* (UI) do DataHub;
- Permite criar e adicionar *tags* em qualquer tipo de entidade, via GraphQL API, ou através da UI;
- Permite pesquisar e navegar pelas diversas *tags* por forma a acelerar a descoberta entre entidades.

Governança de Dados

- Permite atribuir propriedade dos recursos a utilizadores e/ou grupo de utilizadores;
- Permite gerir o controlo de acessos com a elaboração de regras que especifiquem, por exemplo, os privilégios que cada utilizador e/ou grupo de utilizadores têm sobre determinados recursos.

Análise de Qualidade e Utilização de Metadados

- Permite o acesso a dados estatísticos relativos aos metadados e à interação dos utilizadores com a plataforma como, por exemplo, informações sobre o volume total de ativos, número de utilizadores ativos em cada mês, entre vários outros.

A nível de modelos de integração, o DataHub fornece uma grande variedade de integrações possíveis para, por exemplo, permitir a ingestão de metadados. De notar que esta é uma lista que se encontra em constante atualização dado que cada vez mais integrações vão sendo adicionadas e poderão ser consultadas em (*Features | DataHub*, 2021).

O DataHub fornece integradores para fontes de dados, ferramentas de BI, ferramentas de *Extract, Transform, Load* (ETL) ou *Extract, Load, Transform* (ELT), orquestração de *workflows*, observabilidade dos dados, plataformas de *Machine Learning* (ML) e gestão de identidade (*Features | DataHub*, 2021).

O DataHub segue uma filosofia de *model-first* com um foco em permitir a interoperabilidade entre ferramentas e sistemas distintos. Na Figura 1 encontra-se uma das representações da arquitetura de alto nível do DataHub (*Overview | DataHub*, 2021).

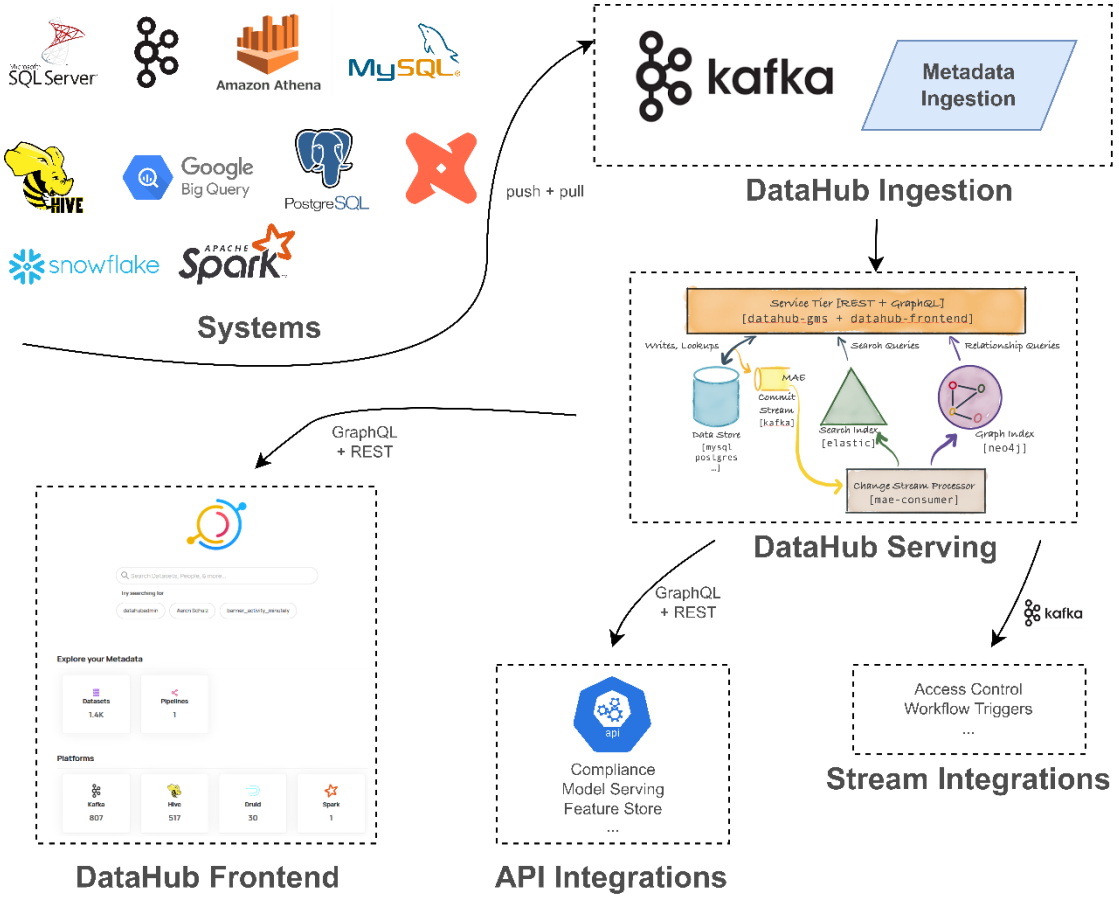


Figura 1 – Arquitetura de alto nível do DataHub. Adaptado de: (Overview | DataHub, 2021)

São agora abordados alguns pontos fulcrais para compreender a arquitetura do DataHub, nomeadamente:

- Modelação de metadados com uma abordagem *schema-first*: O DataHub utiliza uma linguagem de serialização para descrever o modelo de metadados. São suportadas APIs REST e GraphQL. Em adição, é suportada uma API baseada em Avro, sobre o Kafka, para comunicar alterações e proceder a subscrições nos metadados;
- Plataforma de metadados em tempo real: A infraestrutura de metadados do DataHub é orientada ao processamento de *streams*, o que permite que alterações aos metadados se vejam refletivas na plataforma numa questão de segundos;
- *Federated Metadata Serving*: O DataHub apresenta um único serviço de metadados (*Generalized Metadata Service*) no repositório *open-source*. No entanto, também suporta serviços federados de metadados.

Na Figura 2 encontram-se os diferentes componentes que compõem a plataforma DataHub que serão abordados de seguida (*Components | DataHub*, 2021).

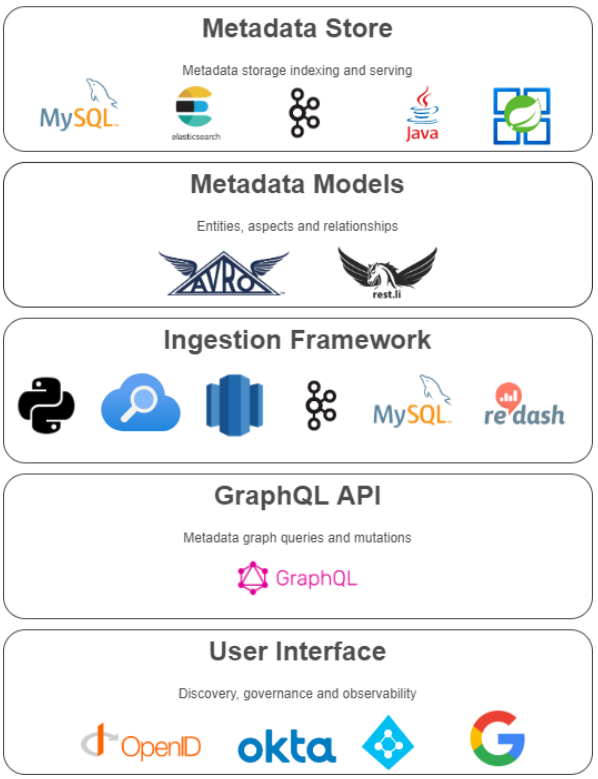


Figura 2 - Componentes da Plataforma DataHub. Adaptado de: (Components | DataHub, 2021).

- Metadata Store: Armazena as entidades e aspetos do grafo de metadados e disponibiliza uma API para a ingestão de metadados, para extrair metadados pela chave primária, para a pesquisa de entidades e para extrair as relações entre as entidades;
- Metadata Models: São os esquemas que definem o grafo de metadados composto por entidades e aspetos e as suas relações. São criados utilizando uma versão estendida da *Pegasus Data Language* (PDL), uma linguagem de modelação.
- Ingestion Framework: Trata-se de uma biblioteca Python, modular e extensível, para extrair metadados de fontes externas, transformá-los para o modelo de metadados do DataHub e colocá-los no DataHub via Kafka ou REST;
- GraphQL API: API orientada às entidades que procura simplificar a interação com as entidades do grafo de metadados. Inclui APIs para adicionar e remover *tags*, *owners*, *links*, entre outras;
- User Interface: Trata-se de uma UI em React, que consome a GraphQL API por forma a permitir a descoberta, governança e *debugging* dos dados.

Os metadados no DataHub são modelados conceptualmente através das seguintes abstrações (Metadata Model | DataHub, 2021):

- Entidades: Uma entidade é o nodo primário do modelo de metadados. É constituído por uma chave primária e grupos de atributos de metadados que são denominados por aspetos;
- Aspetos: Um aspeto é um conjunto de atributos que descreve uma determinada faceta de uma entidade. Estes podem ser partilhados por várias entidades e podem ser atualizados de forma independente;
- Relações: Uma relação representa a ligação entre duas entidades. Estas contêm uma anotação relativa ao tipo de relação entre as entidades e são declaradas através de uma chave estrangeira. É permitido que estas relações sejam percorridas de forma bidirecional;
- Identificadores (chaves e *Uniform Resource Name* (URNs)): Uma chave é um tipo especial de aspeto que contém um campo que identifica de forma única uma entidade. Esta pode ser serializada para URN, que permite a execução de pesquisas sobre entidades com maior facilidade, e as URNs podem ser convertidas de volta para a estrutura do aspeto chave.

Qualquer utilizador do DataHub pode estender o modelo de metadados, quer criando novas entidades ou estendendo entidades existentes. Como realizar essa extensão do modelo é explicado passo a passo em (Extending the Metadata Model | DataHub, 2021).

3.2. Apache Atlas

Apache Atlas é uma ferramenta escalável e extensível que fornece recursos fundamentais para a gestão e governança dos dados e metadados, permitindo às organizações a criação de catálogos de dados e a integração com todo o ecossistema de dados corporativos. O Atlas foi desenvolvido pela Hortonworks em parceria com a *Data Governance Initiative* (DGI) e, em 2015, ingressou no Apache Foundation Incubator, onde cresceu, tornando-se num projeto de alto nível em 2017. O projeto de código open-source continua em desenvolvimento, com contribuições de várias organizações ((What Is Apache Atlas, 2021),(Data Governance and Metadata Framework, 2019)).

O Apache Atlas, ao utilizar o conector para os componentes Hadoop, permite a facilidade de troca de repositórios de metadados, promovendo a interoperabilidade. Esta ferramenta cinge-se a uma abordagem de inovação que permite acelerar a maturidade do produto e o tempo de retorno dos dados para as organizações que priorizam os mesmos (Apache Atlas, 2020).

(Data Governance and Metadata Framework, 2019) apresenta um conjunto de funcionalidades, agrupadas pelas suas características, que estão presentes na ferramenta e garantem a eficácia da mesma na catalogação dos dados.

Tipos de Metadados

- O tipo de dados está predefinido para vários metadados Hadoop e não Hadoop;

- Tem a capacidade de definir novos tipos de metadados;
- Assume atributos primitivos, complexos, referências de objetos;
- Capacidade de capturar informação sobre metadados e os seus relacionamentos;
- Integra APIs REST para trabalhar com tipos e instâncias que permitem uma integração facilitada.

Classificação

- Capacidade de criar automaticamente classificações;
- As entidades podem estar associadas a várias classificações, facilitando a descoberta;
- Capacidade de garantir automatização nas classificações, de forma que estas sigam os dados à medida que estes passam por vários tipos de processamentos.

Linhagem e Pesquisa

- Interface intuitiva para visualizar os dados à medida que passam por vários processos;
- Possibilidade de aceder e atualizar linhagens por meio de APIs REST;
- Utiliza a linguagem *Structured Query Language* (SQL) como linguagem de consulta para metadados.

Segurança

- Segurança para o acesso de metadados, permitindo controlos de acesso e operações como adicionar, atualizar e remover classificações;
- Integração com o Apache Ranger que permite estabelecer autorizações no acesso aos dados com base em classificações associadas a entidades existentes no Apache Atlas.

A Figura 3 apresenta a arquitetura do Apache Atlas, onde são representadas as várias camadas e componentes da ferramenta, detalhadas no decorrer desta subsecção (Apache Atlas – Architecture, 2019).

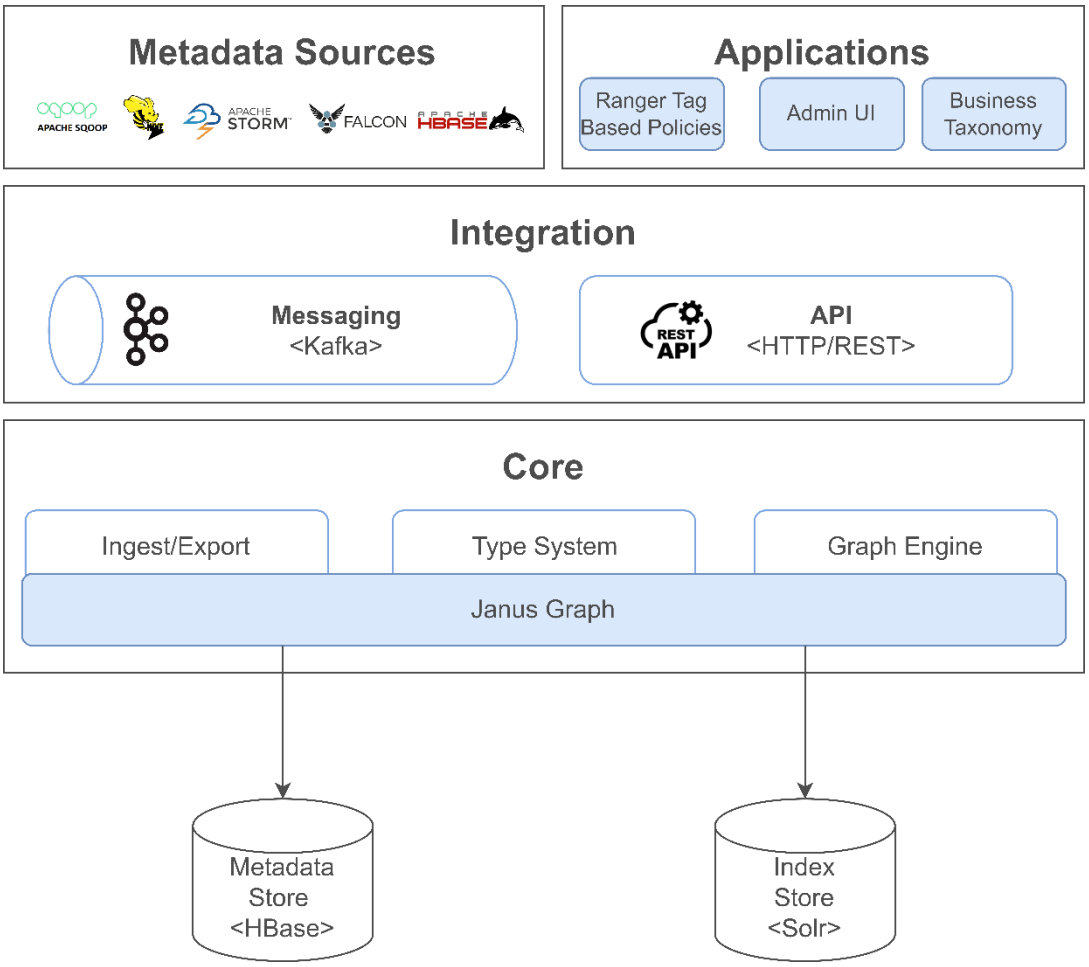


Figura 3 - Arquitetura do Apache Atlas. Adaptado de (Apache Atlas – Architecture, 2019)

Core

- Type System: permite aos utilizadores definirem um modelo para os objetos de metadados, composto por instâncias 'tipos' chamadas de 'entidades'. Esta componente permite definir e gerir os tipos e as entidades;
- Graph Engine: esta componente é responsável pela tradução entre tipos e entidades do sistema e criação de índices apropriados para os objetos para que possam ser pesquisados com eficiência;
- Ingest/Export: permite exportar os dados com as alterações feitas para que o utilizador possa consumir e reagir aos eventos de alteração dos metadados em tempo real.

Integration

- API: todas as funcionalidades são expostas aos utilizadores por meio da incorporação de uma API REST que permite a criação, atualização e exclusão de tipos e entidades;

- Messaging: os utilizadores podem optar por integrar-se à ferramenta através de uma interface de mensagens baseada no Kafka, sendo útil na comunicação de objetos de metadados e no consumo de eventos de alteração.

Metadata Sources

- A ferramenta Apache Atlas oferece suporte à ingestão e gerenciamento de dados provenientes de várias fontes. Para a integração é necessário a existência de modelos de metadados e componentes destinadas à ingestão de objetos de metadados.

Applications

- Atlas Admin UI: consiste numa aplicação *web* que permite aos utilizados descobrir e anotar metadados, com uma interface de pesquisa e uma linguagem de consulta SQL, que pode ser utilizada para consultas;
- Tag Based Policies: componente responsável pela integração do Apache Ranger com o Apache Atlas que garante uma governança eficaz, permitindo aos utilizadores a definição de políticas de segurança;
- Business Taxonomy: com vista a melhorar a capacidade de governança, o Apache Atlas tem uma interface de taxonomia que permite aos utilizadores definir um conjunto hierárquico de termos comerciais.

4. METODOLOGIA DE COMPARAÇÃO TECNOLÓGICA

OSSpal é uma metodologia que sucede, entre outras, à metodologia *Business Readiness Rating* (BRR) fundado em 2005, que tem como objetivo auxiliar as organizações a encontrar software *open-source* gratuito e de alta qualidade, *Free and Open-Source Software* - FOSS (Welcome to OSSpal | OSSPAL, n.d.).

A alteração do projeto de BRR para OSSpal trouxe uma variedade de alterações, entre as quais se destaca, em (Wasserman et al., 2017), o número de categorias que, graças à junção de algumas categorias da metodologia BRR, deram origem a sete áreas de medição de software *open-source*, nomeadamente:

- Funcionalidades: Satisfação das necessidades dos utilizadores ao utilizar o software;
- Características operacionais do software: Segurança, desempenho e escalabilidade do software, qualidade da interface gráfica, facilidade de utilização por parte do utilizador final, facilidade de instalação, configuração, implantação e manutenção;
- Suporte e serviço: Suporte comercial ou de comunidade ao software e existência de serviços de formação e consultoria;

- Documentação: Existência de documentação e tutoriais adequados para o software;
- Atributos tecnológicos do software: Arquitetura do software, modularidade, portabilidade, flexibilidade, extensibilidade, abertura e facilidade em integrar com o software e qualidade do código, *design* e testes;
- Comunidade e adoção: Comunidade ativa de utilizadores e adoção do software pela comunidade, mercado e indústria;
- Processo de desenvolvimento: Nível de profissionalismo no processo de desenvolvimento do software e organização do projeto.

A metodologia OSSpal é composta por quatro fases (Leite et al., 2018):

- Fase 1 – Identificação dos componentes a serem analisados;
- Fase 2 – Alocação de pesos para cada categoria e medidas definidas;
- Fase 3 – Atribuição de pontuação a cada métrica, de 1 (inaceitável) a 5 (excelente);
- Fase 4 – Cálculo da pontuação final da avaliação através das pontuações e pesos atribuídos.

A avaliação da primeira categoria (Funcionalidades) é distinta das restantes uma vez que cada tipo de tecnologia tem um conjunto de características que necessitam de estar presentes no software. Para isso, a avaliação das funcionalidades é feita por comparação com um conjunto de características padrão, que poderão ser definidas de raiz ou citando uma fonte externa (Wasserman et al., 2017). Os seguintes passos devem então ser seguidos para a avaliação desta categoria:

- Atribuição de uma pontuação numa escala de 1 a 3 relativa à importância de cada um dos itens presentes na lista de características padrão escolhida, sendo 1 pouco importante e 3 muito importante;
- Comparação entre a lista de características padrão e a tecnologia a avaliar adicionando a importância corresponde a cada item a uma soma cumulativa. Caso a tecnologia não contenha algum dos itens da lista, retirar pontuação;
- Divisão da soma cumulativa pela pontuação máxima possível obtendo assim a pontuação do software em termos de funcionalidades;
- Normalização da pontuação das funcionalidades para a escala de 1 a 5 da seguinte forma:
    - < 65% = 1 (Inaceitável)
    - 65% a 80% = 2 (Mau)
    - 80% a 90% = 3 (Aceitável)

- 90% a 96% = 4 (Muito Bom)
- > 96% = 5 (Excelente)

5. ANÁLISE COMPARATIVA

A presente secção apresenta uma análise comparativa entre as tecnologias de catalogação de dados, DataHub e Apache Atlas, enumerando alguns critérios essenciais e atribuindo uma classificação às mesmas com o objetivo de selecionar a tecnologia que melhor cumpre os requisitos expectáveis.

OSSpal é a metodologia utilizada para a comparação de ferramentas, que considera sete características, descritas na secção 4. A Tabela 1 atribui pesos a cada uma das categorias definidas consoante a sua importância, onde a soma cumulativa tem de ser obrigatoriamente 100%.

CATEGORIAS	PESO
Funcionalidades	30 %
Características operacionais do software	20 %
Suporte e serviço	15 %
Documentação	15 %
Atributos tecnológicos do software	10%
Comunidade e adoção	5 %
Processo de desenvolvimento	5 %
Total	**100%**

Tabela 1 - Atribuição de pesos a cada uma das categorias

A primeira característica, funcionalidades, é considerada a categoria mais importante, uma vez que consiste na avaliação de características imprescindíveis para alcançar o resultado esperado. Por este motivo foi-lhe atribuído um peso de 30%. As características de software operacional, que dizem respeito à qualidade do software do sistema a nível de desempenho, escalabilidade e usabilidade, são a segunda categoria mais importante com um peso de 20%. Às categorias documentação e suporte e serviço são atribuídos o peso de 15% pois considera-se fulcral que os utilizadores tenham acesso a tutoriais e documentação, de forma a entenderem o funcionamento da ferramenta e estarem a par de alterações e novas versões. Estas categorias representam um papel significativo para os utilizadores que recorrem a ferramentas *open-source*. Os atributos tecnológicos do software apresentam uma importância não tão significativa, todavia é importante que estas tenham a capacidade de integrar ferramentas de armazenamento de grande volume de dados. A esta categoria foi atribuída uma pontuação de 10%. Comunidade e adoção e o processo de desenvolvimento são consideradas as categorias com menos importância com um peso de 5% cada.

As funcionalidades de um catálogo de dados variam consoante as necessidades dos projetos para os quais são construídos. Por isso, é necessário estabelecer algumas características desta categoria que vão servir como indicadores para a avaliação das ferramentas.

(Wells, 2020) afirma que o principal componente de um catálogo de dados são os metadados recolhidos sobre os conjuntos de dados, pelo que as principais características relacionadas com os dados são o acesso e pesquisa dos conjuntos de dados, assim como a avaliação dos mesmos.

As ferramentas de catalogação devem garantir a capacidade de avaliar a qualidade e utilidade dos conjuntos de dados disponíveis para determinado caso de uso sem ter de aceder e descarregar os mesmos. A capacidade de pré-visualizar os conjuntos de dados, visualizar notas adicionadas, ler críticas de outros utilizadores e ter acesso a informação sobre a qualidade dos dados poderá ser bastante útil nesta perspetiva.

Desta forma, o componente fundamental de uma ferramenta de catalogação é o acesso, pesquisa e avaliação de conjuntos de dados, sendo estas características de elevada importância.

(Zaidi et al., 2017) aborda a importância de promover a transparência na utilização de dados através da apresentação da sua linhagem e a análise do seu impacto, permitindo compreender o percurso e a origem dos mesmos, assim como as alterações que sofreram até chegarem ao seu estado atual, melhorando a compreensão do fluxo de dados.

No que toca a características de infraestrutura para ferramentas de catalogação de dados podemos destacar os custos associados à ferramenta. Esta característica é considerada de baixa importância, uma vez que ao nível empresarial é natural a existência de verbas a alocar para as tecnologias necessárias ao bom funcionamento das organizações, pelo que a gratuitidade da plataforma, apesar de útil em vários casos, não se enquadra no grupo de características imprescindíveis de uma plataforma de catalogação de dados.

Em relação ao controlo de acesso dos utilizadores, este serve para evitar que qualquer utilizador possa aceder a todo o conteúdo da plataforma. Esta característica é considerava de importância média uma vez que, apesar de não fazer parte das características *core* de uma plataforma de catalogação de dados, é de grande importância para um funcionamento seguro da plataforma.

Para cada uma das características de funcionalidades foi atribuída uma pontuação de 1 a 3, consoante a sua importância, de acordo com a metodologia OSSpal, sendo 1 o menos importante e 3 muito importante, tal como apresentado na Tabela 2.

CARACTERÍSTICAS	PONTUAÇÃO
Acesso e pesquisa de conjuntos de dados	3
Avaliação de conjuntos de dados	3

Monitorização e rastreamento dos dados	2
Gratuitidade	1
Controlo de acesso dos utilizadores	2

Tabela 2 - Atribuição de pontuações a cada característica das funcionalidades

Em seguida, na Tabela 3 comparam-se as funcionalidades das ferramentas DataHub e Apache Atlas consoante as características estabelecidas, com o objetivo de normalizar a pontuação das funcionalidades de cada ferramenta para uma escala de 1 a 5.

Caso a ferramenta tenha determinada característica ser-lhe-á atribuída a pontuação referente a tal característica, definida na tabela anterior. Caso contrário deverá ser atribuída a pontuação de -1.

CARACTERÍSTICAS	PONTUAÇÃO	DATAHUB	APACHE ATLAS
Acesso e pesquisa de conjuntos de dados	3	3	3
Avaliação de conjuntos de dados	3	3	3
Monitorização e rastreamento dos dados	2	2	2
Gratuitidade	1	1	1
Controlo de acesso dos utilizadores	2	2	2
Soma cumulativa	11	11	11
Pontuação	100 %	100%	100%
		5	5

Tabela 3 - Pontuação das funcionalidades

A Tabela 4 apresenta uma análise comparativa entre as ferramentas de acordo com as sete categorias apresentadas na secção anterior, recorrendo a uma escala de 1 a 5, correspondente à terceira fase da metodologia OSSpal, para classificar a qualidade de cada ferramenta em cada uma das categorias, sendo 1 a pontuação mínima e 5 a pontuação máxima.

CATEGORIAS	DataHub	Apache Atlas
Funcionalidades	5	5
Características operacionais do software	3	3
Suporte e serviço	5	3
Documentação	2	3
Atributos tecnológicos do software	4	4
Comunidade e adoção	4	2
Processo de desenvolvimento	4	2

Tabela 4 – Análise comparativa entre as ferramentas

DataHub

No que diz respeito às funcionalidades, a pontuação atribuída foi já justificada através da informação presente na tabela 3.

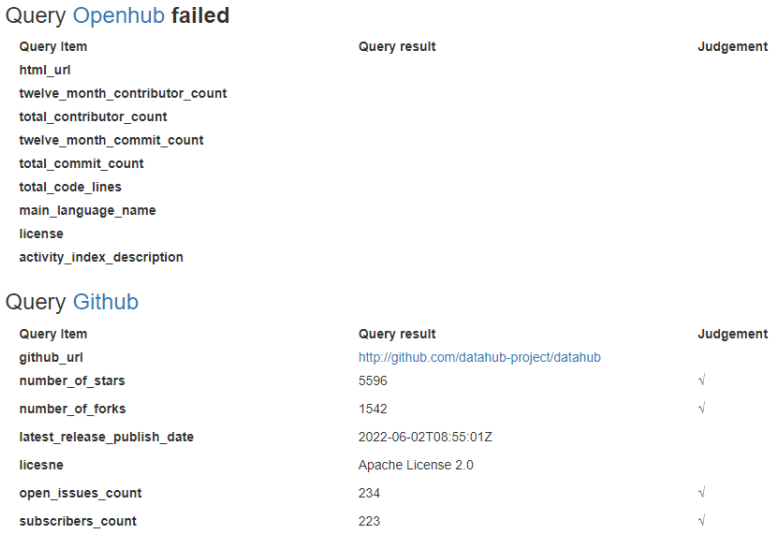
Em relação às características operacionais de software, foram atribuídos 3 pontos. O DataHub assenta em *containers* Docker o que facilita o processo de instalação, configuração, implantação e manutenção. No entanto, uma vez que a aplicação ainda é recente e as atualizações são frequentes é possível encontrar alguns problemas na instalação em determinadas versões. Abordando a interface gráfica, esta contém um aspeto visual simples, intuitivo e de fácil utilização, a conexão a fontes de dados é simplificada uma vez que os ficheiros de configuração necessários para tal são editáveis a partir da UI, porém, é possível encontrar alguns *bugs* na interface gráfica, nomeadamente na apresentação dos metadados.

No que toca ao suporte e serviço, o DataHub encontra nesta categoria a característica em que mais se destaca e, por isso, a classificação de 5 pontos. Esta conta com uma forte componente de suporte ao software uma vez que qualquer utilizador poderá utilizar o canal de Slack dedicado à plataforma para colocar as suas dúvidas relacionadas com qualquer aspeto relativo ao DataHub e ver as mesmas serem respondidas com bastante rapidez, seja por algum membro da comunidade, seja por qualquer colaborador da Acryl Data, organização que dirige o projeto *open-source* e elabora seminários regulares em que apresenta as novidades da plataforma com exemplos práticos para uma melhor aprendizagem por parte dos utilizadores.

Em termos de documentação, o DataHub conta com o seu *website* que contém documentação em relação à plataforma, a sua instalação, arquitetura, entre muitas outras. No entanto, contém uma componente mais restrita de *troubleshooting*, fazendo com que diversos problemas comuns dos utilizadores não consigam ser resolvidos com base na documentação pelo que esta categoria é classificada com 2 pontos.

Em relação às características tecnológicas do software, o DataHub tem uma arquitetura simples, modular e portável dado que assenta em Docker, que modula todos os componentes necessários ao bom funcionamento da plataforma e que pode funcionar em qualquer sistema operativo. O DataHub permite a extensibilidade da plataforma através da extensão do metamodelo de dados. Por estas razões foram atribuídos 4 pontos na avaliação, deixando espaço para crescimento e aperfeiçoamento.

Em relação ao tópico da comunidade e adoção, tomou-se por base a utilização da *Quick Assessment Tool* da metodologia OSSpal que mede alguns dados das páginas de ferramentas *open-source* no GitHub e OpenHub. No caso do DataHub, este apenas apresenta resultados relativos ao GitHub e podem ser vistos na Figura 4. Como pode ser visto na figura pelo número de *forks*, estrelas, subscritores, *issues* e pela data do último lançamento, o DataHub conta tanto com uma comunidade bastante ativa como com níveis de adoção elevados e, por isso, foi-lhe atribuído 4 pontos.



Query Openhub **failed**

Query Item	Query result	Judgement
html_url		
twelve_month_contributor_count		
total_contributor_count		
twelve_month_commit_count		
total_commit_count		
total_code_lines		
main_language_name		
license		
activity_index_description		

Query Github

Query Item	Query result	Judgement
github_url	http://github.com/datahub-project/datahub	
number_of_stars	5596	√
number_of_forks	1542	√
latest_release_publish_date	2022-06-02T08:55:01Z	
licesne	Apache License 2.0	
open_issues_count	234	√
subscribers_count	223	√

Figura 4 - *Quick Assessment Tool* (DataHub)

Por fim, relativamente ao processo de desenvolvimento, o projeto conta com uma organização que lidera o projeto, a Acryl Data, pelo que grande parte do desenvolvimento de software se encontra profissionalizado, não esquecendo as contribuições da comunidade, o que contribui para os 4 pontos atribuídos à plataforma.

Apache Atlas

A ferramenta Apache Atlas baseia-se em containers Docker para a sua instalação, tendo como pré-requisitos uma máquina virtual com Docker, Google Cloud ou Azure e imagens do Apache Atlas, Zookeeper, Kafka, Hadoop e Hive PostgreSQL. Após a extração das imagens, adiciona-se uma compilação Maven do Atlas.

A nível de segurança, suporta recursos como SSL, os quais permitem uma comunicação segura entre o *site* e o navegador, autenticações de serviço, permitindo que a plataforma interaja com outras plataformas, como por exemplo o HDFS e autenticação HTTP responsável pelo controlo de acesso e autenticação. A integração com o Apache Ranger é uma vantagem da ferramenta pois permite estabelecer autorizações no acesso aos dados.

A interface é bastante intuitiva e simples, o que facilita a compreensão das suas funcionalidades por parte do utilizador e facilidade no manuseamento da plataforma, permitindo, através de APIs REST, visualizar e atualizar a linhagem dos dados, assim como pesquisar critérios com condições mais complexas. Todavia, existem alguns erros na plataforma como, por exemplo, com a criação de várias tabelas Hive na mesma base de dados com um único comando: o Atlas não consegue capturar todos os eventos da criação da tabela. Também eventos simultâneos podem produzir entidades duplicadas, assim como a hora de criação de uma tabela Hive não é refletiva na ferramenta.

A extensa lista de pré-requisitos e dificuldades de instalação, devido ao surgimento de novas versões, contribuem para a descida da classificação da ferramenta quanto às características operacionais de software, classificada com 3 pontos.

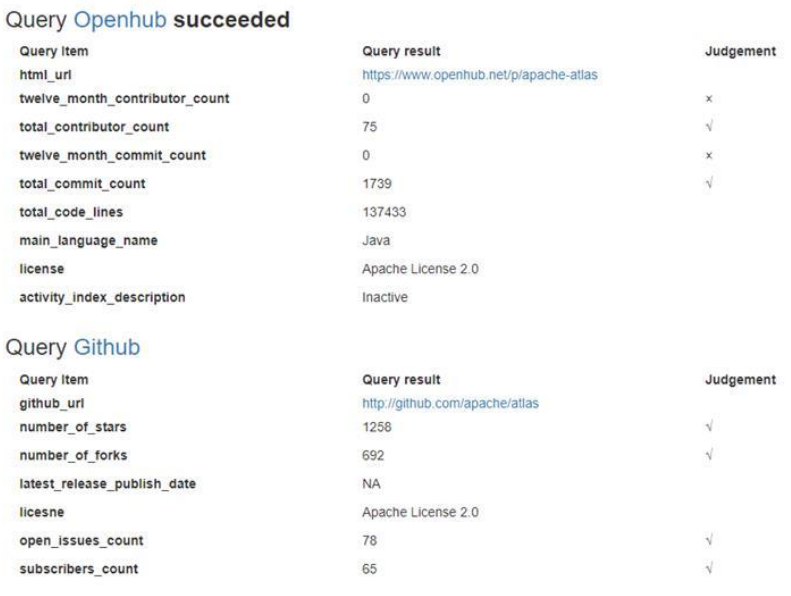
Na área suporte e serviço, a ferramenta não tem nenhuma plataforma própria disponível para interação entre utilizadores e resolução de problemas, deixando essas tarefas para plataformas de terceiros, como o Stack Overflow e Youtube, que disponibilizam vídeos, tutoriais e respostas a questões dos utilizadores. Destarte, atribui-se uma classificação de 3 pontos.

A ferramenta apresenta uma larga documentação de todas as versões, requisitos e tutoriais de instalação, bugs registados e possíveis soluções e detalhe sobre a arquitetura, características, API REST, informações do projeto, entre outras, traduzindo-se num ponto forte no uso da ferramenta. A documentação tem como suporte diversas imagens, como descrição da arquitetura e até comandos a serem utilizados. Toda esta informação está presente no site oficial. Desta forma, atribui-se a pontuação de 3 pontos à ferramenta na área de documentação e tutoriais adequados para o *software*.

O Atlas é normalmente utilizado em ambientes Hadoop, porém possibilita a integração com outros ambientes, pois possui uma arquitetura escalável e extensa. Esta ferramenta suporta atributos primários e complexos, permite a criação dinâmica de classificações e suporta a ingestão de dados provenientes de diversas fontes.

O Apache Atlas, com recurso a um modelo gráfico, permite grande flexibilidade e eficiência nos relacionamentos entre os objetos dos metadados. Os utilizadores podem também optar por integrar uma interface de mensagens baseada no Kafka, aconselhável a quem pretende uma maior escalabilidade e confiabilidade. São atribuídos 4 pontos à área tecnológica da ferramenta.

A Figura 5 apresenta resultados relativamente ao OpenHub e Github do Apache Atlas. Apesar desta ser uma plataforma estável e com alguns anos de existência, em comparação com a ferramenta DataHub, esta não tem muitos contributos nem uma comunidade muito ativa, o que diminui o suporte, pelo que na secção comunidade e suporte é apenas atribuída uma pontuação de 2 pontos.



Query Openhub **succeeded**

Query Item	Query result	Judgement
html_url	https://www.openhub.net/p/apache-atlas	
twelve_month_contributor_count	0	×
total_contributor_count	75	√
twelve_month_commit_count	0	×
total_commit_count	1739	√
total_code_lines	137433	
main_language_name	Java	
license	Apache License 2.0	
activity_index_description	Inactive	

Query Github

Query Item	Query result	Judgement
github_url	http://github.com/apache/atlas	
number_of_stars	1258	√
number_of_forks	692	√
latest_release_publish_date	NA	
licesne	Apache License 2.0	
open_issues_count	78	√
subscribers_count	65	√

Figura 5 - Quick Assessment Tool (Apache Atlas)

Devido ao número de contributos por parte dos utilizadores não ser elevado, considera-se uma pontuação de 2 pontos para a área de processo e desenvolvimento.

Na Tabela 5 encontram-se as pontuações finais de cada uma das ferramentas, bem como os respetivos cálculos através das pontuações e pesos atribuídos a cada categoria.

CATEGORIAS	**PONTUAÇÃO**	
	DATAHUB	**APACHE ATLAS**
Funcionalidades	5 x 0.30 = 1.50	5 x 0.30 = 1.50
Características operacionais do software	3 x 0.20 = 0.60	3 x 0.20 = 0.60
Suporte e serviço	5 x 0.15 = 0.75	3 x 0.15 = 0.45
Documentação	2 x 0.15 = 0.30	3 x 0.15 = 0.45
Atributos tecnológicos do software	4 x 0.10 = 0.40	4 x 0.10 = 0.40
Comunidade e adoção	4 x 0.05 = 0.20	2 x 0.05 = 0.10
Processo de desenvolvimento	4 x 0.05 = 0.20	2 x 0.05 = 0.10
Pontuação Final	**3.95**	**3.60**

Tabela 5 - Pontuações parciais e finais das ferramentas

6. CONCLUSÕES E TRABALHO FUTURO

Este artigo avaliou duas tecnologias de catalogação de dados *open-source*, DataHub e Apache Atlas. Para essa análise foi utilizada a metodologia de avaliação de software *open-source*, OSSpal. Os resultados obtidos, baseados numa análise multifatorial, atribuem uma pontuação final de 3.95 para a tecnologia DataHub e 3.60 para a plataforma Apache Atlas. A ferramenta DataHub destaca-se

principalmente nas categorias de suporte e serviço, comunidade e adoção e processo de desenvolvimento ficando atrás da ferramenta Apache Atlas na categoria de documentação.

De alertar que a análise elaborada neste artigo tem, naturalmente, um certo grau de subjetividade, que a ferramenta de *Quick Assessment Tool* tenta diminuir, sendo essa uma limitação do presente trabalho, pelo que se salvaguarda que tal análise pode ser passível de diferentes interpretações.

Para trabalho futuro é importante que seja mantido o acompanhamento destas duas tecnologias, uma vez que as pontuações aqui atribuídas incidem apenas sobre o estado atual das ferramentas que, com sucessivas atualizações, irão certamente evoluir, o que poderá alterar as pontuações atribuídas. Por este facto será também necessária nova utilização da metodologia OSSpal, sempre que seja necessário comparar as duas ferramentas.

7. AGRADECIMENTOS

Este trabalho foi suportado pela FCT – Fundação para a Ciência e Tecnologia, no âmbito das unidades de I&D, Centro ALGORITMI, Projeto UIDB/00319/2020 e foi elaborado com o apoio do Centro de Computação Gráfica, no contexto do projeto "City Catalyst – Catalisador para as Cidades Sustentáveis", referência POCI/LISBOA-01-0247-FEDER-046119, cofinanciado pelo Fundo Europeu de Desenvolvimento Regional (FEDER), através do Portugal 2020 (P2020).

REFERÊNCIAS BIBLIOGRÁFICAS

About DataHub. (2021). https://blog.datahubproject.io/about

Apache Atlas – Architecture. (2019). https://atlas.apache.org/1.2.0/Architecture.html

Apache Atlas – Data Governance and Metadata framework for Hadoop. (2019). https://atlas.apache.org/2.0.0/index.html

Apache Atlas | Cloudera. (2020). https://www.cloudera.com/products/open-source/apache-hadoop/apache-atlas.html

Apache License, Version 2.0. (2004). https://www.apache.org/licenses/LICENSE-2.0

Calcada, A., & Bernardino, J. (2019). Evaluation of couchbase, couchdb and mongodb using osspal. *IC3K 2019 - Proceedings of the 11th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, *1*(Ic3k), 427–433. https://doi.org/10.5220/0008345104270433

Carvalho, G., Mykolyshyn, S., Cabral, B., Bernardino, J., & Pereira, V. (2022). Comparative Analysis of Data Modeling Design Tools. *IEEE Access*, *10*, 3351–3365. https://doi.org/10.1109/ACCESS.2021.3139071

Components | DataHub. (2021). https://datahubproject.io/docs/components

Costa, M. I. P. (2019). *Etiquetagem e rastreio de fontes de dados num Big Data Warehouse*. https://repositorium.sdum.uminho.pt/handle/1822/70190

Dibowski, H., & Schmid, S. (2021). Using Knowledge Graphs to Manage a Data Lake. *Informaitk 2020, Lecture Notes in Informatics (LNI), January*, 41–50.

Extending the Metadata Model | DataHub. (2021). https://datahubproject.io/docs/metadata-modeling/extending-the-metadata-model/

Features | DataHub. (2021). https://datahubproject.io/docs/features/

Inmon, B. (2016). Data Lake Architecture: Designing the Data Lake and avoiding the garbage dump. Technics publications, 2016. In *Technics Publications*.

Leite, N., Pedrosa, I., & Bernardino, J. (2018). Open source business intelligence platforms' assessment using osspal methodology. *ICETE 2018 - Proceedings of the 15th International Joint Conference on e-Business and Telecommunications*, *1*(Icete), 190–196. https://doi.org/10.5220/0006910101900196

Marques, J. F., & Bernardino, J. (2019). Evaluation of asana, odoo, and projectlibre project management tools using the osspal methodology. *IC3K 2019 - Proceedings of the 11th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, *2*(Ic3k), 397–403. https://doi.org/10.5220/0008351903970403

Metelo, M., Bernardino, J., & Pedrosa, I. (2021). *Avaliação de Ferramentas Open Source para Data Science usando a Metodologia OSSpal*. 588–607.

Open Sourcing DataHub: LinkedIn's Metadata Search and Discovery Platform | LinkedIn Engineering. (2020). https://engineering.linkedin.com/blog/2020/open-sourcing-datahub--linkedins-metadata-search-and-discovery-p

Overview | DataHub. (2021). https://datahubproject.io/docs/architecture/architecture/

Preziuso, D., Sempreviva, A., & Orrell, A. (2021). *Deliverable D12-Distributed Wind Data Catalog Development Guide and Instruction Manual*. https://www.ntis.gov/about

The Metadata Model | DataHub. (2021). https://datahubproject.io/docs/metadata-modeling/metadata-model

Wasserman, A. I., Guo, X., McMillian, B., Qian, K., Wei, M. Y., & Xu, Q. (2017). OSSpal: Finding and evaluating open source software. *IFIP Advances in Information and Communication Technology*, *496*, 193–203. https://doi.org/10.1007/978-3-319-57735-7_18

Welcome to OSSpal | OSSPAL. (n.d.). Retrieved June 9, 2022, from https://www.osspal.org/

Wells, D. (2020). *Introduction to Data Catalogs*. www.eckerson.com

What is Apache Atlas: Capabilities and How It Works | Atlan. (2021). https://atlan.com/what-is-apache-atlas/

Zaidi, E., De Simoni, G., Edjlali, R., & Duncan, A. D. (2017). Data Catalogs Are the New Black in Data Management and Analytics. *Gartner*, *December*, 1–16.