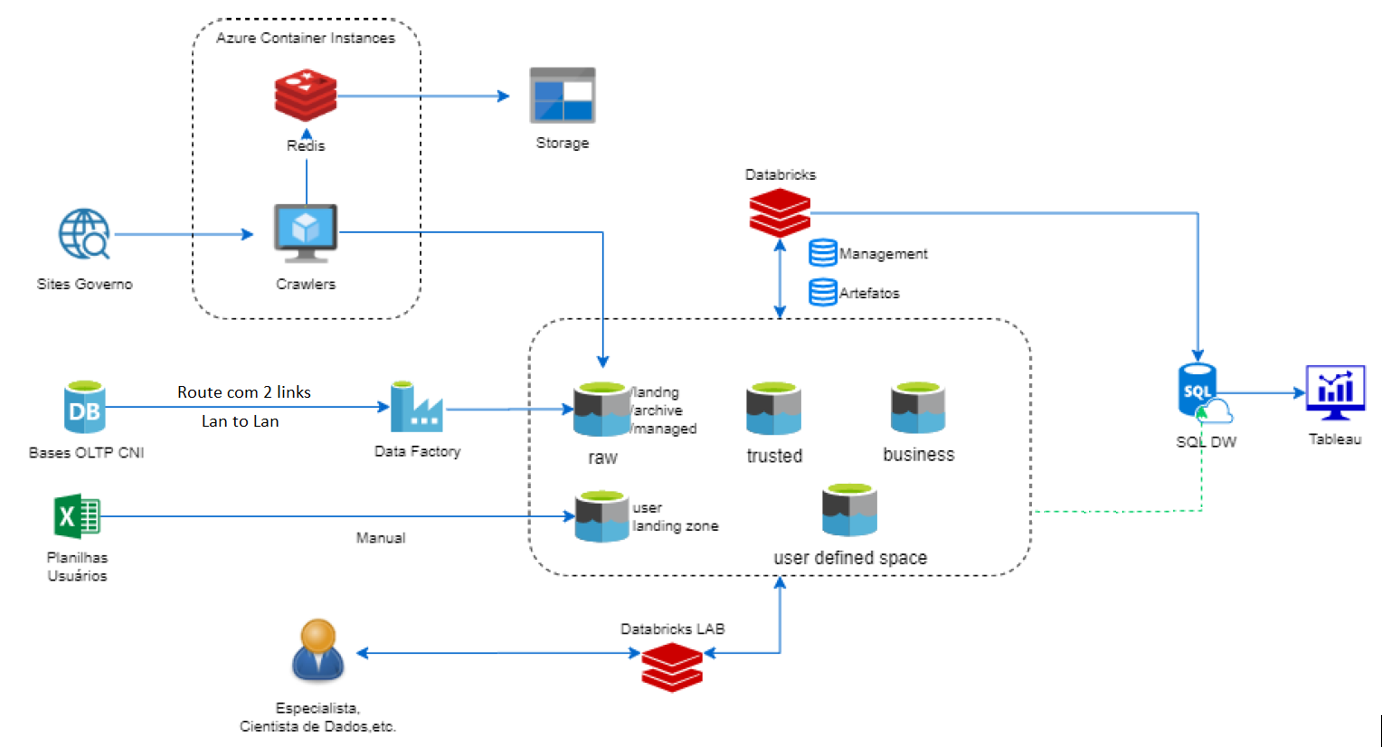
Anexo 1a – Arquitetura de Big Data

Uma arquitetura de Big Data foi projetada para lidar com a ingestão, o processamento e a análise de dados grandes ou complexos demais para sistemas de banco de dados tradicionais. Os dados podem ser processados em lote ou em tempo real.

O diagrama a seguir apresenta os componentes técnicos que se inserem na arquitetura de Big Data das ENSI, cujo ambiente computacional é provido por meio da nuvem da Microsoft Azure.

**Arquitetura Técnica (componentes utilizados)**



**Arquitetura do Enterprise Data Hub**

O objetivo de um Enterprise Data Hub (EDH) é permitir que a instituição tenha uma fonte de dados centralizada e unificada que possa fornecer rapidamente informações a diversos usuários de negócio, apoiando a tomada de decisão.

O Enterprise Data Hub inclui:

**Reservatório de Dados (**[**Data Lake**](http://www.cienciaedados.com/data-lake-a-fonte-do-big-data/)**):** Coleta de dados brutos que antes tinham alto custo para armazenamento e processamento. Os dados de diferentes fontes são gerenciados e governados no Data Lake, que também pode atuar como um arquivo online para dados acessados ​​com menos frequência.

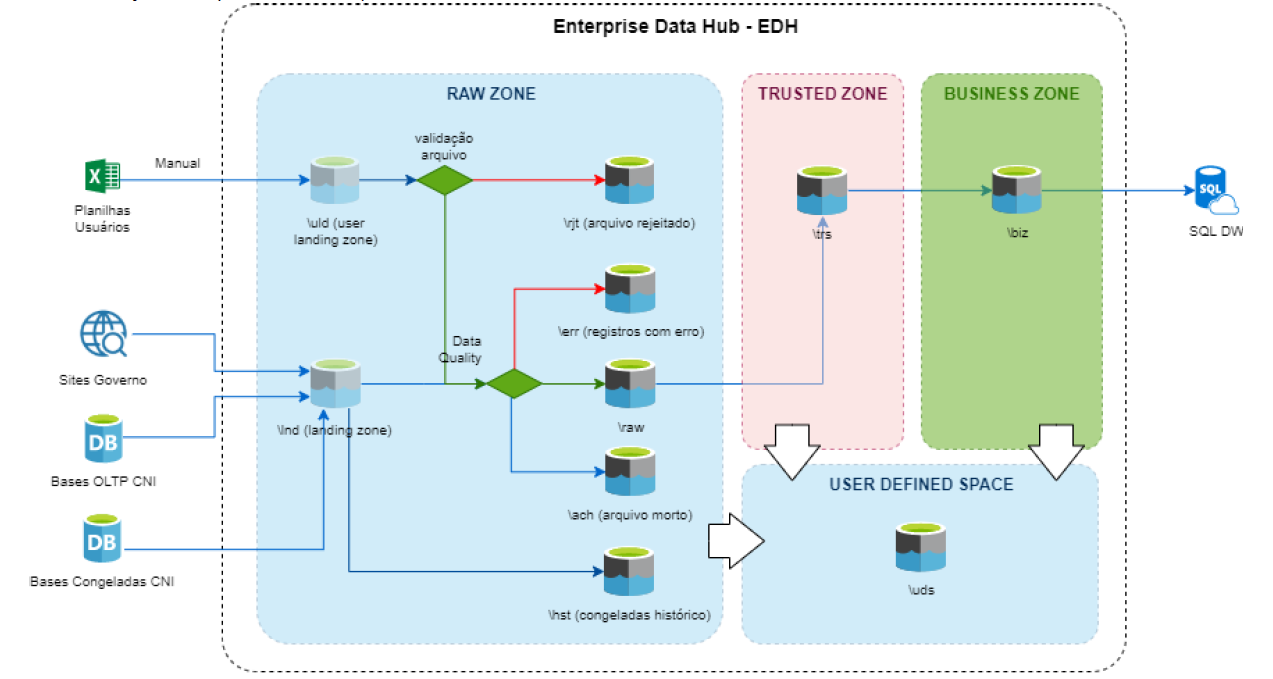
**Refino de dados**: Otimizar o processo de integração de diversos tipos de dados de várias fontes para descobrir as relações. Analisar, limpar, transformar e integrar dados.

**Exploração do Big Data**: Realizar análises investigativas em grandes volumes de dados de valor desconhecido, aplicando uma combinação de [Machine Learning](http://www.cienciaedados.com/analise-de-sentimentos-e-machine-learning/" \t "_blank), estatísticas e técnicas de análise de gráfico SQL-on-Hadoop para descobrir novos conhecimentos e melhorar a análise operacional, tais como detecção de anomalia e sistemas de recomendação.

**Fácil acesso aos dados**: Os mais variados tipos de dados podem ser facilmente acessados em um EDH, garantindo uma fonte única para o trabalho de analytics.

**Armazenamento de dados em formato nativo**: Talvez uma das principais vantagens do EDH. Ao iniciar o trabalho de analytics, há a garantia que os dados estão em seu estado bruto, evitando distorção no processo de análise.

A seguir a imagem da arquitetura técnica do EDH:



**Camadas de dados no Data Lake**

Os produtos de desenvolvimento de código fonte entregues deverão disponibilizar os dados na plataforma Big Data em 3 camadas:

* Dados Brutos (Raw): cópia dos dados no mesmo formato da origem;
* Dados confiáveis (Trusted): dados padronizados, atendendo às regras de negócio definidas pelos usuários;
* Dados transformados (Business): dados modelados e otimizados para melhor performance de consumo da área usuária.

**Resumo dos principais componentes da plataforma de Big Data**

* **Portal Azure**

Portal Azure é o local unificado para a gestão dos recursos da nuvem Azure na assinatura do Big Data-CNI pelos administradores da conta.

* **Redis**

Redis é um banco de dados NoSQL com suporte a vários tipos de estrutura de dados e muito rápido no retorno das requisições. É comumente utilizado para cache e gravação de estado de objetos.

No projeto é utilizado de duas formas:

1. Nos Crawlers: banco chave-valor para guardar os status de coleta das fontes externas e permitir a retomada da coleta a partir do ponto de parada. Essa questão é crucial para APIs que fazer paginação, como FB e Youtube, por exemplo.
2. ADLS ALC Reporter: Cache para os dados dos usuários e grupos do AAD

* **DataFactory**

O Data Factory é um serviço de ETL de integração de dados baseado em nuvem.

No projeto ele é utilizado para executar os fluxos de automatização de cargas, desde ingestão dos dados, passando pela movimentação dos arquivos no DataLake, executando os scripts criados na esteira de produção via DataBricks para a transformação dos dados e gravando os dados do SQLDW.

* **SHIR (Self Hosted Integration Runtime)**

SHIR é a infraestrutura de computação usada pelo Azure Data Factory para fornecer funcionalidades de integração de dados entre diferentes ambientes de rede.

No projeto ele é utilizado em uma VM, conforme orientações da Microsoft, para realizar a ponte entre as bases de dados dos sistemas (OLTP) e a nuvem, ou seja, é uma espécie de servidor confiável que tem autorização de buscar os dados dos sistemas e disponibilizá-los na nuvem (via DataFactory) de forma segura.

* **Containers**

Container é uma tecnologia que padroniza, empacota e torna portável uma determinada aplicação. Ele agrupa o código de um aplicativo, seus respectivos arquivos de configuração e as bibliotecas necessárias para a sua execução.

No projeto os containers são utilizados para o desenvolvimento e a execução dos crawlers que buscam dados de sites externos (arquivos, raspagens, APIs, outros) para o Data Lake.

* **ADLS (Azure Data Lake Storage Gen1)**

ADLS é um repositório escalável para cargas de trabalho analíticas de big data. O Azure Data Lake permite a captura de dados de qualquer tamanho, tipo e velocidade de ingestão em um único lugar para análises operacionais e exploratórias.

No projeto ele é utilizado como o repositório oficial dos dados de toda a instituição, cujos arquivos podem ser acessados via DataBricks e DataFactory.

* **DataBricks**

O Databricks é uma plataforma de análise baseada no Apache Spark otimizada para a plataforma de serviços de nuvem. Ele possibilita o processamento dos dados em memória distribuída, tornando rápidas as soluções de cálculos complexas.

No projeto ele é utilizado como serviço de transformação de dados e de laboratório de ciência de dados. Para tanto é utilizada a linguagem python por meio da biblioteca pyspark.

* **SqlDW (atualmente chamado de** [**Azure Synapse Analytics**](https://azure.microsoft.com/services/synapse-analytics/)**)**

O Synapse é um sistema de data warehouse empresarial baseado em nuvem que aproveita o processamento paralelo maciço para executar rapidamente consultas complexas em petabytes de dados.

No projeto ele é utilizado para concentrar os conjuntos de dados agregados para a utilização no Tableau (ferramenta institucional de Self Service BI).

* **Tableau**

O Tableau é uma plataforma integrada de Self Service BI que permite ao próprio usuário de negócio, realizar por conta própria, a criação de painéis de análise e visualização de dados.

No projeto ele é utilizado exatamente conforme sua definição. Trata-se da ferramenta de BI institucional e possibilita ao próprio usuário da área de negócio realizar suas análises e divulgar seu trabalho.

* **Function**

O Azure Functions é um serviço de computação sem servidor que permite a execução de scripts de código disparado por meio de eventos e sem a necessidade de provisionar explicitamente ou gerenciar a infraestrutura.

No projeto as funções são utilizadas para integrar alguns componentes (auxiliar no cadastro de metadados, iniciar e finalizar as instâncias de containers, iniciar e finalizar os clusters) e realizar chamadas de serviços tais como crawlers e proxy.