Introdução

As empresas estão explodindo com dados, inclusive de bancos de dados e aplicativos, e com dados de streaming de comércio eletrônico, mídia social, aplicativos e dispositivos conectados na Internet das Coisas (IoT). Eles estão procurando maneiras de alavancar esses dados para transformar seus negócios - trazendo novos insights de negócios para criar futuros produtos e serviços, revolucionar o atendimento ao cliente, agilizar operações, descobrir novos fluxos de receita e muito mais.

As empresas estão percebendo que as tecnologias tradicionais não podem suportar suas transformação e não permitirá que atendam às novas necessidades de negócios. Como um Como resultado, muitas organizações estão modernizando suas plataformas de dados, voltando-se para o scale-out arquiteturas como lagos de dados.

Mas um lago de dados mal arquitetado também é limitante. Isso pode resultar em um pântano de dados onde os dados são despejados indiscriminadamente, proporcionando visibilidade limitada e, valor. Arquitetar um lago de dados para o sucesso requer uma abordagem cuidadosa. Este papel descreve uma arquitetura de referência para um lago de dados que pode definir você à direita caminho.

Este artigo irá discutir:

· Arquitetura de referência do lago de dados de Zaloni

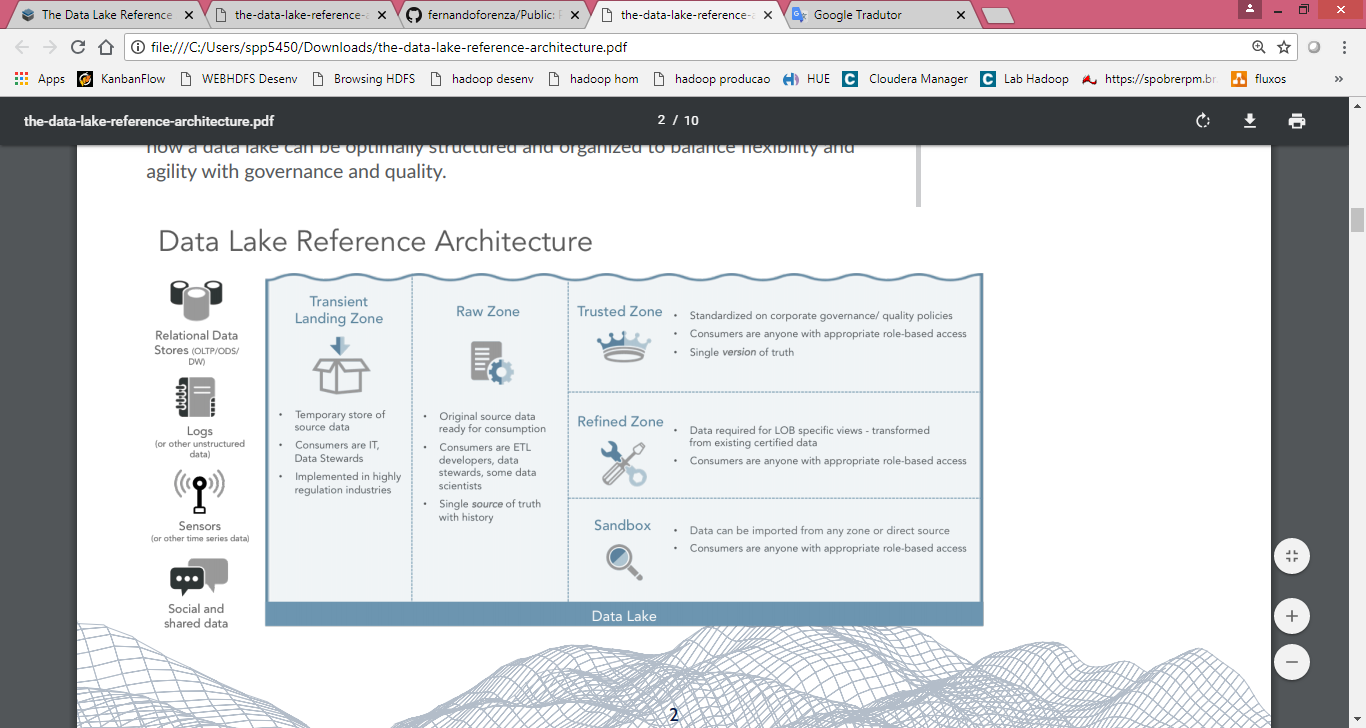
· Preparar o futuro da sua pilha de dados

· Exemplos de estudos de caso com fluxos de dados

Arquitetura de referência do Data Lake

Uma arquitetura de referência é uma estrutura que pode ser referenciada para 1) compreensão melhores práticas do setor, 2) acompanhar um processo, 3) fornecer um modelo para a solução, e 4) compreender estruturas e elementos.

A arquitetura de referência do lago de dados de Zaloni fornece uma visão funcional e mostra como um lago de dados pode ser otimamente estruturado e organizado para equilibrar a flexibilidade e agilidade com governança e qualidade.

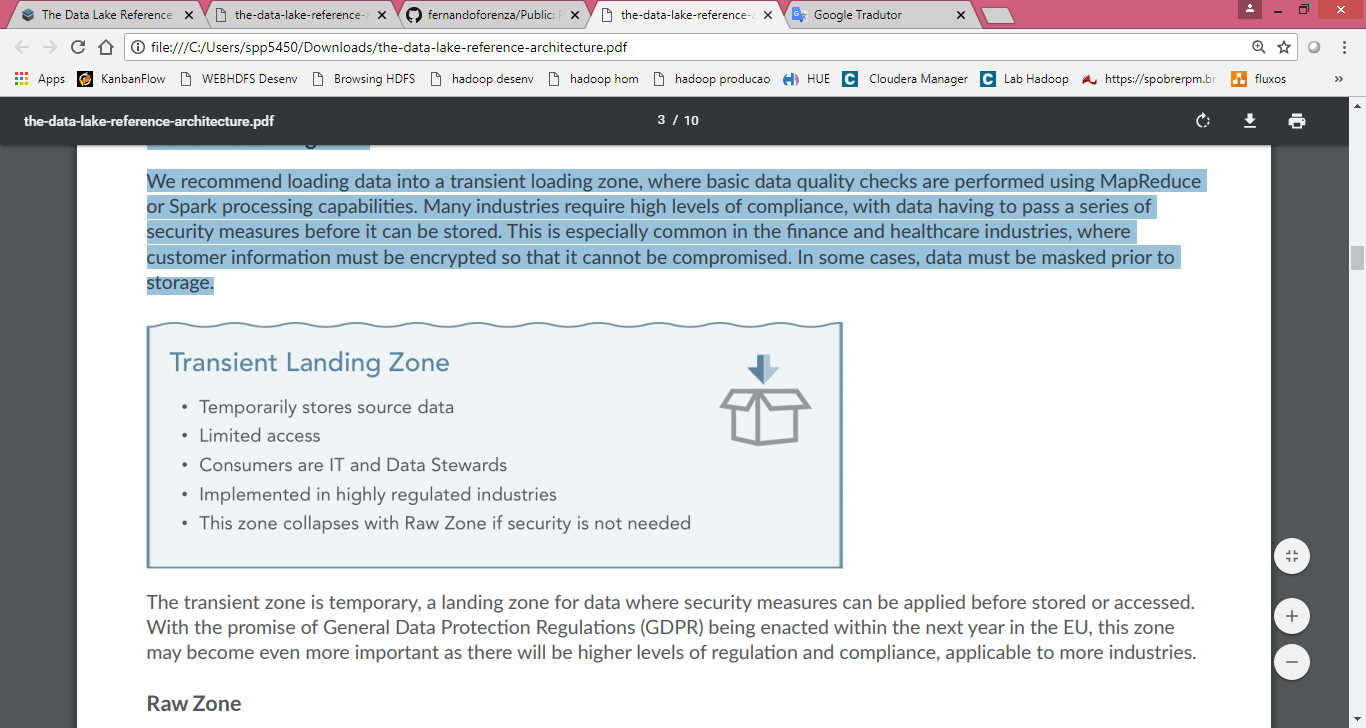


Nossa arquitetura de referência é organizada em quatro zonas, além de uma sandbox. Em todas as zonas, os dados são rastreados, validados, catalogados, metadados atribuídos, refinados e muito mais. Essas capacidades e as zonas em que ocorrem nos ajudam a entender em que estágio os dados estão e quais medidas foram aplicadas a eles até o momento.

A principal vantagem dessa arquitetura é que os dados podem entrar no data lake de qualquer lugar, incluindo transações on-line. processamento (OLTP) ou sistemas operacionais de armazenamento de dados (ODS), um DW, logs ou outros dados da máquina, ou de serviços na nuvem. Estes Os sistemas de origem incluem muitos formatos diferentes, como dados de arquivos, dados do banco de dados, ETL, fluxo de dados e até dados nas APIs.

Zona transitória

Recomendamos o carregamento de dados em uma zona de carregamento transitório, onde as verificações básicas de qualidade de dados são realizadas usando o MapReduce ou recursos de processamento do Spark. Muitas indústrias exigem altos níveis de conformidade, com dados tendo que passar por uma série de medidas de segurança antes que possam ser armazenadas. Isso é especialmente comum nas indústrias de finanças e saúde, onde as informações do cliente devem ser criptografadas para que não possam ser comprometidas. Em alguns casos, os dados devem ser mascarados antes de armazenamento.

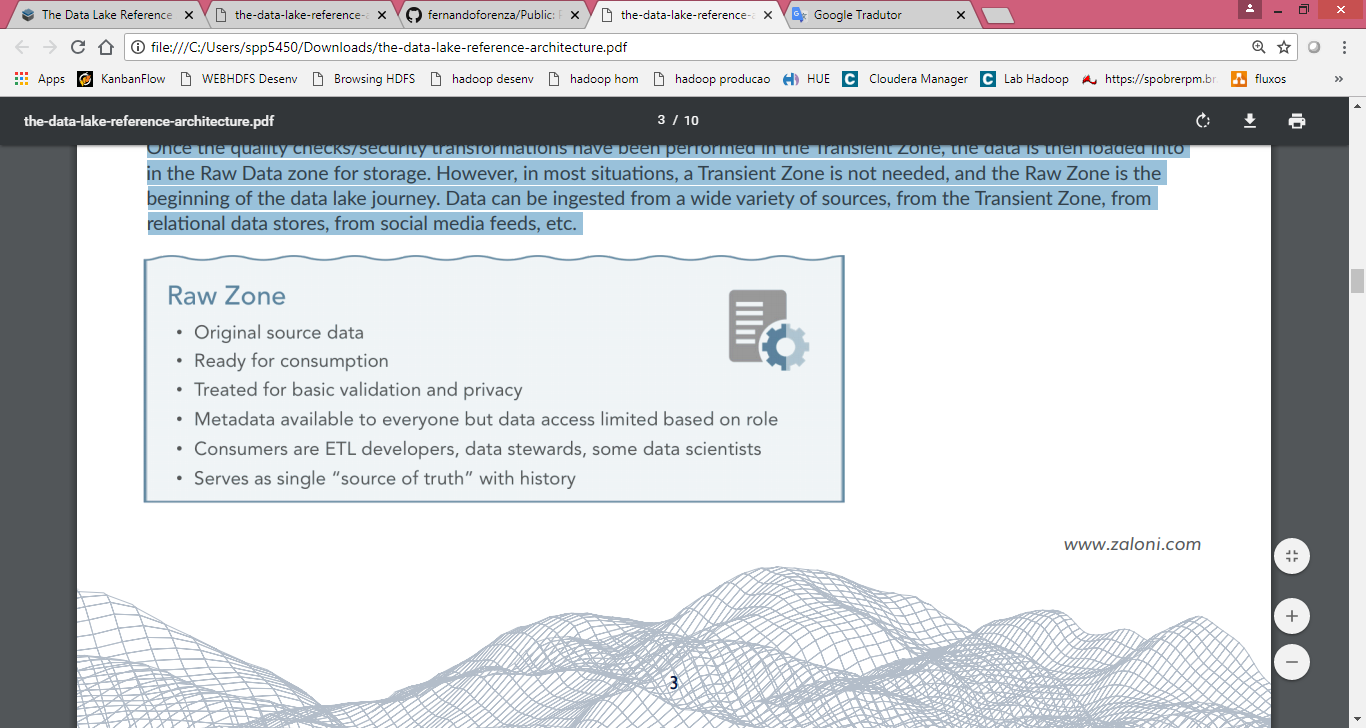


A zona transitória é temporária, uma zona de pouso para dados em que medidas de segurança podem ser aplicadas antes de serem armazenadas ou acessadas.

Com a promessa de que o Regulamento Geral de Proteção de Dados (GDPR) seja promulgado no próximo ano na UE, essa zona pode se tornar ainda mais importante, pois haverá níveis mais altos de regulamentação e conformidade, aplicáveis a mais indústrias.

Zona RAW

Assim que as verificações de qualidade / transformações de segurança tiverem sido realizadas na Zona Transiente, os dados serão carregados em na zona de dados brutos para armazenamento. No entanto, na maioria das situações, uma zona transitória não é necessária e a início da jornada do lago de dados. Os dados podem ser ingeridos a partir de uma ampla variedade de fontes, da Zona transitória, armazenamentos de dados relacionais, de feeds de mídia social etc.

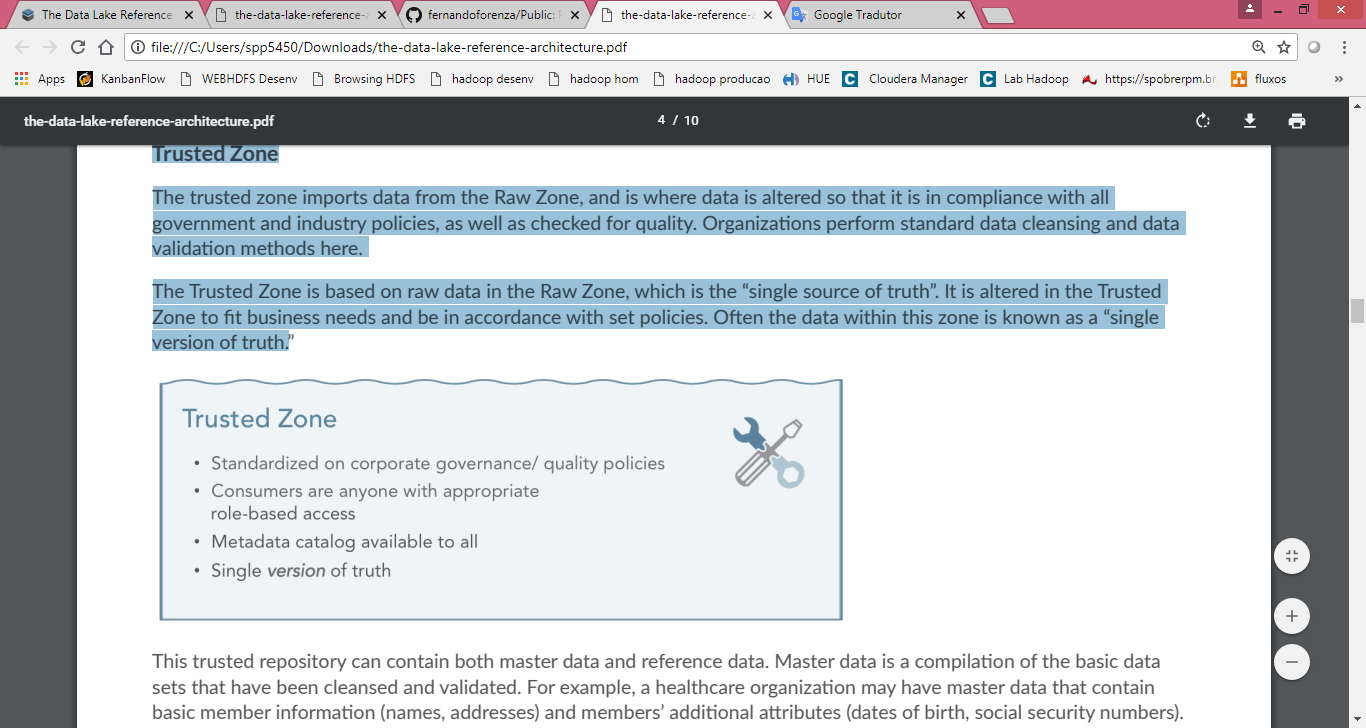


Dentro dessa zona, os dados são mascarados / tokenizados conforme necessário, adicionados aos catálogos e os metadados são aplicados. Na Raw Zone, dados é armazenado permanentemente e em sua forma original, por isso é conhecido como "a única fonte da verdade". Cientistas de dados e negócios analistas podem mergulhar nessa zona para conjuntos de dados a serem descobertos.

Zona Confiável

A zona confiável importa dados da Raw Zone e é onde os dados são alterados para que estejam em conformidade com todos políticas do governo e da indústria, bem como a verificação da qualidade. Organizações executam limpeza e dados padrão de dados métodos de validação aqui.

A Zona Confiável é baseada em dados brutos na Zona RAW, que é a "única fonte de verdade". Está alterado no Trusted Necessidade de negócios de zona a pé e estar de acordo com as políticas definidas. Muitas vezes, os dados dentro desta zona são conhecidos como “single versão da verdade”.



Esse repositório confiável pode conter dados mestre e dados de referência. Dados mestres são uma compilação dos dados básicos conjuntos que foram limpos e validados. Por exemplo, uma organização de saúde pode ter dados mestre que contenham informações básicas dos membros (nomes, endereços) e atributos adicionais dos membros (datas de nascimento, números de segurança social).

Uma organização precisa garantir que os dados mantidos na zona confiável estejam atualizados usando os mecanismos de captura de dados e alteração (CDC).

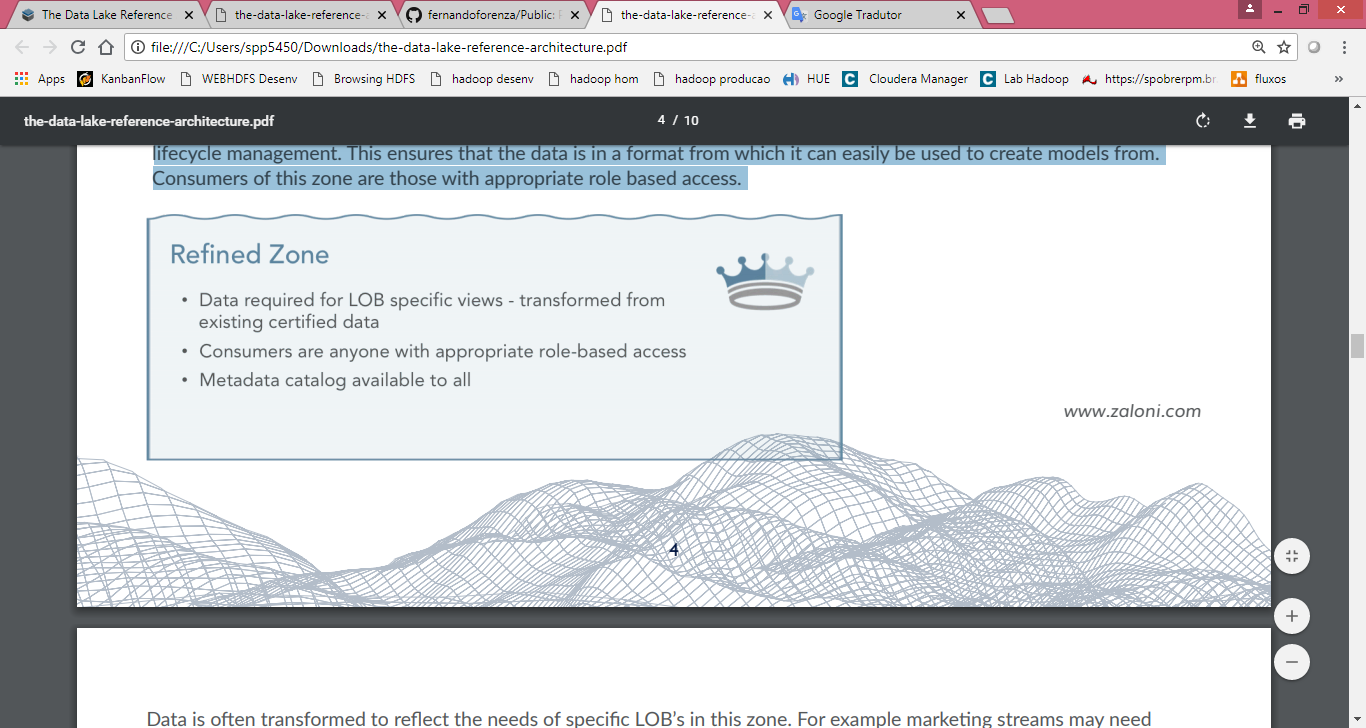
Os dados de referência, por outro lado, são considerados a versão única da verdade para conjuntos de dados misturados mais complexos. Para por exemplo, a organização de serviços de saúde pode ter um conjunto de dados de referência que mescla informações de várias origens. Tabelas no armazenamento de dados mestre, como os atributos de informações básicas e de membros adicionais do membro para criar uma “única versão da verdade” para dados de membros. Qualquer pessoa na organização que precise de dados de membros pode acessar esses dados de referência e sabe que eles podem depender disso.

Zona refinada

Dentro da Zona Refinada, os dados estão passando por suas últimas etapas antes de serem usados ​​para obter insights. Os dados aqui são integrado em um formato comum para facilidade de uso, e passa por detokenizaton possível, verificações de qualidade adicionais, e

Gerenciamento do ciclo de vida. Isso garante que os dados estejam em um formato a partir do qual possam ser facilmente usados ​​para criar modelos.

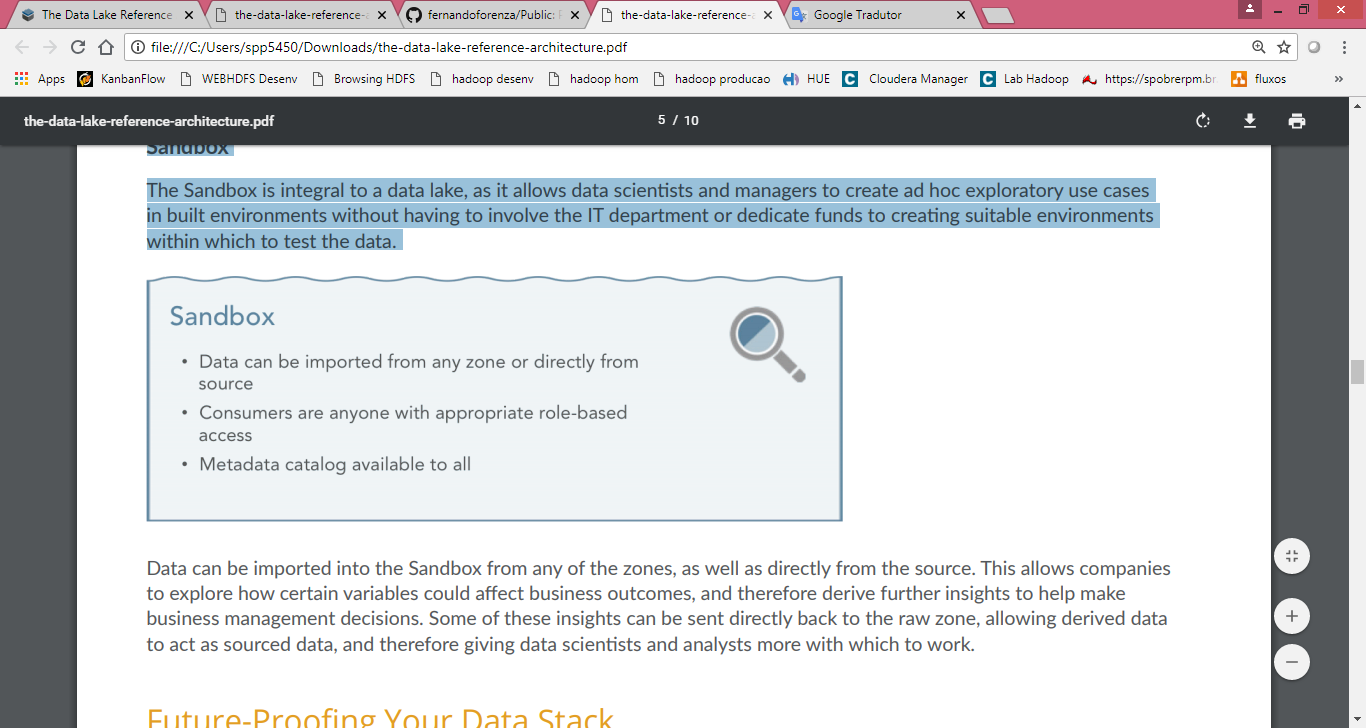
Os consumidores desta zona são aqueles com acesso adequado baseado em função.



Os dados costumam ser transformados para refletir as necessidades de LOBs específicos nessa zona. Por exemplo, os fluxos de marketing podem precisar para ver o ROI de certos compromissos para avaliar seu sucesso, enquanto os departamentos financeiros podem precisar de informações apresentado sob a forma de balanços. Os consumidores desta zona incluem aqueles com acesso baseado em funções, que podem ser determinado pela empresa.

SandBox

O Sandbox é parte integrante de um lago de dados, pois permite que os cientistas e gerentes de dados criem casos de uso exploratórios ad hoc em ambientes construídos sem ter que envolver o departamento de TI ou dedicar fundos para criar ambientes adequados dentro do qual testar os dados.



Os dados podem ser importados para o Sandbox de qualquer uma das zonas, bem como diretamente da origem. Isso permite que as empresas para explorar como certas variáveis podem afetar os resultados do negócio e, portanto, obter mais informações para ajudar nas decisões de gestão de negócios. Algumas dessas informações podem ser enviadas diretamente de volta para a zona não processada, permitindo dados derivados para atuar como dados de origem e, portanto, fornecer mais dados aos cientistas e analistas para trabalhar.

Evolução futura da sua pilha de dados

Determinar quais tecnologias empregar ao construir sua pilha de dados é uma tarefa complexa. Você deve:

Considere armazenamento, processamento, gerenciamento de dados, etc.

No passado, a maioria dos lagos de dados residia no local. Isso passou por uma tremenda mudança recentemente, com a maioria das empresas olhando para a nuvem para substituir ou aumentar suas implementações. Armazenamento no local, armazenamento em nuvem, nuvem múltipla e modelos híbridos são todos possíveis com muito poucos ajustes na arquitetura de referência correspondente. Muitas vezes o funcional as arquiteturas parecem geralmente as mesmas, enquanto as arquiteturas de componentes são alteradas para refletir as plataformas de armazenamento em nuvem e as aplicações com as quais eles são mais compatíveis.

Enquanto o armazenamento e o processamento nas instalações fornecem um controle mais rígido sobre segurança de dados e privacidade de dados, os sistemas de nuvem pública oferecer recursos de armazenamento e computação altamente dimensionáveis ​​e elásticos para atender às necessidades das empresas de processamento em larga escala e armazenamento de dados sem ter as despesas gerais de aprovisionamento e manutenção de infraestrutura dispendiosa. Além disso, com o ferramentas e tecnologias em constante mudança no ecossistema, os lagos baseados em nuvem também podem ser usados ​​como incubadora de Dev / test ambientes para avaliar todas as novas ferramentas e tecnologias em um ritmo rápido antes de escolher o caminho certo para colocar em produção, seja na nuvem ou no local.

Armazenamento do Data Lake:

Para os lagos de dados locais, o HDFS parece ser o armazenamento de escolha, pois fornece dados distribuídos com replicação. Isso permite um processamento mais rápido de casos de uso de big data. O HDFS também permite que as empresas criem ters de armazenamento para permitir gerenciamento do ciclo de vida de dados, alavancando esses custos para economizar custos, mantendo políticas de retenção de dados e requisitos.

O armazenamento em nuvem oferece uma vantagem única, pois permite o armazenamento de dados dissociados da necessidade de computar, permitindo assim que as empresas economizem nos custos de processamento e alavanquem diferentes poderes de demandas de casos de uso. O armazenamento na nuvem também permite usar o armazenamento limitado para otimizar o custo e retentar os dados e regulamentar requisitos.

Processamento de Data Lake:

O Hadoop tem sido central para os lagos de dados no local, pois permite o processamento distribuído de grandes conjuntos de dados processamento de clusters para a empresa. Ele também pode ser implantado em um lago de dados baseado em nuvem para permitir um lago de dados híbrido usando uma única distribuição (por exemplo, Cloudera, Hortonworks e MapR).

O Apache Spark fornece um mecanismo mais rápido para processamento de dados em grande escala, aproveitando a computação na memória. Pode ser executado em Hadoop, Mesos, na nuvem, ou em um ambiente independente para criar uma camada de computação unificada em toda a empresa.

O Apache Beam fornece um resumo sobre o cluster de processamento. Ao usar o Beam, as empresas podem desenvolver suas pipelines de processamento de dados usando o Beam SDK e, em seguida, escolha um Beam Runner para executar o pipeline em uma grande escala específica.

sistema de processamento de dados. O corredor pode ser qualquer coisa desde um Direct Runner, Apex, Flink, Spark, Datafow e Gearpump.

(incubação). Esse design permite que o duto de processamento seja portátil entre os diferentes corredores,

fexibilidade para as empresas alavancar a melhor plataforma para atender aos requisitos de processamento de dados em uma

caminho.

Gestão de dados:

A necessidade de gerenciar dados na variada pilha de tecnologia de um lago de dados amplia a importância de dados unificados plataforma de gestão. Uma robusta controladora de gerenciamento de dados permite que as empresas gerenciem seus dados em vários armazenamento, computação e processamento de camadas, mantendo um controle claro dos dados em todo o seu ciclo de vida. Além disso, pode automatizar o movimento e o processamento de dados dentro e entre as zonas.

Isso não apenas fornece uma maneira eficiente e rápida de obter insights, como também permite que as empresas atendam às suas regulamentações.

requisitos em relação à privacidade, segurança e governança de dados. Abaixo estão quatro principais recursos de gerenciamento de dados que devem ser considerado quando arquitetar um lago de dados da próxima geração.

Gerenciamento de metadados:

Catalogar metadados no lago de dados é essencial. Metadados permite uma maneira de categorizar dados e fornece contexto para quando foi enviado, de onde veio e que linhas de negócios para as quais é relevante. Isso permite que um para mais facilmente consultar dados, bem como organizá-los. Os metadados servem de base para todo o processamento e controle de dados em todo o zonas.

Segurança:

Protegendo dados no lago é critcal. Você pode aproveitar o acesso baseado em função e pode fazê-lo dentro das zonas. Além disso, mascarando e tokenizaton são duas formas comuns de garantir que as informações sensíveis permaneçam em segredo até que tenha sido analisada da maneira correta e incorporada nos modelos certos. O mascaramento envolve a criação de um “estruturalmente semelhante, mas inauthentc ”de dados para que possa ser usado para teste e treinamento. Tokenizaton, envolve a substituição de dados protegidos com um valor que se refere aos dados sem expô-lo.

Qualidade dos dados:

Isto se refere a assegurar que os dados sejam tão reflexivos quanto possível, e que haja uma fonte de verdade que possa ser referida como por todas as LOBs, mesmo se as entradas iniciais fossem mais variáveis.

Gerenciamento de Ciclo de Vida de Dados:

O gerenciamento do ciclo de vida dos dados envolve designações frias, quentes e quentes, nas quais se rastreia quanto tempo os dados estado no lago de dados e por quanto tempo ele deve continuar sendo antes de ser reciclado.

Para abordar esses recursos nas empresas de data lake, considere uma plataforma de gerenciamento de dados, como a Zaloni Plataforma de Gerenciamento de Dados (ZDP).

Estudos de caso com fluxos de dados

As empresas estão dando passos sérios modernizando suas arquiteturas de dados com lagos de dados, resolvendo problemas antigos de novas maneiras e criando novas oportunidades para melhorar seus negócios, sua fidelidade de clientes e sua competitividade vantagem. Abaixo descrevemos dois estudos de caso distintos que alavancaram uma arquitetura de referência para construir um Data Lake e foram bem recompensados ​​pelo esforço.

Uma dessas implementações foi no local e a outra utilizou uma estratégia em nuvem na AWS. Para cada um Nestes exemplos, mostraremos um fluxo de dados para explicar quais ferramentas são necessárias para realizar um caso de uso específico.

Nesses fluxos de dados, processos como ingestão, armazenamento, operações e segurança são abordados, assim como subseções diferentes, como operações de processamento em lote ou de painel. Para cada uma dessas subseções, recomendações são feitas e a tecnologia é montada para garantir uma implementação perfeita.

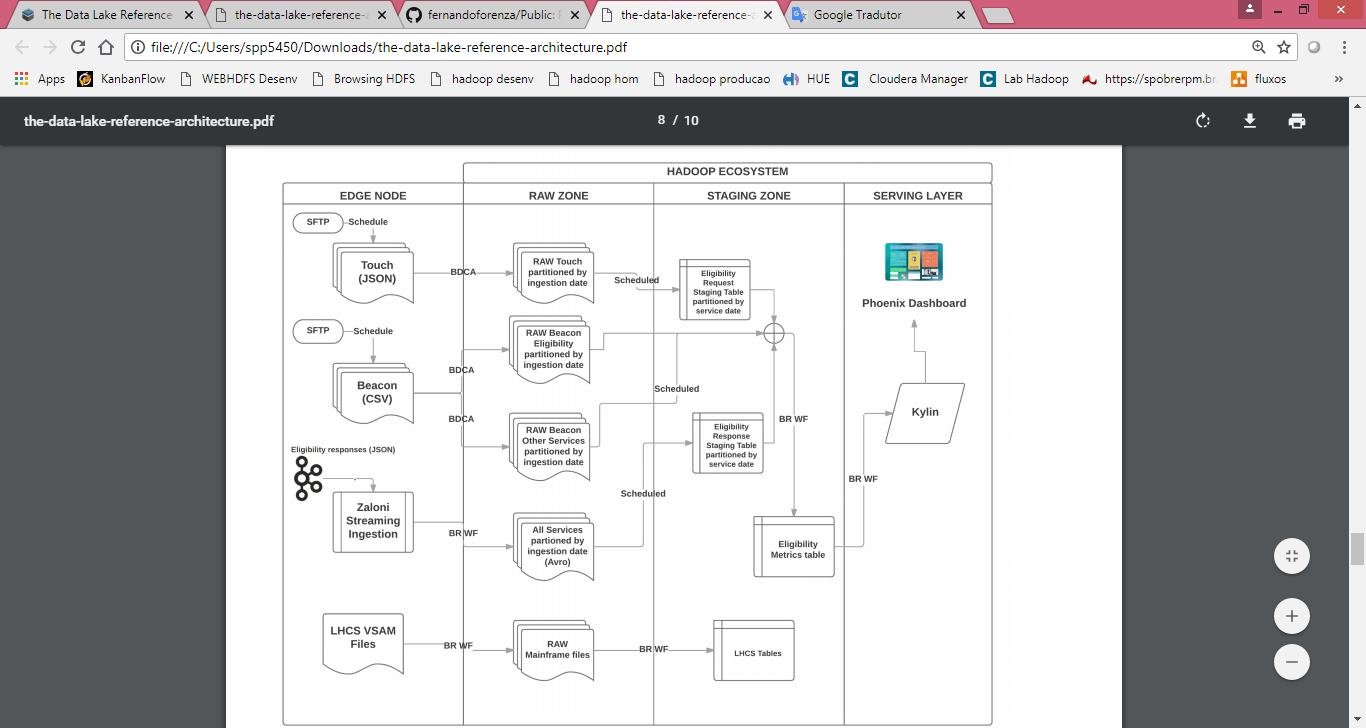
Lago de dados no local para diagnósticos de assistência médica

Uma das maiores empresas de diagnósticos de saúde do mundo, tinha um grande ambiente de mainframe para lidar com altos volumes de dados de pacientes usados ​​em relatórios para milhares de clientes. O ambiente era inflexível, impedindo a

a empresa de ser capaz de escalar para adicionar novos clientes ou aumentar a capacidade de computação para acelerar os relatórios para atender demanda do cliente existente para relatórios mais frequentes. O cliente também não conseguiu gerar relatórios personalizados para solicitações de; por exemplo, um relatório para um paciente específico. O cliente precisava de uma nova infraestrutura para automatizar processos e dimensionar a capacidade de processamento conforme necessário.

Com o data lake: Zaloni adotou uma abordagem de várias etapas para o projeto. Primeiro a equipe construiu um lago de dados e o colocou produção, alavancando o ZDP para ingestão, catalogação, gerenciamento de dados e governança. Novos dados do cliente foram ingeridos

no data lake, e um código personalizado foi desenvolvido para gerar relatórios de acordo com as preferências dos clientes. O cliente agora começou a adicionar clientes existentes à nova plataforma, ingerindo arquivos de mainframe / dados históricos no Hadoop. Fluxo da arquitetura de referência: o lago de dados está operando dentro de um ecossistema Hadoop maior com recursos que variam da ingestão para processamento em lote e fluxo. Escolhendo uma implementação local, o cliente utilizou o sistema HDFS baseado em java como seu sistema de arquivos distribuído e uma forma de armazenamento de dados, bem como uma infinidade de outros programas.



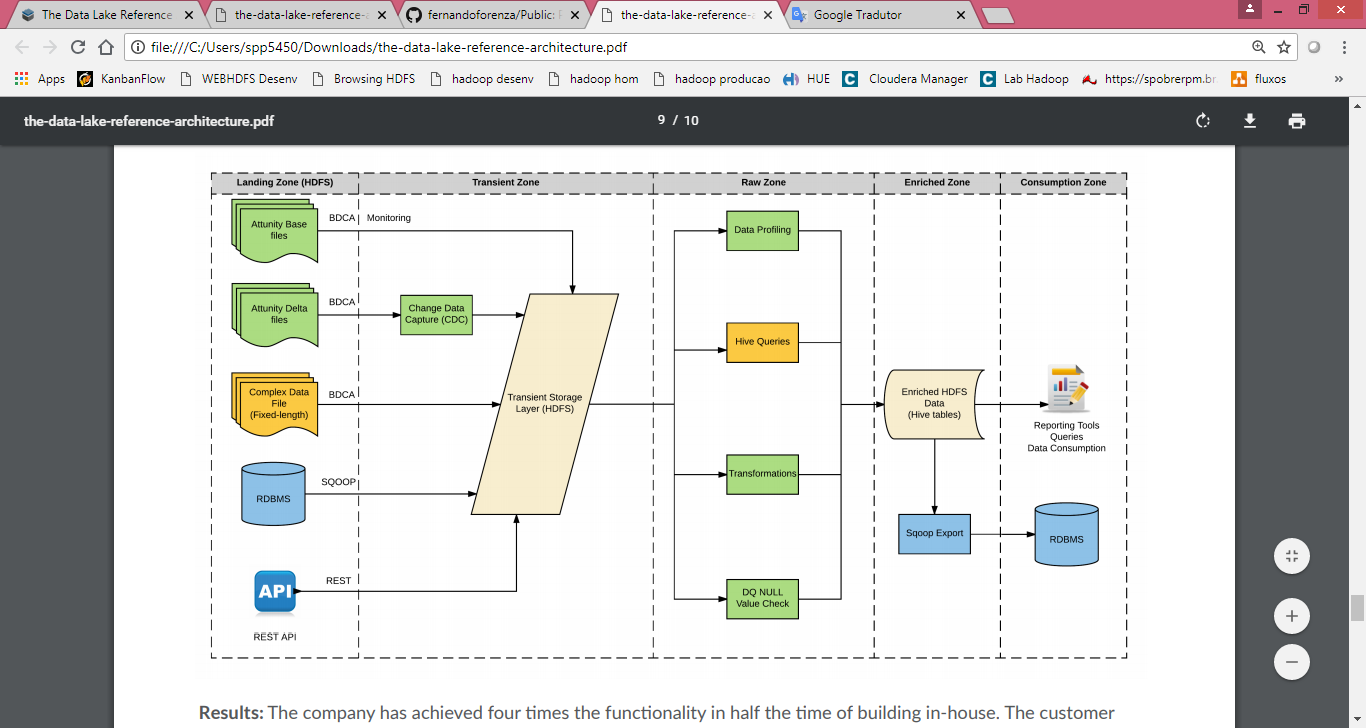
Resultados: Com flexibilidade e agilidade de relatórios significativamente maiores, o cliente é capaz de atender à demanda dos clientes diariamente. relatórios. A empresa percebeu custos reduzidos de processamento, armazenamento, licenciamento e manutenção de dados, e terá

capacidade quase ilimitada de adicionar novos clientes no futuro. Lago de dados em nuvem para iniciativa do Customer 360 em empresa de publicação Uma das maiores editoras da América do Norte queria aumentar seu modelo de negócios monetizando seus 45 anos de dados do assinante para permitir uma abordagem do Customer 360 e fornecer informações de marketing aos seus clientes. A empresa tinha décadas de valiosos dados de assinantes de suas revistas, assinaturas on-line e materiais de marketing, inexplorados e sentado em sistemas de silos em toda a organização. Não foi possível combinar esses grandes volumes de dados internos com dados de terceiros para criar fontes de receita adicionais a jusante.

Com o data lake: Um novo data lake baseado em nuvem forneceu um repositório centralizado para dados internos e de terceiros.

Em apoio à estratégia Customer 360 da empresa, o data lake permitiu que a empresa coletasse dados de todos os seus sistemas e gerenciar e governá-lo com uma plataforma de dados unificada, ZDP.

Fluxo da arquitetura de referência: a empresa optou por hospedar seu data lake na nuvem, utilizando o Simple Storage da Amazon Atendimento como armazenamento de dados e Redshift como armazenamento de dados relacional. A empresa escolheu implementar uma zona transitória como os dados podem ser protegidos e qualificados antes de serem permanentemente armazenados.



Resultados: A empresa alcançou quatro vezes a funcionalidade na metade do tempo de construção interna. O cliente espera obter economias significativas devido à redução do tempo de lançamento no mercado e à capacidade de dimensionar dados com perfeição gerenciamento e análise na velocidade do negócio.

Arquitetando seu lago de dados para o sucesso

Os lagos de dados têm uma série de habilidades que podem atender a uma ampla variedade de casos de uso e indústrias em várias linhas de negócios. No entanto, para extrair o valor máximo do lago de dados, é crucial garantir que eles estejam arquitetado e gerenciado.

Usar uma arquitetura de referência é uma boa maneira de estruturar um lago de dados, bem como planejar sua implementação e funções. A arquitetura de referência do lago de dados Zaloni serve como um bom modelo funcional a partir do qual desenhar.

À medida que você arquitetar seu lago de dados, independentemente do modelo de implantação, lembre-se de abordar o gerenciamento de dados e governança. Sugerimos uma plataforma integrada, como o Zaloni Data Management Platform (ZDP), para construir um Data Lake que pode atender às suas necessidades de negócios hoje e que pode escalar com você no futuro.

A Zaloni Data Platform (ZDP) é uma plataforma de dados de autoatendimento premiada que fornece os recursos necessários para gerenciamento, governança e autoatendimento para fornecer um data lake pronto para produção que pode ser implantado em ambientes com várias nuvens. A plataforma operacionaliza o gerenciamento de dados e elimina silos de dados para o gerenciamento central de todas as fontes de dados corporativas, independentemente do local, e oferece um catálogo de dados de autoatendimento em toda a empresa, por meio do qual descubra e envolva conjuntos de dados e obtenha insights de negócios transformacionais usando análises avançadas.

Simplifique e amplifique sua infraestrutura de dados moderna

Nossa plataforma totalmente integrada minimiza a dor de modernizar sua infraestrutura de dados e ajuda a otimizar sua

pilha de dados complexa e fragmentada:

• Aproveita qualquer loja, qualquer distribuição, qualquer implantação - nuvem, multi-nuvem, local ou híbrida

• Integra-se facilmente à sua arquitetura de dados existente

• Simplifica o uso, o gerenciamento e a integração de tecnologias e ferramentas de big data em constante mudança

Benefícios da plataforma

Os benefícios do ZDP são significativos. Nas várias implementações de clientes da Zaloni, foi constatado que a Zaloni faz isso

75% mais rápido para implantar um lago de dados. Além disso, pesquisas de terceiros mostraram que alavancar a plataforma da Zaloni pode resultar em

um retorno de investimento de 650% versus a implementação de uma solução DIY de ferramentas legadas e domésticas.

VALOR DE VELOCIDADE DE ESCALA

Acelere o tempo e o esforço para

construa o lago de dados em 75%.

Reduza o tempo de percepção de

semana a minutos.

Desfrute de um retorno de 650% sobre

investimento vs. construa o seu próprio.

Sobre o Zaloni

Operacionalize seu lago de dados. Acelere os insights de negócios.

O Zaloni simplifica o big data para insights de negócios transformadores. Trabalhamos com empresas pioneiras para modernizar seus dados

Arquitetar e operacionalizar seus lagos de dados para incorporar dados às práticas comerciais cotidianas. A Plataforma de Dados Zaloni

(ZDP) fornece controle total em todo o fluxo de dados, desde a ingestão até a análise, com gerenciamento de dados abrangente,

capacidades de preparação de dados de governança e autoatendimento para usuários de TI e de negócios. Um líder em big data por mais de uma década,

A experiência da Zaloni é profunda, abrange vários setores e se mostrou inestimável para os clientes em muitas das principais empresas do mundo.

Estamos orgulhosos de sermos reconhecidos pela lista Big Data 100 da CRN em 2018, pelas 20 maiores empresas de big data da Forbes, e pela Red Herring.

Prêmio Top 100 da América do Norte.