UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
CURSO DE ESPECIALIZAÇÃO EM INDÚSTRIA 4.0

LUCIANO LEMES

DATA LAKE INDUSTRIAL

Ponta Grossa

2020

LUCIANO LEMES

DATA LAKE INDUSTRIAL

Trabalho de Conclusão de Curso de Especialização apresentado como requisito parcial à obtenção do título de Especialista em Industria 4.0, da Universidade Tecnológica Federal do Paraná, Campus Ponta Grossa.

Orientador: Prof. Diego Roberto Antunes

Ponta Grossa

2020

Ministério da Educação
UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
CÂMPUS PONTA GROSSA
Departamento Acadêmico de Engenharia de Produção

TERMO DE APROVAÇÃO DE TCCE

Data Lake Industrial

Luciano Lemes

Este Trabalho de Conclusão de Curso de Especialização (TCCE) foi apresentado em oito de fevereiro de 2020 como requisito parcial para a obtenção do título de Especialista em Industria 4.0. O candidato foi arguido pela Banca Examinadora composta pelos professores abaixo assinados. Após deliberação, a Banca Examinadora considerou o trabalho aprovado.

_____
Prof. Diego Roberto Antunes
Prof. Orientador

_____
Prof. Max Mauro Dias Santos
Membro titular

_____
Prof. Dr. Marcelo Vasconcelos de Carvalho
Membro titular

RESUMO

Com o aumento massivo na geração de dados e a necessidade de vantagens competitivas no setor industrial, o uso de ferramentas de Big Data para a análise de dados e para o suporte no processo de tomada de decisão mostra-se fundamental para o crescimento das companhias. Neste contexto, este trabalho apresenta o processo de criação de um Data Lake para uma indústria de papel e celulose. A arquitetura em nuvem desenvolvida promoveu a solução de problemas tais como: Baixa performance, segurança na rede de automação, sendo necessário a liberação de usuários para acesso a rede industrial, dificuldade na criação de novos relatórios gerenciais, devido a várias bases de dados espalhado pelo ambiente de automação, solucionando problemas de alto consumo de Hardware nos servidores, devido aos múltiplos acessos que eram realizados simultaneamente. Os resultados obtidos com a solução proposta incluem: A criação de um Data Lake com uma arquitetura corporativa da Companhia, sendo possível cruzar dados de diversos sistemas e bases para elaboração de relatórios, visando a melhoria e maior confiabilidade da planta fabril, tornando o ambiente e a infraestrura escalável e mais segura, possibilitando maior performance na elaboração dos relatórios de fábrica para tomada de decisões. Os resultados foram observados instantaneamente, com a alta performance na criação e acesso mais simplificado aos relatórios, dando maior visibilidade e segurança nas informações adquiridas através da extração dos dados de alarmes e forces dos equipamentos. Levando a satisfação do cliente interno e a confiança no desenvolvimento de novos projetos utilizando as tecnologias e conceitos de Big Data, possibilitando novas descobertas, pesquisas e inteligências devido ao alto poder computacional que é a utilização e o tratamento dos dados em um ambiente escalável como é a nuvem.

Palavras-chave: Data Lake. Big Data. Inteligência Artificial. Nuvem. Automação. Redes industriais

ABSTRACT

With the massive increase in data generation and the need for competitive advantages in the industrial sector, the use of Big Data tools for data analysis and support in the decision-making process is essential for the growth of companies. In this context, this work presents the process of creating a Data Lake for a pulp and paper industry. The cloud architecture developed promoted the solution of problems such as: Low performance, security in the automation network, requiring the release of users to access the industrial network, difficulty in creating new management reports, due to several databases spread across the world. automation environment, solving problems of high consumption of Hardware on the servers, due to the multiple accesses that were performed simultaneously.

The results obtained with the proposed solution include: The creation of a Data Lake with a corporate architecture of the Company, making it possible to cross data from different systems and bases for reporting, aiming at the improvement and greater reliability of the manufacturing plant, making the environment and the scalable and more secure infrastructure, enabling greater performance in the preparation of factory reports for decision making. The results were observed instantly, with high performance in the creation and more simplified access to reports, giving greater visibility and security in the information acquired through the extraction of data from alarms and forces of the equipment. Taking internal customer satisfaction and confidence in the development of new projects using Big Data technologies and concepts, enabling new discoveries, research and intelligence due to the high computational power that is the use and treatment of data in a scalable environment as it is the cloud.

Keywords: Data Lake. Big Data. Artificial intelligence. Cloud. Automation. Industrial networks

SUMÁRIO

1 INTRODUÇÃO

Com geração intensa de dados e a necessidade de vantagens competitivas no setor industrial, o uso de ferramentas de Big Data para a análise de dados e para o suporte no processo de tomada de decisão mostra-se fundamental para o crescimento das companhias. Neste caso a criação de um Data Lake vem para atender a alta demanda de dados trafegados na rede sendo compilado em um único lugar de armazenamento, trazendo benefícios e facilitando o acesso de informações muitas vezes dispersas passando para um controle e gestão mais acessível destas informações, de forma mais rápida, variável e escalável, tirando qualquer necessidade e preocupação quanto a espaço em armazenamento acesso de vários lugares, possibilitando o cruzamento de dados e geração de múltiplos relatórios, com uma escala dinâmica para corresponder as prioridades do setor corporativo.

1.1. JUSTIFICATIVA

Modelo atual apresentava baixa performance e segurança na rede industrial, gerando dificuldades para gerar novos relatórios, sendo necessário a contratação de recursos terceiros para análise dos dados em várias bases espalhadas em diversos servidores e sistemas, sendo necessário permitir acessos a vários usuários na rede de automação comprometendo a segurança dos dados e do ambiente industrial.

1.2. OBJETIVO GERAL

- Construção de um Data Lake em nuvem para área de automação industrial, integrando várias fontes de dados da fábrica.

1.3. OBJETIVOS ESPECIFICOS

- Criar ETL (Extract, transform, load) para extrair, tratar e enviar os dados para nuvem;
- Criar relatórios gerenciais de Alarmes de equipamentos da indústria;

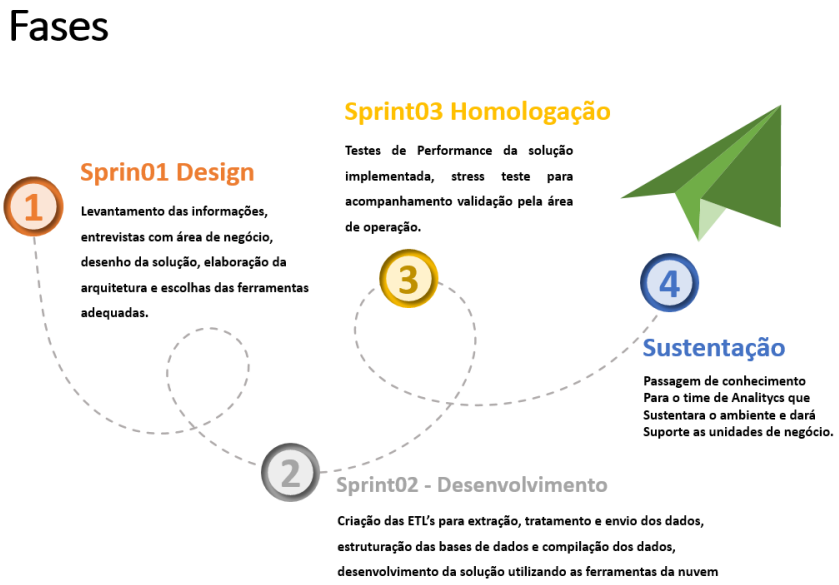
- Relatório de Gestão de ativos das áreas da Linha de fibras, Recuperação e Utilidades e Máquinas de celulose;
- Cruzar dados de diversos sistemas da empresa;
- Criar um modelo de arquitetura para Data Lake corporativo na companhia;
- Criar um Roadmap de atualizações futuras.

1.4. METODOLOGIA

Foi utilizado metodologia Ágil para o desenvolvimento do projeto, divido todo o cronograma em sprints semanais (**Figura 1**), realizando entregas mais rápidas com maior qualidade, flexibilidade de escopo do projeto, entregas rápidas gerando valor de acordo com as necessidades dos usuários, como principais focos:

- Foco no cliente;
- Trabalhar com pequenos avanços incrementais, chamados de interações;
- Testes de progresso e validação.

Figura 1, fases de execução do projeto



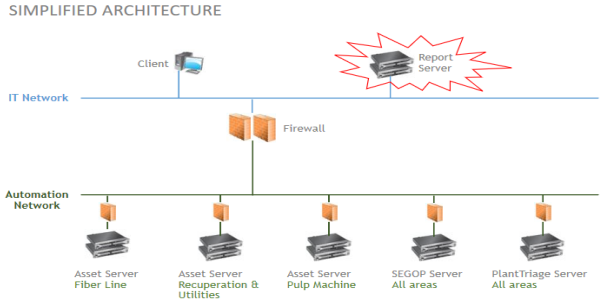
Autoria própria

2 DESENVOLVIMENTO

2.1. Situação atual

Com a grande massa de geração de dados o formato atual sofre como o modelo de utilização dos usuários, múltiplos acessos simultâneos, afetando a operação e performance do sistema de gerenciamento de alarmes e forces, chegando a momento de intertravamentos e interrupções dos sistema, dificultando a exploração, compilação dos dados e emissão de relatórios para analises e tomadas de decisões, conforme a arquitetura simplificada na **figura 2**, mostra como o acesso é realizado, usuários conectam de uma rede corporativa, passando por uma camada de Firewall, separando as redes de negócio da rede industrial, permitindo liberações de acesso, acessando diretamente as fontes de dados na raiz, correndo o risco de corrompimento de algum sistema ou banco de dados, os Asset Servers (Fiber Line, Recuperation & Utilities, Pulp Machine) mencionados na imagem, são correspondentes aos servidores que gerenciam todos os ativos pertencentes as áreas de Linha de Fibras, Recuperação e Utilidades e Maquinas de celulose, SEGOP SERVER é referente aos alarmes e forces dos equipamentos, trazendo informações de segurança operacional da fábrica. PlantTriage Server monitora continuamente sua planta para identificar problemas sempre que eles ocorrerem. Em seguida, ele prioriza as informações com base em fatores técnicos. A solução ajuda a encontrar a causa raiz dos problemas e fornece um conjunto completo de ferramentas de análise, de modo que você possa se aprofundar e solucionar a causa do problema. O PlantTriage também inclui uma poderosa ferramenta de sintonia de malhas integrada e análise de instrumentos e válvulas, de maneira que você possa escolher os melhores parâmetros para otimizar seu processo.

Figura 2, arquitetura simplificada da rede industrial



Autoria própria

2.2. Modelo desenvolvido

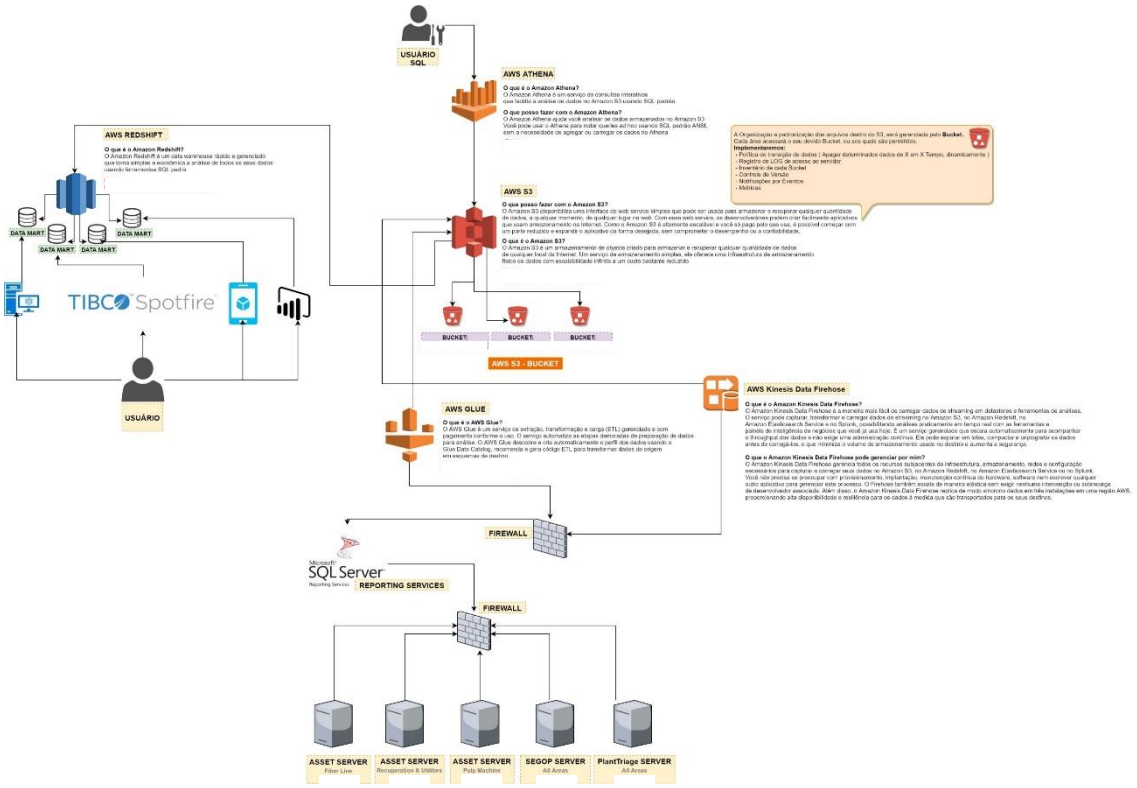
Modelo de arquitetura em nuvem proposto de compilação dos dados (**Figura 3**), unificando em base uniforme, o Glue (AWS Glue, 2020) é um serviço de extração, transformação e carga (ETL) gerenciado que facilita a preparação e a carga de dados para análises pelos clientes. Você pode criar e executar uma tarefa de ETL com apenas alguns cliques. Basta indicar ao Glue os dados armazenados na nuvem que ele os descobre e armazena os metadados associados. Uma vez catalogados, os dados são disponibilizados imediatamente para pesquisas, consultas e relatórios.

Utilizando como solução em armazenamento o modelo (Simple Storage Service AWS S3, 2020) para armazenar qualquer tipo e volume de dados em grande escala e variedade em casos de usos. O AWS S3 fornece recursos de gerenciamento fáceis de usar, de maneira que você possa organizar os dados e configurar os controles de acesso refinados para atender a requisitos específicos conforme a sua necessidade.

O Athena (AWS Athena, 2020) é um serviço de consultas interativas que facilita a análise de dados usando SQL padrão, não precisa de servidor, portanto, não há infraestrutura para gerenciar. A solução permite criar um repositório de metadados unificado em vários serviços, realizar uma pesquisa avançada nas fontes de dados para descobrir esquemas e preencher o catalogo com definições novas e modificadas de tabelas e partições, além de manter o versionamento do esquema. O AWS Redshift permite executar consultas de alta performance em petabytes de dados estruturados com facilidade e economia para criar relatórios e painéis avançados usando as ferramentas de inteligência de negócios que a companhia possua, segmentando cada área da companhia por DATA MART para que que exista mais facilidade nas análises e consultas dos dados de forma mais eficaz e rápida.

Para a transferência, coleta, processamento e a análise de dados de streaming em tempo real, foi utilizado o Kinesis (AWS Kinesis, 2020) permitindo que você obtenha insights oportunos e reaja rapidamente às novas informações, é um recurso essencial para processar dados de streaming em qualquer escala de forma econômica, além da flexibilidade de escolher as ferramentas mais adequadas aos requisitos dos aplicativo, conforme ilustração na **figura 3**.

Figura 3, arquitetura de Data Lake em Nuvem desenvolvida



Autoria própria

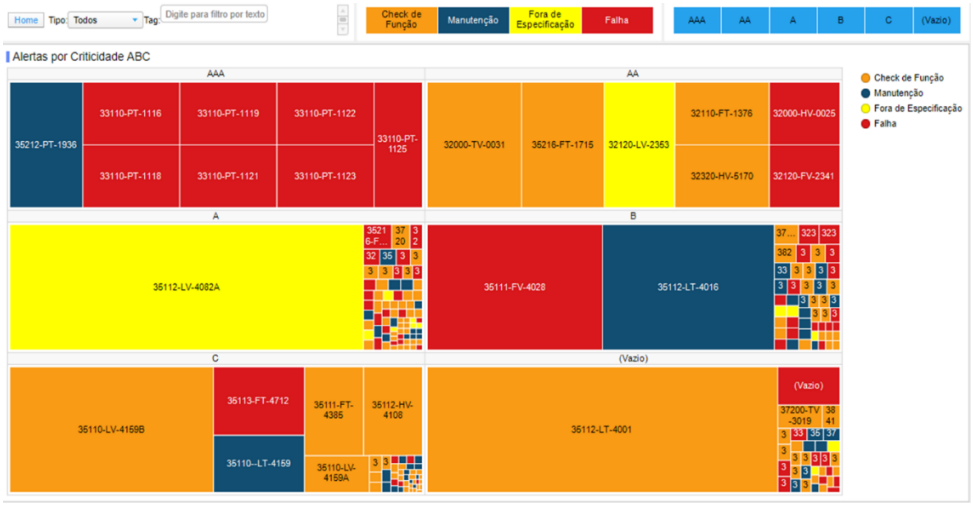
2.3. Os resultados Obtidos

Relatórios para tomada de decisões extraídos de forma bem mais rápida devido aos dados necessários para realizar análise de falha de equipamentos de criticidades A, B e C estarem compilados em nuvem, conforme **figura 4,** podemos observar o status de cada equipamento monitorado, na cor laranja podemos verificar Check de função, equipamentos em azul mostram que já estão em manutenção são classificados com Criticidade C, em amarelo criticidade B são equipamentos que podem afetar áreas e reduzir a performance da fábrica gerando problemas e diminuindo a produtividade, já os em vermelho com um nível A, maior de criticidade, são equipamentos bem críticos que podem causar a parada de um setor ou de toda a planta industrial se a falha não for corrigida a tempo. As TAGs de numeração descritas na tela indicam informações da área onde aquele determinado equipamento pertence, Ex: 32120-LV-2353 podendo o usuário clicar sobre o quadrado e obter mais informações técnicas daquele determinado equipamento como informações e dados de processo.

Como ferramenta analítica para desenvolver a dashboard foi utilizado o TIBCO SPOTFIRE como solução para construção do relatório de forma simplificada e rápida, conectado diretamente no Data Lake desenvolvido em Nuvem, gerando as informações necessárias em apenas alguns segundos, sendo alimentado em tempo real.

Permitindo a customização de diversos modelos de relatórios, de acordo com o perfil e gosto dos usuários.

Figura 4, relatório de alarmes dos equipamentos de criticidade A, B e C



Autoria própria

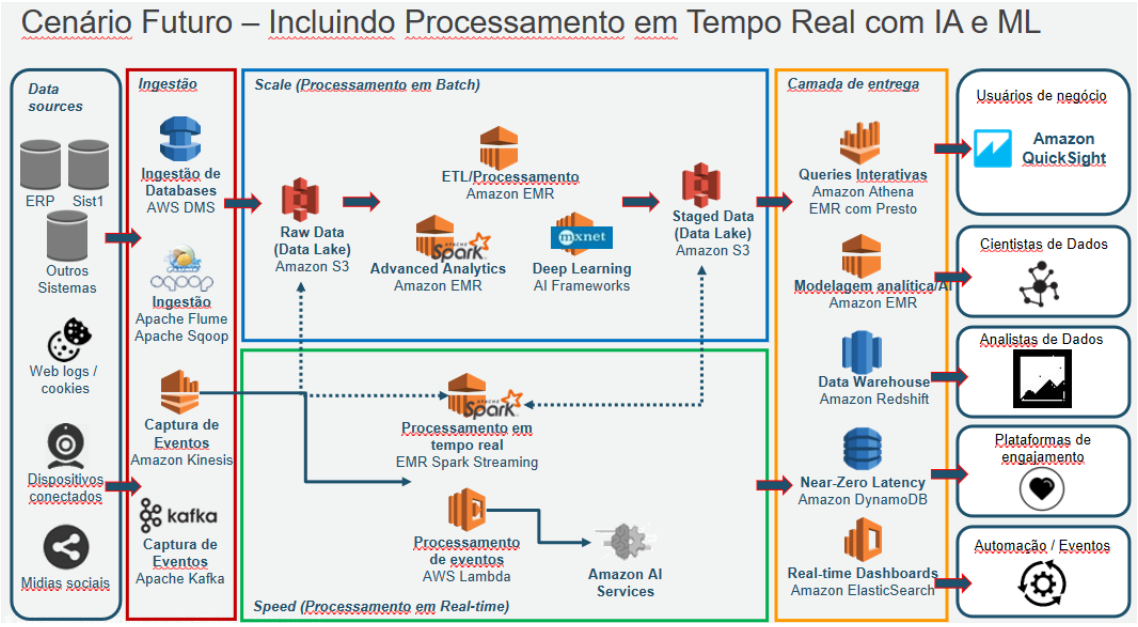
3 CONSIDERAÇÕES FINAIS

O valor agregado é visível rapidamente após a entrega e início da operação da solução implementada, com a agilidade no processo de criação dos relatórios gerenciais, para tomada de decisões, ganho de performance, avanço tecnológico, mais segurança dos dados ingeridos, e a satisfação dos usuários é clara com feedbacks positivos e ideias de melhorias e outras oportunidades já estão em discussões.

3.1. Cenários futuros

Modelo futuro corporativo de ferramentas para criacão de ingestão de dados em tempo real coletando de diversas fontes geradoras de dados alimentando o Big Data, como dados de notas de manutençcão alimentados a partir do ERP da companhia, todos os dados de variaveis de processo que controlam a planta industrial, retornando com modelos matemáticos, analiticos avançados, aprendizados de máquinas e inteligencia artificial, como todas as ferramentadas que compoem o BigData mencionadas na **figura 5**.

Figura 5, arquitetura referência corporativa, ferramentas Inteligência Artificial, Machine Learnig, Deep Learnig, Analitycs, mensageria e fontes de dados.



Autoria própria.

REFERÊNCIAS

Amazon Redshift. Disponível em https://docs.aws.amazon.com/redshift/index.html
Acesso em 24 de julho de 2019.

Amazon ec2. Disponível em https://docs.aws.amazon.com/ec2/index.html
Acesso em 24 de julho de 2019.

Amazon Glue. Disponível em https://docs.aws.amazon.com/glue/index.html
Acesso em 24 de julho de 2019.

Amazon S3. Disponível em https://docs.aws.amazon.com/s3/index.html
Acesso em 24 de julho de 2019.

Amazon Kinesis. Disponível em https://docs.aws.amazon.com/kinesis/index.html
Acesso em 24 de julho de 2019.

Amazon Athena. Disponível em https://docs.aws.amazon.com/athena/index.html.
Acesso em 24 de julho de 2019.

Planttriage. Disponível em https://www.metso.com/br/servicos/valves-services-/control-performance/planttriage/
Acesso em 28 de novembro de 2019.

Ágil. Disponível em https://digital.fispaltecnologia.com.br/gest-o/metodologia-agile-o-que-ela-pode-ajuda-na-sua-companhia
Acesso em 01 de dezembro de 2019.