UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
DIRETORIA DE PESQUISA E PÓS-GRADUAÇÃO
CURSO DE ESPECIALIZAÇÃO EM INDÚSTRIA 4.O

RAQUEL SOUZA GOULART

BIG DATA COM DATA LAKE: UM CASO DA INDÚSTRIA DE PAPEL

TRABALHO DE CONCLUSÃO DE CURSO DE ESPECIALIZAÇÃO

PONTA GROSSA
2020

RAQUEL SOUZA GOULART

BIG DATA COM DATA LAKE: UM CASO DA INDÚSTRIA DE PAPEL

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

Área de Concentração: Gestão do Conhecimento e Inovação e Gestão da Produção e Manutenção.

Orientador(a): Porf. Dr. Max Santos

PONTA GROSSA

2020

TERMO DE APROVAÇÃO DE TCCE

Big Data com Data Lake: Um caso da Indústria de Papel

Raquel Souza Goulart

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. Dr, Max Santos
Prof. Orientador

_____
Prof. Dr. Rui Tadashi Yoshino
Membro titular

_____
Prof. Dr. Marcelo Vasconcelos de Carvalho
Membro titular

A FOLHA DE APROVAÇÃO ENCONTRA-SE ASSINADA NA SECRETARIA DO CURSO

RESUMO

GOULART, Raquel. Big Data com Data Lake: um caso da Indústria de papel. 2020. 17 f. Monografia (Especialização em Indústria 4.0) - Universidade Tecnológica Federal do Paraná. Ponta Grossa, 2020.

Os pilares da Indústria 4.0 apresentam uma diversidade de tecnologias que promovem a jornada de transformação digital nas empresas, onde busca-se a inovação por meio de tecnologias como: IIoT (Internet Industrial das Coisas), Integração de Sistemas, computação em nuvem e Big Data. Este trabalho apresenta uma proposta para o desenvolvimento de uma camada de dados estruturais, concentrando os diferentes dados em um Data Lake, integrando diferentes sistemas como o Sistema de Gerenciamento de Informações da Planta e combinando dados de fontes de dados de processo, manutenção e qualidade. Essas correlações permitirão à empresa estabelecer condições através das diferentes fontes de dados, com o objetivo de fazer melhores previsões no processo e restabelecer o processo com mais eficiência, aumentando ao mesmo tempo a estabilidade do processo. Com isso poderá, consequentemente, reduzir a reclassificação de produtos. A análise de previsão de manutenção aumenta a vida útil do ativo, a confiabilidade do processo e a produtividade. A padronização proposta do modelo Data Lake torna o ambiente industrial escalável com maior segurança, através da elaboração de análises preditivas de alto desempenho, relatórios gerenciais e indicadores.

Palavras-chave: Indústria 4.0, Data Lake, IIOT, Manufatura Inteligente, Big Data.

ABSTRACT

GOULART, Rachel. Big Data with Date Lake: A Case of The Paper Industry. 2020. 17 p. Monograph (Especialization in Industry 4.0) - Federal Technological University of Paraná. Ponta Grossa, 2020.

The pillars of Industria 4.0 present a diversity of technologies that promote the journey of digital transformation in the company where the company that strives to find innovation through technologies such as: IIoT (Industrial Internet of things), Systems Integration, cloud computing and Big Data. This work presents a proposal for the development of a structural data layer concentrating the different data in a data lake, integrating different systems such as Information Management System of the Plant and combining data from sources like process, maintenance and quality data. These correlations will allow us to establish conditions through the different data sources with the purpose of making better predictions in the process, and reestablishing the process more efficiently, while at the same time increasing stability. This could consequently reduce the reclassification of products. Maintenance prediction analysis extends asset life, process reliability, and productivity. The proposed standardization of the Data Lake model makes the industrial environment scalable with greater security through the elaboration of high-performance predictive analysis, management reports and indicators.

Keywords: Industry 4.0, Data Lake, IIOT, Intelligent Manufacturing, Big Data, Supply Chain.

SUMÁRIO

1 INTRODUÇÃO

Com a convergência das tecnologias emergentes da Industria 4.0 como *Big Data, Cloud e IOT*, estamos iniciando uma jornada da transformação digital nas unidades da empresa onde queremos fábricas mais inteligentes e conectadas que permitam aumentar nossa produtividade, disponibilidade, eficiência operacional e antecipar os riscos com essas tecnologias através do uso de sistemas integrados buscando a consolidação de dados em uma única plataforma de modo que estes dados possam ser analisados e correlacionados de forma estruturada ou não estruturada oriundos das diversas fontes disponíveis na indústria como: PIMS - *Plant Information Management System*, LIMS - *Laboratory Information Management System*, ISRA – *Images Applications System*, SAP - *Data Processing Systems, Applications and Products*. Possibilitando trabalhar de forma preditiva nas análises de processos e equipamentos atuando antes que a falha aconteça.

2 REFERENCIAL TEÓRICO

2.1 BIG DATA

O Big Data permite obter vantagens em um ambiente industrial altamente competitivo. O desafio do Big Data é encontrar conjunto necessários de habilidades, ferramentas, técnicas e recursos exigidos para lidar com a complexidade advinda pela enorme quantidade de dados geradas pela indústria 4.0. Na literatura, o Big Data é representado através de 4 conceitos:

i) Volume: Consiste na quantidade de dados gerados e armazenados. O tamanho dos dados determina valor e o potencial de transforma-los em informações. Além disso, o volume de dados decide se pode ser aplicado conceito Big Data ou não de determinado conjunto de dados;

ii) Variedade: O tipo e a natureza de dados são variados, como por exemplo: imagens, textos, vídeos, áudios, etc. Consequentemente, aumenta a complexidade de analises para gerar. As vantagens de ter uma grande variedade de dados obtidos de um mesmo processo é aplicar técnicas de fusão de dados para preencher ‘peças perdidas’ entre uma análise e outra;

iii) Velocidade: Os dados inseridos no contexto de Big Datas são produzidos de forma mais continua em relação à conjunto de dados pequenos. Devido ao grande volume e variedade de dados, o processamento para que os dados sejam gerados e processados para atender a demandas de uma indústria deve ser veloz;

iv) Veracidade: É a definição estendida para Big Data, que se refere à qualidade, valor e utilidade dos dados obtidos (HILBERT, 2015).

Os dados coletados e analisados com técnicas de mineração permitem realizar atividades de rotina com o objetivo de inspecionar e testar a presença de condições de avisos que indicam que um componente está prestes a falhar. Com isto, é possível programar uma manutenção correlativa para substituir, reparar ou revisar o componente antes da falha prevista, e alguns casos eminente, minimizando riscos de atraso a produção **(**LEE; KAO; YANG, 2014).

O desafio que as organizações enfrentam é desenvolver mecanismos de governança, políticas e estruturas que atinjam um equilíbrio entre criação de valor e exposição a riscos diante de quantidades crescentes de dados e inovação que oferecem tecnologias de armazenamento com propostas cada vez melhor, mais rápida e mais barata. Um estudo recente do Centro de Sistemas de Dados em Grande Escala da Universidade da Califórnia, em San Diego, relatou que a quantidade de dados nos datacenters corporativos continua a crescer, em média, 40% ao ano. Em alguns setores principalmente os de saúde, produtos farmacêuticos, energia, telecomunicações e transporte os gerentes relatam um crescimento de datacenter superior a 100% ao ano (TALLON,2013).

2.2 DATA LAKE

A definição de *Data Lake* consiste em um repositório centralizado que permite armazenar todos os dados em formato raw. Podendo ser dados estruturados de banco de dados relacional, dados semiestruturados, como por exemplo CSV, XML e JSON e dados não estruturados (e-mails, PDFs). Um *Data Lake* geralmente é armazenamento único de todos os dados corporativos, incluindo cópias brutas de dados do sistema de origem. A partir de um *Data Lake*, é possível executar diversos tipos de análises em tempo real, processamentos de big data e aprendizado de máquina com o objetivo de auxiliar na tomada de decisões. A diferença do *Data Lake* para o *Data Warehouse* é que enquanto o *Data Warehouse* armazena dados em arquivos ou pastas, o *Data Lake* utiliza uma arquitetura plana para armazenar os dados.

2.2.1 Data Lake vs Data Warehouse

As vantagens do *Data Lake* em relação ao *Data Warehouse* encontram-se no Quadro I:

Quadro 1 – Vantagens do Data Lake em relação ao Data Warehouse

	Data Warehouse	Data Lake
Dados	Estruturado; Processado.	Estruturado / Semiestruturado / Não Estruturado; Não processado (bruto).
Processamento	Esquema de dados gerado no momento da gravação.	Esquema de dados gerado no tempo de leitura.
Armazenamento	Alto custo para alto volume de dados.	Projetado para ser barato, independentemente do volume de dados.
Agilidade	Configuração fixa levemente ágil.	Muito ágil, pode ser configurado e reconfigurado conforme necessário.
Segurança	Estratégias de segurança muito maduras.	Ainda precisa melhorar a segurança dos dados e o modelo de acesso.
Usuários	Analistas de negócios.	Cientistas e analistas de dados.

Fonte: Lee, Kao e Yang (2014).

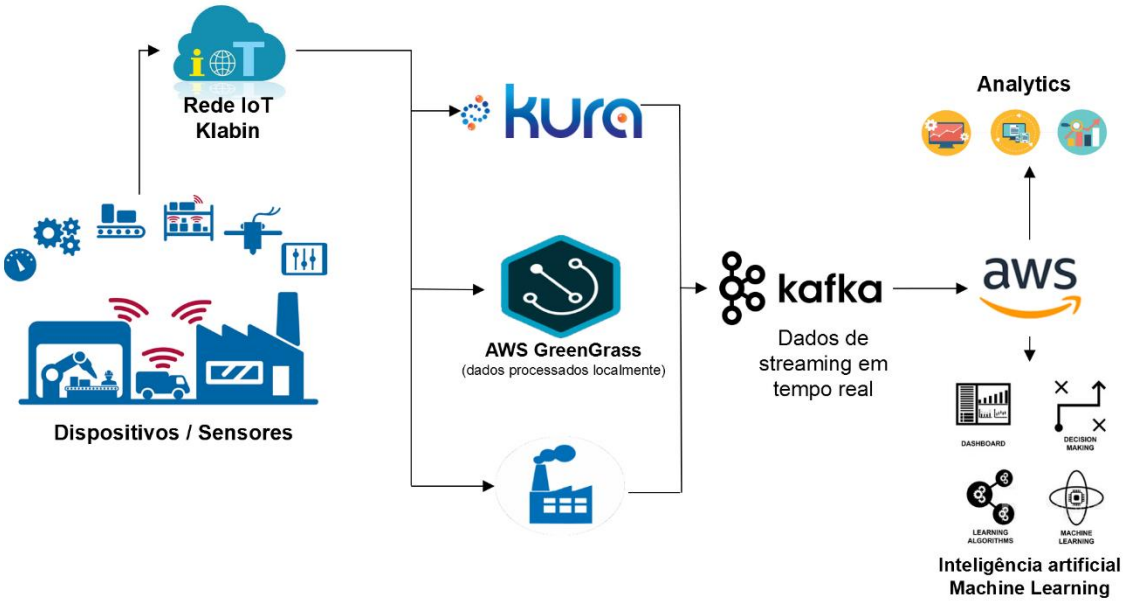
De acordo com o Quadro I, um *Data Warehouse* armazena dados modelados com um alto custo para alto volume de dados e com o esquema de dados gerados no momento da escrita utilizando o conceito *schema-on-write*. Entretanto, o *Data Lake* armazena todos os dados estruturados ou não em qualquer escala em sua forma bruta. A estrutura do *Data Lake* é definida no momento da utilização dos dados, utilizando o conceito *schema-on-read.* Outra vantagem do *Data Lake* em relação ao *Data Warehouse* é que o *Data Lake* pode ser configurado dinamicamente conforme necessidade, fazendo com seja mais ágil em relação ao *Data Warehouse*, o qual utiliza configurações fixas.

Data Lake fornecem uma plataforma completa para criar um repositório para centralizar, analisar, transformar e desenvolver aplicativos que extraem valor dos dados. Os dados estão disponíveis para revisão por todas as equipes da organização, podemos armazenar tabelas relacionais, não relacionais, documentos de texto e dados estruturados ou não estruturados definidos para interpretação automática (por exemplo, imagens ou áudio). Isso diminui o custo de transformar dados e aumenta a agilidade dos analistas para explorar e extrair novos insights de dados (MARQUESONE, 2016).

3 METODOLOGIA

Neste capítulo veremos a metodologia utilizada para o processo de criação de um *Data Lake* através de dados gerados por dispositivos e sensores da fábrica de papel até a utilização de ferramentas de análises do *Data Lake* pelo *Amazon Web Services (AWS).* A solução proposta visa coletar dados dos sensores da SKF utilizando o protocolo MQTT e enviar os dados para a nuvem aws onde serão armazenados no S3 para posterior analise e geração de Insights. A Figura 1 ilustra o processo de criação do *Data Lake.*

Figura 1: Arquitetura de Referência para Data Lake



Fonte: Autora

De acordo com a Figura 1, os dispositivos e sensores estão conectados com a Rede *IoT* da empresa. Os dados gerados serão processados localmente pelo *AWS IoT GreenGrass*.

3.1 REDE IOT

Com o objetivo de melhorar o processo produtivo, a empresa construiu uma rede dedicada para o sensoriamento *IoT,* onde encontram-se mais de 15.000 sensores e atuadores sem fio em toda sua linha de produção para o monitoramento dos equipamentos e priorização de manutenção preventiva nas máquinas críticas, as quais são responsáveis por 70% do volume de produção.

3.2 AWS GREENGRASS

A ferramenta AWS *IoT Greengrass* permite que os sensores interligados pela rede IoT da empresa processem os dados gerados localmente, sem a necessidade de estar conectado à internet, para que posteriormente quando a conexão for estabelecida seja possível gerar análises pelo AWS Amazon.

3.3 APACHE KAFKA

O Apache Kafka é um sistema de código aberto com envio distribuído de mensagens para a criação de aplicativos em tempo real usando dados de streaming. Os dados de streaming gerados pelo AWS GreenGrass serão armazenados no cluster do Apache Kafka para serem distribuídos para aplicativos de processamento de streams.

3.4 AWS AMAZON

A Solução foi desenvolvida utilizando as tecnologias descritas abaixo:

- ✓ Sensor de temperatura SKF
- ✓ Gateway com Sistema operational Ubuntu 18.10 (64-bit)
- ✓ Protocolo MQTT
- ✓ Apache Kafka
- ✓ Aws Iot

- ✓ AWS Lamba
- ✓ AWS S3

3.5 SENSORES SKF

Utilizamos neste processo o sensor SKF *DataFly*. Este sensor envia uma mensagem no formato JSON que contém a Data da mensagem, Temperatura entre outras informações tal como mostrado abaixo:

{ "Date" : 1566828585, "Pck" : 0, "QtdPck" : 21, "Acel_pk" : 0.040272, "Acel_RMS" : 0.028476, "Vel_RMS" : 0.163420, "EnvE1_pkpk" : 0.000000, "EnvE2_pkpk" : 0.000000, "EnvE3_pkpk" : 0.083112, "Temp" : 18.27800, "Voltage" : 2.832691, "Alarm_Status" : 10250, "QtdPts" : 52400, "PtsType" : "AW", "TColeta" : 2.000000, "dT" : 38.16793, "Retry_ADC" : 0, "RSSI_%" : 87}

3.4 LEITURA DOS DADOS DOS SENSORES

Por Padrão os sensores enviam os dados via wifi para a nuvem da SKF. Para viabilizar esta solução os dados foram redirecionados via regra de Firewall para um Gateway da Klabin, onde foi desenvolvido os seguintes processos:

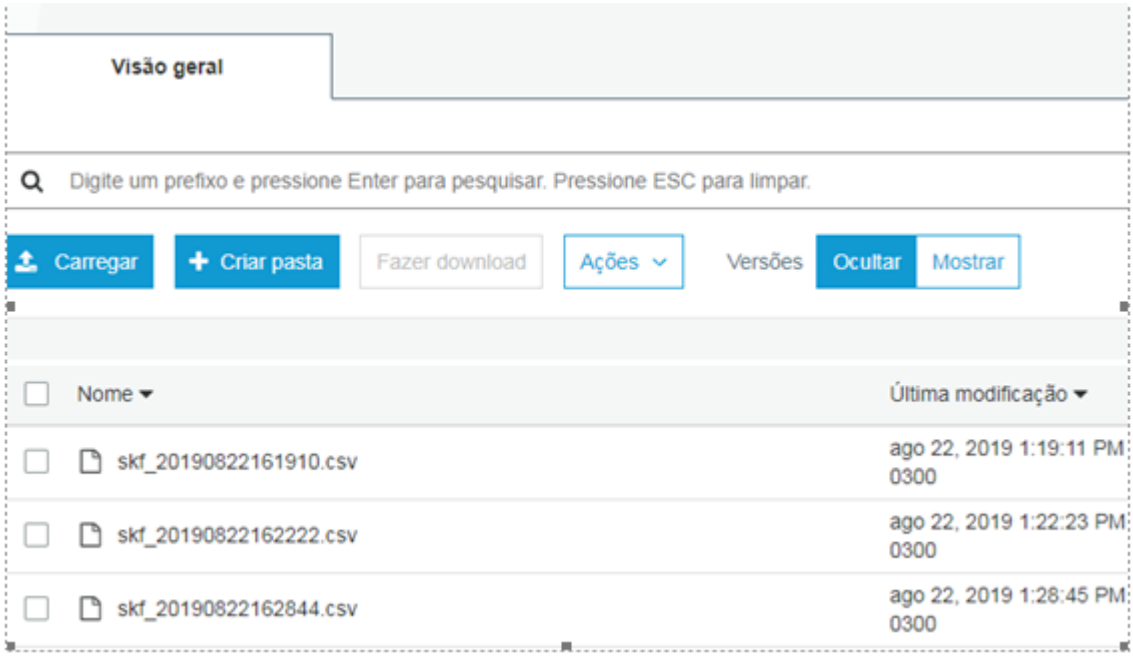
Sensor_to_kafka.py: Script responsável por receber os dados no formato Json, em uma porta especifica (20010) utilizando o protocolo MQTT, o ID do sensor é extraído do nome do tópico Mqtt e adicionado ao Json afim de identificar a qual sensor pertence a mensagem enviada. Os dados são publicados em um tópico do Kafka e redirecionado tal como recebido para a nuvem skf. O script é executado via linha de comando tal como mostrado abaixo permanecendo em execução em um looping infinito: /home/osboxes/Downloads/aws_iot/sensor_to_kafka.py

Kafka_to_awsiot.py: Script responsável ler o tópico do kafka e enviar os dados para a nuvem aws utilizando o Serviço AWS IOT core. Para utilização do serviço AWS IOT é necessário criar um device no modulo aws iot, fazer download dos certificados e do sdk do aws iot que deverão armazenados na mesma pasta do projeto no nosso caso o diretório: /home/osboxes/Downloads/aws_iot/aws-iot-device-sdk-python/.

3.5 ARMAZENAMENTO EM NUVEM AWS

Os dados são transmitidos para a AWS utilizando o serviço AWS IOT, onde foi desenvolvido uma função lambda para recebimento e armazenamento dos dados. Função Lambda responsável por receber os dados e fazer a conversão dos dados do formato .json pra o formato .csv que são armazenados em um bucket no S3 com a nomenclatura skf_[datetime-da-gravação-do-dado]. A Figura 2 ilustra o armazenamento dos dados no *Data Lake.*

Figura 2: Armazenamento dos Dados no Data Lake



Fonte: Autora

4 CONCLUSÃO

O teste de estruturação do envio de dados para um *Data Lake* foi aplicado em uma empresa fabricante de papel localizada no sul do Brasil, com quadro de, aproximadamente, 900 funcionários. Por questões de confidencialidade, a pedido da empresa, não serão fornecidos outros dados que possam caracteriza-la ou identifica-la.

Após a realização dos testes de comunicação através da arquitetura de referência criada e o processo de desenvolvimento dos s*cripts* para o envio e armazenamento em *cloud* aplicamos o teste de performance e equalização técnica avaliando requisitos de memória, disco, consumo de rede, carga, tipos de protocolos de comunicações, tempo de resposta de cada transação, tempo de resposta entre cliente e servidor, gerenciamento de sessões e alta disponibilidade.

E concluimos que utilizando o Sistema *Data Lake* contribuimos para maior eficiencia operacional, onde introduzimos maior automação, conectividade e técnicas flexíveis de produção e qualidade com o armazenamento de diferentes fontes de dados. Outros ganhos com a implementação da arquitetura de um Data Lake:

- ✓ Padronização da estrutura de um Data Lake corporativo;
- ✓ Solução escalável que pode ser aumentada ou diminuida conforme a necessidade;
- ✓ Tornar o ambiente e a estrutura escalável e mais segura;
- ✓ Alta Performance para a elaboração de análises de predição de Processos e/ou Manutenção;
- ✓ O historiador de informações em nuvem agiliza a extração e aumenta a disponibilidade dos dados, uma vez que estes se tornam independentes de servidores de automação, eliminando seus gargalos;
- ✓ Redução de custos de Hardware.

5 OPORTUNIDADES DE MELHORIAS

Após unificar todos os dados em uma única base, o próximo passo é fornecer inteligência e gerar informações que geram valor por meio de modelos de aprendizado de máquina, análise preditiva e análise de confiabilidade nas áreas de processo, manutenção e qualidade utilizando ferramentas como o PRISM, PowerBI e Spotfire. Desta forma os dados gerados pelos sensores possibilitam a análise industrial com o intuito de transformar dados em valiosos *insights*. Estes *insights* possibilitam ações mais precisas e melhores planos para manutenção preditiva.

Com estas aplicações conseguiremos mais inteligência na tomadas de decisão, pois através de uma rede IIoT de dispositivos inteligentes conectados e disponibilizados irá permitir que organizações industriais conectem as pessoas, dados e processos do chão de fábrica a todos os níveis organizacionais da empresa, auxiliando portanto na produtividade dos gestores e tomadas de decisão.

REFERÊNCIAS

HILBERT, Martin**. Big Data for Development: A Review of Promises and Challenges.** Development Policy Review, [s.l.], v. 34, n. 1, p.135-174, 13 dez. 2015. Wiley. http://dx.doi.org/10.1111/dpr.12142.

LEE, Jay; KAO, Hung-an; YANG, Shanhu. **Service Innovation and Smart Analytics for Industry 4.0 and Big Data Environment**. Procedia Cirp, [s.l.], v. 16, p.3-8, 2014. Elsevier BV. http://dx.doi.org/10.1016/j.procir.2014.02.001.

MARQUESONE, R. **Big Data - Técnicas e Tecnologias para Extração de Valor dos Dados**. Casa do Código, São Paulo, 2016.

TALLON, Paul P. **Corporate governance of big data: Perspectives on value, risk, and cost**. Computer, v. 46, n. 6, p. 32-38, 2013.