



Data Science Academy

www.datascienceacademy.com.br

Data Lake – Design, Projeto e Integração

Data Lake - Casos de Uso
Capturando e Analisando Dados de
Sensores



Escolha qualquer peça de eletrônica à sua frente - provavelmente há sensores nela. E não apenas peças eletrônicas, agora os sensores estão mesmo em árvores e edifícios. E eles estão lá por uma razão: as pessoas extraem valor de suas leituras. Então, se você acha que os sensores do equipamento de uma fábrica também são potencialmente lucrativos, você está certo. Mas o simples sentimento não é suficiente. Para extrair o valor real dos dados de sensores, antes de recorrer à consultoria de Big Data, você precisa entender os fundamentos da análise de sensores. Esta é a proposta deste estudo de caso.





O "porquê"

O “porquê” de analisar os dados de sensores depende do que a empresa precisa. Geralmente, os fabricantes analisam os dados de sensores para monitorar / otimizar processos ou projetar produtos. Vejamos o pouco sobre o “porquê”:

1. Monitoramento

O monitoramento geralmente pressupõe o seguinte curso de ação: você cria um modelo para determinar os parâmetros que definem falhas, analisa os dados do sensor, identifica padrões com falhas baseados no modelo e corrige imediatamente os erros para restaurar o fluxo normal de seus processos.

Em um exemplo específico, ficaria assim: um fabricante de produtos de plástico que possui sensores em seus equipamentos pode monitorar todas as etapas de seu processo de produção. Assim, se a temperatura do plástico derretido atingir um valor máximo admissível, eles podem baixar a temperatura do plástico e evitar a propagação no produto final. Isso permite minimizar as despesas relacionadas com produtos defeituosos e desfrutar de uma melhor garantia de qualidade.

Além disso, o monitoramento de processos é uma boa oportunidade para configurar a manutenção preventiva dentro de uma fábrica (por meio de análise preditiva). A análise de dados do sensor pode avisar sua equipe de manutenção sobre qualquer padrão defeituoso no chão de fábrica. E isso lhes dará a oportunidade de consertar o equipamento antes que ele avarie e evitar tempos de inatividade dispendiosos.

2. Otimização

A otimização de processos de manufatura significa analisar todo o ciclo de produção (ou uma certa parte dele), identificando os principais parâmetros de influência e ajustando-os para aumentar o rendimento, estabilizar a produção ou melhorar a qualidade.

Agora, vamos ver como isso funciona para um fabricante de pneus. Depois de analisar os dados de sensores, sua ferramenta de análise de Big Data revela uma peculiaridade do processo. Se um pneu contém 6% mais enxofre e a temperatura do vapor bombeado para o balão do molde do pneu for 9% menor, a qualidade da saída é 12% superior à média. A equipe de análise também descobriu que economizar nos custos de aquecimento a vapor supera as despesas adicionais relacionadas ao enxofre. Além disso, quanto menor a temperatura do vapor, mais tempo leva para os moldes dos pneus se desgastarem. E usando esse insight, o fabricante ajusta seus principais parâmetros de fabricação de acordo e consegue economizar e melhorar a qualidade ao mesmo tempo.



3. Design do produto

Levando em consideração os resultados da análise de sensores, os fabricantes podem projetar produtos melhores, permitindo as características de desempenho dos produtos em uso.

Vamos considerar os fabricantes de smartphones. Além de outros insights de design, para criar novos modelos, eles usam dados de sensores implantados em protótipos testados e smartphones já em uso. Os dados do protótipo mostram se o fabricante está projetando algo tecnicamente bom. E os sensores em uso mostram quais intervalos de modelos anteriores devem ser corrigidos e quais recursos técnicos não muito populares podem removidos.



O "quando"

Quando você analisa os dados de sensores, também deve se basear nas tarefas específicas que você precisa realizar. Talvez você pense que será suficiente apenas uma vez, como no caso de uma otimização de processo única. Mas tendo provado a doce torta de análise de Big Data junto com o lucro que ela traz, você provavelmente desejará mais. E há dois modelos que descrevem quando os dados de sensores são analisados: ad-hoc e em tempo real.

1. Análise ad-hoc de sensores

A análise de dados ad-hoc significa analisar os dados de sensores sob demanda, somente quando for necessário. Normalmente, é realizado por uma equipe de Cientistas de Dados ou Analistas.

Aqui está um exemplo. Um fabricante de produtos químicos não tem certeza sobre a frequência com que precisa mudar seus filtros industriais de ar e água. Eles fazem isso a cada 3 meses, como as instruções dizem. Mas eles não sabem se os cálculos do fabricante permitem as especificidades de sua produção química, o que pressupõe a emissão de gases tóxicos.

Analisando os dados de equipamentos e sensores de filtro, é possível que a fábrica de produtos químicos modifique todos os meses. Além disso, eles devem escolher um tipo de filtro diferente, pois o que eles usam não é bom o suficiente para lidar com o lixo tóxico. Isso permite que a fábrica evite um enorme escândalo ecológico e uma multa governamental.

2. Análise de sensores em (near) real-time

Em contraste com ad-hoc, o mero nome da análise em tempo real dá o ar da rapidez. E faz isso sem intenção de enganar: a análise de sensores em tempo real fornece um fluxo constante de resultados de análise.

No entanto, empresas diferentes entendem a palavra "tempo real" de maneira diferente. Para alguns, pode significar um intervalo de 40 milissegundos de coleta e análise de dados. Enquanto para outros, um intervalo de 30 minutos seria bastante em tempo real também. E como muitos problemas podem ter a análise de dados de sensores, escolher o intervalo certo para a análise em tempo real é definitivamente um deles.

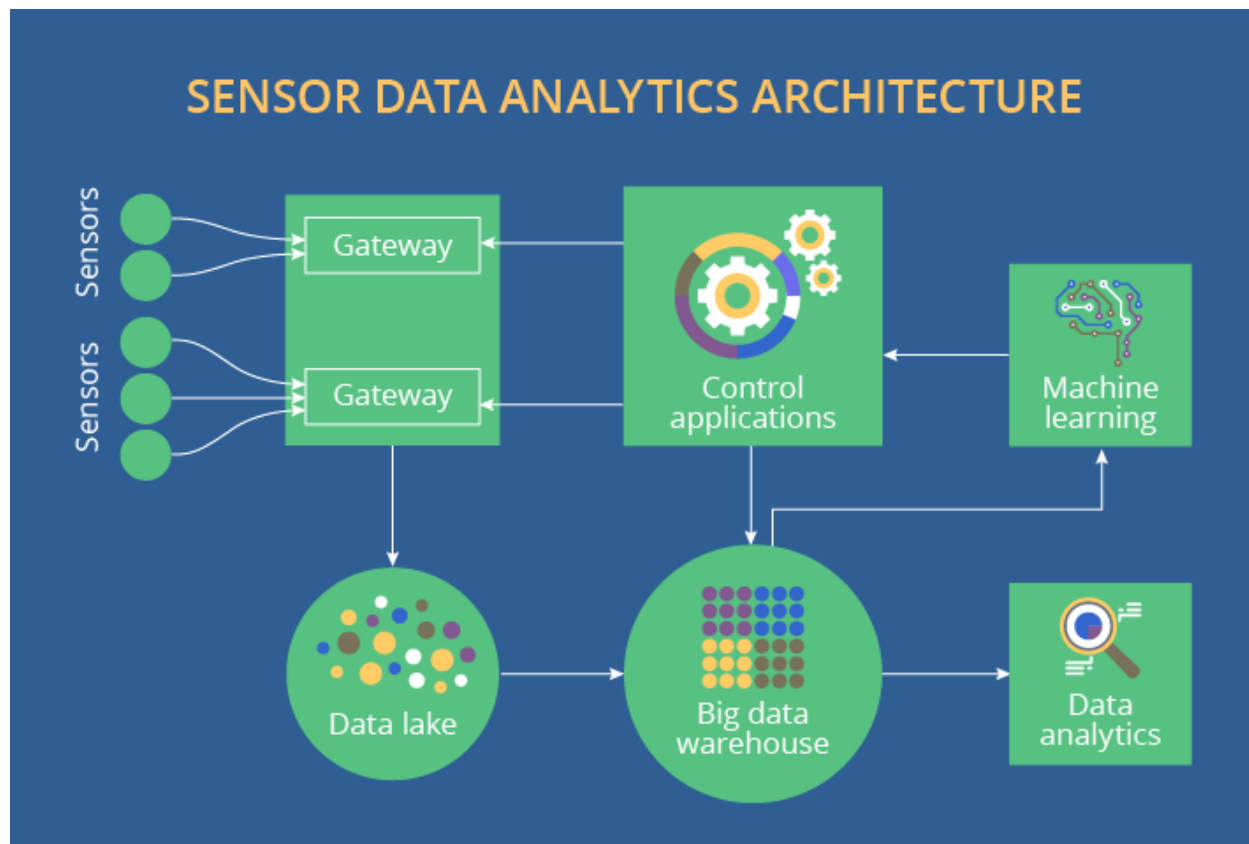
Um exemplo perfeito do intervalo de 40 milissegundos seria a análise em tempo real de uma turbina eólica. Enquanto coleta e analisa dados a uma taxa de 40 milissegundos, a ferramenta analítica usa, por exemplo, para encontrar a melhor maneira de ajustar o tom das lâminas. E é fácil justificar essa alta velocidade de coleta de dados: depende da própria natureza da capacidade de alteração do vento.



Enquanto para o intervalo de 30 minutos, o exemplo poderia ser um parque solar. Definitivamente seria muito para transmitir dados de conversão de energia solar a cada 40 milissegundos. Como o ângulo dos raios solares muda com o tempo, o painel solar pode ajustar sua posição ao Sol para converter mais energia. Levando em conta esse fato, um intervalo de coleta e análise de dados de 30 minutos poderia ser útil.

O 'como'

Como você analisa os dados de sensores também, adivinhe, depende de suas necessidades, tarefas e contexto específicos. Mas existem algumas boas práticas. A arquitetura abaixo funciona para praticamente qualquer solução de análise de dados de sensores.



O ponto de partida é um sensor. Quando coleta dados de seu "host" e os transmite para gateways, os dados são filtrados e movidos para um Data Lake.

O Data Lake é um reservatório que armazena dados em seu formato natural até que você precise analisá-los. Em seguida, os dados são extraídos, transformados ("limpos e organizados") e carregados em um Big Data Warehouse.

O Big Data Warehouse é o local que armazena os dados limpos e organizados que são usados para análise. Além de um Data Lake, um Big Data Warehouse obtém dados dos aplicativos de controle que controlam os atuadores. Ele também armazena dados sobre suas configurações de máquinas, os locais onde determinados sensores são implantados e todas as outras informações que colocam os dados do sensor no contexto. Dessa forma, o Warehouse



está "ciente" não apenas do que os sensores estão transmitindo, mas também de onde eles estão e do que seu sistema disse aos seus atuadores para fazer.

E obviamente há um segmento de análise de dados. É aí que a própria análise acontece. É a verdadeira fonte de todos os valiosos insights de negócios que você pode obter.

A última e mais "inteligente" forma de análise de dados de sensores é o aprendizado de máquina (Machine Learning). Ele observa os dados do seu sensor, percebe novos padrões, cria novos modelos para aplicativos de controle e os coloca em ação. Dessa forma, sua análise de sensores é sempre atualizada.



Um pouco mais sobre Data Lake vs. Big Data Warehouse

A principal diferença entre um Data Lake e um Big Data Warehouse é a abordagem para armazenar dados:

- Com um Big Data Warehouse, tudo é difícil: antes de carregá-lo, você precisa filtrar, processar, integrar e modelar. Você precisa dar aos dados uma forma e estrutura adequadas de "aparência". Essa abordagem é chamada de esquema na gravação (schema on write).
- Com um Data Lake, tudo fica mais fácil: você não precisa de muito barulho preparando e testando dados. Você só pega e carrega no lago. Essa abordagem é chamada de esquema na leitura (schema on read).

Agora, você pode ver porque o armazenamento de todos os seus dados em um Big Data Warehouse é caro: imagine quantos dados de sensores você precisará refinar e quantos recursos serão necessários.

As diferenças entre um Data Lake e um Big Data Warehouse estão todas nesta tabela abaixo:

	 DATA LAKE	 BIG DATA WAREHOUSE
APPROACH	Schema-on-read	Schema-on-write
STORAGE	Cheap	Costly
SECURITY	Underdeveloped	Mature
USERS	Data specialists only	Business people via analytical applications; data analysts
ARCHITECTURE	Flexible	Strict and rigid
TECHNOLOGIES	Cloud: AWS, Azure Storage: HDFS, Cassandra, HBase, DynamoDB, MongoDB Streaming: Apache Kafka, Apache Storm Processing: Spark, Hadoop MapReduce * In a big data warehouse , storage can also be organized using Redis, RedShift, Impala, Apache Kudu, etc.	
DATA STATE	All kinds: raw, unstructured, semi-structured, structured	Only structured

Um pouco mais sobre aprendizado de máquina

Veja como funciona. Um Cientista de Dados, junto com um tecnólogo de engenharia, digamos, em uma fábrica de motores aeronáuticos, seleciona um conjunto de parâmetros influentes do processo. Em seguida, o algoritmo ML passa por um conjunto enorme de dados de sensores para esses parâmetros e cria modelos. Em termos de motores de aeronaves, o resultado poderia significar algo assim: se a pressão é 18% menor que a média e a liga contém 7% a mais de alumínio, então com 78% de probabilidade isso leva a um aumento na qualidade do produto final. Depois disso, a equipe testa manualmente o modelo e, se o modelo merece, ele é aplicado pelos aplicativos de controle.

O principal benefício dos modelos de aprendizado de máquina é a precisão. Esta é a principal razão para a rivalidade entre especialistas e ML. Se eles disserem que a temperatura deve ser de 15 ° C, um algoritmo ML pode achar que deveria ser de fato 15,4 ° C. E uma



ferramenta de análise de Big Data dirá a você que a diferença de 0,4 ° C fará uma diferença financeira substancial.

As tecnologias utilizadas para o aprendizado de máquina são a Apache Spark MLlib, o Aprendizado de Máquina da Amazon, Microsoft Azure Machine Learning, o TensorFlow, Torch, etc.

Coletar e analisar dados de sensores não difere muito da coleta e análise de qualquer outro tipo de dado, sendo o volume e a velocidade de geração os pontos que merecem mais atenção.

Para expandir seu trabalho de pesquisa neste tema, acesse os links abaixo usados como referência para este estudo de caso:

Sensors, Environment and Internet of Things (IoT)

<https://www.talend.com/blog/2016/12/08/sensors-environment-and-internet-of-things-iot/>

Data Lakes: The biggest big data challenges

Why data lakes are an important piece of the overall big data strategy

<http://analytics-magazine.org/data-lakes-biggest-big-data-challenges/>

A Case Study of Sensor Data Collection and Analysis in Smart City: Provenance in Smart Food Supply Chain

<http://journals.sagepub.com/doi/full/10.1155/2013/382132>

A Data Lake Architecture with Hadoop and Open Source Search Engines

Using Enterprise Data Lakes for Modern Analytics and Business Intelligence

<https://www.searchtechnologies.com/blog/search-data-lake-with-big-data>

Data lakes swim with golden information for analytics

<https://searchenterpriseai.techtarget.com/feature/Data-lakes-swim-with-golden-information-for-analytics>

Collecting, Analyzing, and Using Fine-Grain Sensor Data with Mobile Platforms

<http://www.cs.cmu.edu/~mmv/papers/WangThesis.pdf>

Analyze HVAC Machine and Sensor Data

<https://hortonworks.com/hadoop-tutorial/how-to-analyze-machine-and-sensor-data/>

Sensor Data Analytics Powered by Deep Learning

<https://www.tcs.com/sensor-data-analytics-powered-by-deep-learning>

Techniques for High-Volume Sensor Data Analysis



<https://software.intel.com/en-us/articles/analysis-of-high-volume-sensor-data>

Sensors, IOT data and Spark

<http://sparklinedata.com/2017/04/28/sensors-iot-data-and-spark/>

Driving Real-Time Insight

The Convergence of Big Data and the Internet of Things

<http://www.oracle.com/us/solutions/internetofthings/big-data-and-iot-wp-3098381.pdf>