Braucht Industrie 4.0 ein Data Warehouse oder ein Lakehouse?

Abb.1: Warehousing im Umfeld Industrie 4.0

Bekannt ist, dass sich mit IIoT (Industrial Internet oft Things) eine Produktion in vieler Hinsicht überwachen und damit auch optimieren lässt. Durch Monitoring werden Ausfälle an den Maschinen vermieden oder sogar Ausfälle genügend früh vorhergesagt und damit ungeplante Stillstände in ein geplantes Wartungsfenster überführt. Das nennt man dann «vorausschauende Wartung» oder englisch «Predictive Maintenance». Auch lässt sich durch Energie-Monitoring herausfinden, wo die Hebel am grössten sind, um im Produktionsprozess den Energieverbrauch bzw. CO_2-Ausstoss schrittweise zu vermindern. Letzteres ist in Kombination mit Rückverfolgbarkeit ein Thema, welches besonders im letzten Jahr von industriellen Unternehmen aufgenommen wurde. Solche Anwendungen von IIoT sind aber nur die eine Seite der Medaille.

Ein noch viel grösserer Hebel der Optimierung der Produktion ist die Reduktion der Stillstandszeiten durch verbesserte Produktionsführung und -planung. Nochmals einen Schritt weitergedacht erweitert sich die Shopfloor-Planung auf eine übergeordnete Planung in Verbindung mit der Intra- und Extralogistik. Das sind alles datengetriebene Prozesse, die heutzutage in unterschiedliche, granulare Planungseinheiten auseinanderdividiert werden. Mit Künstlicher Intelligenz (KI) besteht das Potential eine Gesamtplanungen von kurzfristig (Tages bis Wochenplanung) bis mittelfristig (Monats- bis Jahresplanung) summarisch auszuführen.

«Realtime» Daten und immer ausgeklügeltere Datenanalytik & Datenoptimierung ermöglichen eine Prozessoptimierung, die umso bessere Ergebnisse liefert, je besser Daten entlang des gesamten Prozesses zentral verfügbar gemacht werden. Dieses Thema beleuchten wir in diesem Artikel detaillierter.

1. Warum geht es nicht mit ERP und MES allein?

Mit der Ausreizung der Automation (also gegen Ende von Industrie 3.0), bestand um die Jahrtausendwende die Meinung, eine nahezu maximale Effizienz im Produktionsprozess erreicht zu haben. Retrospektiv ist dies eine Fehleinschätzung, nicht nur im Bereich der Werkstattproduktion, sondern genauso in der Linien- bzw. Massenproduktion. Mit dem Aufkommen von IIoT vor 20 Jahren wurde die prinzipielle Grundlage geschaffen, den gesamten Liefer- und Produktionsprozess online zu überwachen und damit die **OEE** (Overall Equipment Effektivness) nicht nur in der Produktion, sondern über den gesamten Prozess vom Lieferanten bis zum Kunden als **OPE** (Overall Process Effectiveness) in Richtung Echtzeit zu optimieren. Je mehr sich der Markt und die Möglichkeiten in Richtung «Chargengrösse Eins» bzw. personalisierte Produkte hinbewegen, umso wichtiger aber auch gleichzeitig komplexer wird die Optimierung der OPE.

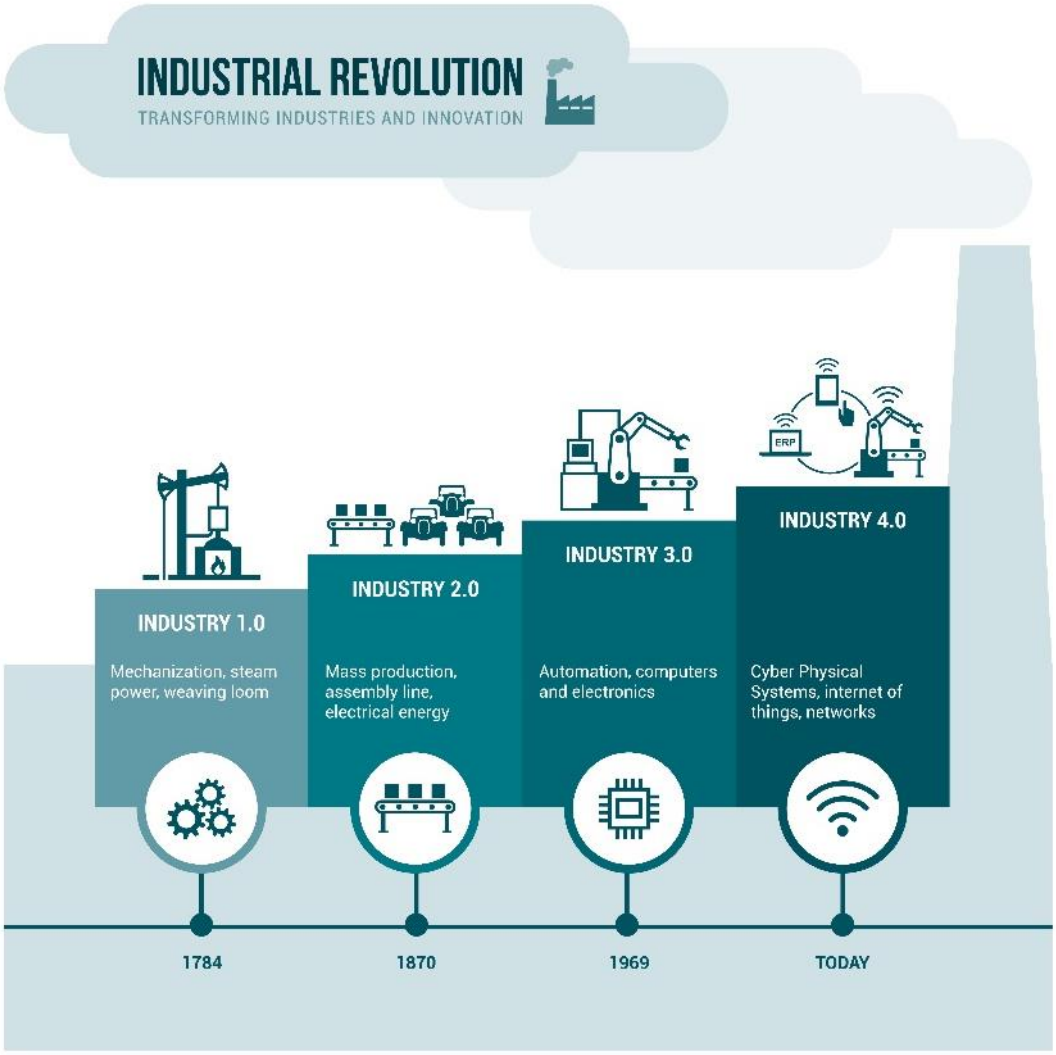

Abb. 2: Das viel gesehene Bild der industriellen Revolutionen neu erklärt

Industrie 4.0 nach Abb. 2 bedeutet auf einen zentralen Nenner gebracht «**Effizienzsteigerung der Produktion und Logistik auf der Grundlage von Daten**». Gegenüber IIoT ist Industrie 4.0 ein erweiterter Begriff aus dem deutschsprachigen Raum, der die Daten-Intelligenz verstärkt auf eine dezentrale Ebene bringt (Cyberphysikalische Systeme mit Einsatz von Künstlicher Intelligenz).

Doch sogar 12 Jahre nach Prägung des Begriffs Industrie 4.0 stehen die meisten industriellen Unternehmen, auch in der Schweiz, in der Umsetzung immer noch am Anfang. Oder positiv ausgedrückt, es schlummert weiterhin ein riesiges Potential in der noch möglichen Effizienzsteigerung industrieller Gesamtprozesse auf der Grundlage von IIoT und Künstlicher Intelligenz.

Es besteht Grund zu fragen: Wenn das Potential der Effizienzsteigerung so hoch sein soll, warum geht es denn nicht schneller? Das liegt aus unserer langjährigen Erfahrung daran, dass die industriellen Logistik- und Produktionsprozesse einen hohen individuellen Charakter für jedes Unternehmen aufweisen. Käufliche ERP und MES-Systeme sind Standardprodukte und leider nicht in der Lage, die Cyberebene von Industrie 4.0, also die Datenverarbeitungsebene, individualisiert abzubilden. ERP und MES für sich enthalten viele wichtige Daten, aber ist eine

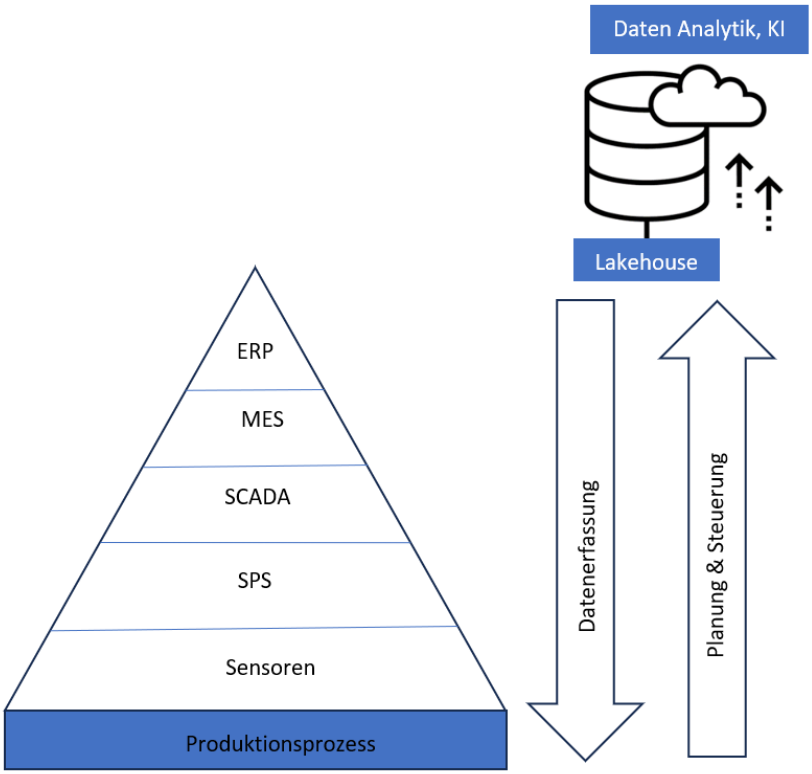

Abb. 3: Automatisierungspyramide wird erweitert durch Lakehouse

reine Zusammenführung dieser Daten nicht ausreichend, um die Anforderungen von Industrie 4.0 zu erfüllen. Auch hat sich die Hoffnung, dass Software-Produkte neuer Startups dies als

Plug & Play anbieten, bis jetzt nicht erfüllt. Jede Produktion ist einzigartig, jede dazugehörende Cybersicht auch. Deswegen bleibt eine solche Plug & Play Lösung eine weite Zukunftsvision.

Die herkömmliche Automatisierungspyramide nach Abb.3 hat weiterhin seine Daseinsberechtigung. Aber die Datenerfassung und Datenaufbereitung bis hin zur Datenanalytik sowie die Planung und Steuerung geschehen ausserhalb dieser Pyramide auf spezifischen Datenplattformen, über die im nächsten Kapitel gesprochen wird.

2. Vom Data Warehouse zum Data Lake und schliesslich zum Lakehouse

Viele Produktionsfirmen erkennen die Notwendigkeit von zentralen Datenplattformen und planen diese für die Zukunft ein. Früher sprach man vom Data Warehouse und verwendete nur strukturierte Daten. Unter Einbezug der «Realtime» Aspekte und unstrukturierter Daten wie Bilder, Audio und Texte wird heutzutage von Lakehouse gesprochen. Für ein besseres Verständnis möchte ich kurz einen geschichtlichen Abriss der unterschiedlichen Technologie von Datenplattformen geben.

Mit dem Ausbau der IT in den 90-iger Jahren erkannte man, dass auf der Grundlage von Daten bessere unternehmerische Entscheidungen möglich waren. Das waren die Anfänge von BI (Business Intelligence) und Data Warehousing. Wir selbst haben in den 90iger Jahren grosse und kleine Data Warehouses umgesetzt, auf unterschiedlichster Software und auch bereits prozessorientiert auf der Basis von Zeitstempeln. Da damals noch die Technologien für eine unstrukturierte Datenverarbeitung fehlten, wurde BI in erster Linie auf strukturieren Daten und relationalen Datenbanken aufgesetzt. Aus Gründen der Performance hat man Star- und Snowflake-Data Marts nach Kimball gebaut, häufig auch im 3-Schichten Prinzip, was aber auch zu hohen Projekt- und Betriebsaufwänden führte. Mit Grund haben KMUs die Finger von solchen Umsetzungen gelassen.

Anwendungsfälle wie Qualitätsüberwachungen auf Basis von Bildern, Rückverfolgbarkeitsbetrachtungen auf stark verteilten Daten und Verarbeitung grosser Datenmengen wie beispielsweise von Akustiksensoren lassen sich heutzutage effizient aufsetzen. No-SQL und Zeitdatenbanken sowie Plattform-Lösungen mit einem hohen Automatisierungsgrad lassen jeden erdenklichen individuellen Use Case betriebswirtschaftlich sinnvoll nutzen.

Die Firma Snowflake war ab 2012 einer der ersten, die die Automatisierung von DWHs vorangetrieben hat und durch die Reduktion von Projekt- und Betriebskosten grossen Erfolg vorweisen konnte. Business Intelligence und Data Warehousing wurde damit auch für KMUs erreichbar.

Gleichzeitig wurden aber auch Technologien geschaffen, um mit unstrukturierten Daten umgehen zu können, z.B. NoSQL Datenbanken und unterschiedlichste Tools und Formate wie bspw. Apache Spark, Deltalake und Iceberg. Mit dem Data Lake Ansatz erschlug man zwei Fliegen gleichzeitig: die Kombination strukturierter und unstrukturierter Daten sowie die

gleichzeitig stark vereinfachten Datenmodelle eines Data Lake. In einer solche Umgebung wurde zum Beispiel Databricks sehr erfolgreich.

Die Nachteile der Data Lake Lösungen erkannte man bald, denn solche Datenstrukturen sind weit weg von den Self-Service Anforderungen der Geschäftsbereiche. Die Geschäftsbereiche

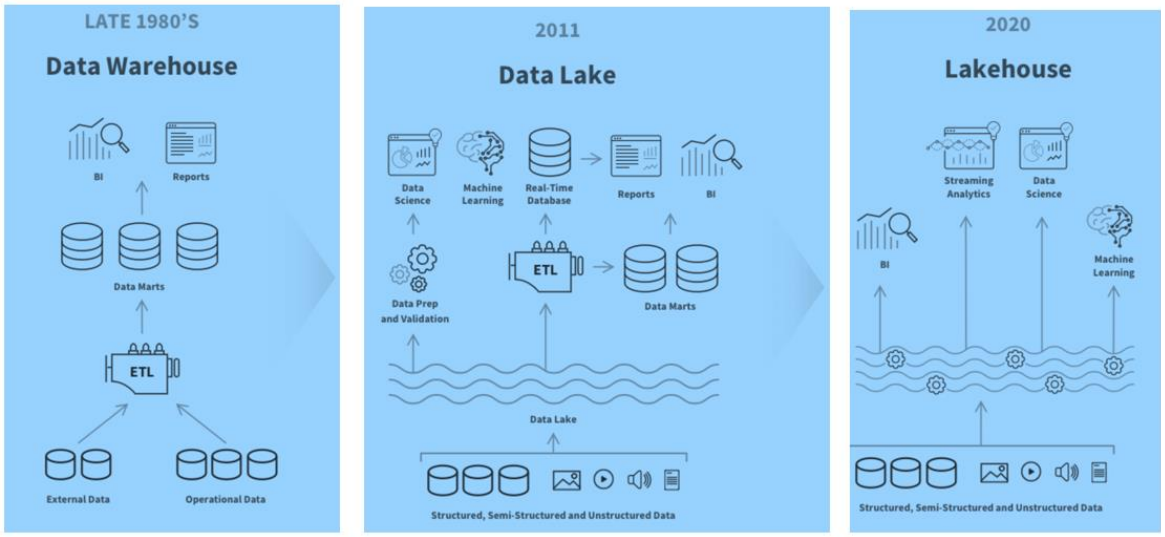

Abb.4: Vom Data Warehouse über Data Lakes zu Lakehouse

wurden ungewollt abhängig von zentralen Datenkompetenzen der IT, obwohl man versuchte, diese eng an das Business anzuschliessen.

Das Streaming von Daten wurde wie oben beschrieben für Realtime Lösungen immer wichtiger und gleichzeitig sollte der Self-Service Ansatz für die Geschäftsbereiche wieder stärker ins Zentrum gestellt werden. Das Konzept des Lakehouse war geboren. Die bisher erfolgreichen Softwareanbieter wie Microsoft, Snowflake und Databricks entwickelten sich dann in Richtung Lakehouse weiter- mit teilweise unterschiedlichen Architekturen.

Um noch einen dritten Plattformanbieter im Bereich Lakehouse zu nennen: Erst vor kurzem wuchs Dremio zu einem Einhorn heran und bietet im Moment noch eine lizenzfreie Einsteigerlösung an. Dremio ist ausgerichtet auf die Datenabfrage (Data Engine mit Caching).

Eine klare Aussage zu den Gesamtkosten für ein Unternehmen beim Einsatz einer dieser vier erfolgreichen Lösungen ist nur möglich, wenn die spezifischen Anforderungen geklärt sind. Wir raten diese vorgängig toolunabhängig aufzunehmen, wobei wir beim nächsten Thema wären.

3. «Bottom up», «Top Down» oder wie hält man Projektaufwände gering?

Als langjähriger Partner von Industrie 2025 haben wir mitgeholfen die Arbeitsgruppe «Smart Data» aufzubauen, haben viele Industrie 4.0 Use Cases im Use-Case Finder der Industrie2025 beigesteuert und sind nun auch im Praxiszirkel Smart Factory involviert. Wir sind als Datenspezialist LeanBI und nun als Substring bei vielen Umsetzungen mit dabei, nicht nur im industriellen Bereich. Schaut man sich die spannenden 49 Use Cases im Use-Case Finder an, kann man zusammengefasst schliessen, dass:

- die beschrieben User Cases im Zusammenhang Industrie 4.0 recht unterschiedlich sind,
- und damit auch die Priorisierung der Unternehmung bezüglich Industrie 4.0 individuell gefasst wird,
- und die Form der Umsetzung und die IT-Architekturen auf unterschiedlicher Basis erfolgt.

was aber nicht bedeutet, dass auf einer längerfristigen Industrie 4.0-Roadmap nicht auch viele Überschneidungen über die Firmen hinweg vorliegen.

Viele Use Cases sind bei schweizerischen Produktionsfirmen erst am Entstehen. Neue Maschinen liefern typischerweise sehr viel mehr Daten und sehen auch standardisierte Schnittstellen vor. Normal ist aber auch, dass Logistik und Produktion weiterhin Maschinen umfassen, die mehrere Jahrzehnte alt sind. Diese Produktionen lassen sich durch datengetriebenen Retrofit (Retrofit 4.0) aufrüsten, was einen Bruchteil einer neuen Maschine kostet. Neue Anwendungen wie Lyra/Moonstone von unserem Partner GradeSens kommen auf den Markt und vereinfachen den Retrofit 4.0. Trotzdem müssen aber die Daten übergeordnet zusammengeführt werden.

Wir sehen es als richtigen Weg an, die Umsetzungen von Industrie 4.0 Projekte auf der Basis von einzelnen Use Cases auszuführen. Aber was ist richtig? «Bottom up», indem jeder Use-Case separat als technisches Projekt umgesetzt wird, oder ein «Top Down» Ansatz, indem zuerst eine Daten-Gesamtarchitektur aufgebaut wird, auf welcher sich dann alle Use Cases umsetzen lassen. Die Wahrheit liegt wie so oft dazwischen. Ein reiner Top Down Ansatz ist nicht zu empfehlen, da sich Technologie rund um Datenverarbeitung und KI weiterhin rasant entwickelt. Aufwand und Zeit beim Aufsetzen einer Gesamtarchitektur weist ein hohes Verlustrisiko auf. Wir empfehlen die Erstellung eines Big Pictures einer Datenarchitektur die mindestens 4 bis 5 Jahre Bestand hat. Diese Architektur sollte so flexibel gehalten sein, dass sich einzelne Bausteine ohne Komplikationen ersetzen lassen. Auf diesem Big Picture wird dann der erste Use Case umgesetzt und damit Erfahrungen gesammelt. Ein erster Use Case sollte eine kurze Durchlaufzeit von wenigen Monaten haben. Sollten für einen zweiten Use Case Adaptionen notwendig sein, dann sind diese im Allgemeinen nicht kostenintensiv. Die gemeinsame Datenarchitektur mehrerer Use Cases hat so eine gute Basis, hat aber gleichzeitig auch die Möglichkeit zu wachsen.

4. Was macht Daten-Architekturen rund um IIoT und Lakehouses erfolgreich?

Unsere Erfahrung zeigt, dass im industriellen Umfeld viele Firmen einen Best of Breed Ansatz wählen. Manchmal wird auch ein reiner Open Source Ansatz gewünscht. Es gibt auch jene Unternehmungen, die einen einzelnen Produktanbieter also eine Suite wie Microsoft, Google, Amazon, SAP, usw. auserkoren haben und so über Jahre hinweg Software grösstenteils aus einer Hand beziehen. Beide Ansätze «Best of Breed» und «Software Suite» führen schon jahrzehntelang zum Erfolg. Der Best of Breed Ansatz hat den Vorteil, dass die neuesten und besten Entwicklungen am Markt in die Architektur aufgenommen werden, andererseits ist etwas mehr Aufwand bei der Integration der einzelnen Tools vorzusehen. Heutzutage sind jedoch Integrationsprobleme sehr viel kleiner als noch vor vielen Jahren. Beim «Best of Breed» Ansatz ist die Lieferantenbeziehung einfacher, da gewöhnlich von wenigen Ansprechpartnern ausgegangen wird.

Wir setzen bei unseren Umsetzungen stark auf Open Source, trotzdem ist zu berücksichtigen, dass Open Source nicht gleichbedeutend mit «Gratis» ist. «Managed Services» machen Sinn, wobei Open Source bei ausgewachsenen Anbieterlösungen meistens ein Einsteigerangebot bietet. Über die letzten 30 Jahre habe ich auch schon häufiger erlebt, dass besonders grössere Open Source Lösungen mit der Zeit in ein Lizenzmodell wechseln. Dann kann es sogar richtig teuer werden. Also ist etwas Vorsicht bei der Auswahl von Open Source geboten.

Die hier besprochene Architektur der Abb.5 zeigen sowohl den «Best of Breed» Ansatzes als auch im Falle einer Suite den Microsoft Ansatz, der in der Schweiz häufig Anwendung findet. Auch bei Microsoft ist es nicht eine Gesamtlösung, sondern die Kombination unterschiedlichster Software-Services. Zu berücksichtigen ist auch, dass die aufgezeigten Produkte den Stand 2024 widerspiegeln. Auch bei uns als Daten-Umsetzer ist die Produktauswahl mit der Technologieentwicklung mittelfristig im Fluss. Als nächstes einige Kommentare zur Abb.5:

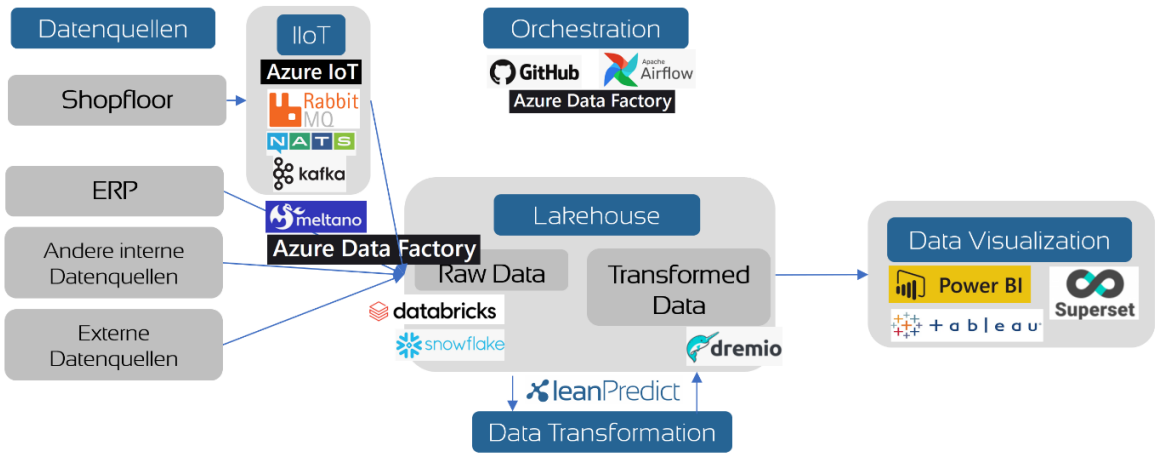

Abb.5: Daten-Architektur im industriellen Umfeld mit typischen Software-Vertretern

Datenquellen und IIoT

Die Datenquellen auf dem Shopfloor (Produktionsebene) sind die Steuerungen (SPS) und übergeordneten SCADA-Systeme. Neue Industriesysteme haben als Schnittstelle standardisiert OPC-UA, an welchem diverse Open Source Message Broker (wie bspw. Rabbit MQ) auf der Basis des MQTT-Protokolls problemlos andocken können. Auf der Basis von MQTT lassen sich Daten in Form von Datenpaketen sicher verschicken. Schnittstellen älteren Typs wie Profibus, die Weiterentwicklung Profinet sowie andere Industrieprotokolle können in der Anbindung etwas komplizierter sein. Kommerzielle Plattformen wie bspw. die Schweizer IoT-Plattformanbieter Stemys weisen Konnektoren auf, die auch hier ein Anschliessen einfach machen. Möchte man grosse Datenmengen streamen, dann lässt sich die Open Source Lösung NATS einsetzen - sehr beliebt ist auch Apache KAFKA, besonders bei grösseren Unternehmen. Azure IoT ist ein Produkt, welches dann verwendet wird, wenn eine Firma eine Microsoft-(Cloud)-Strategie verfolgt. Azure IoT kann an OPC-UA andocken, Daten streamen oder Daten als Pakete verschicken.

Wie die Automatisierungspyramide von Abb.3 zeigt, reichen die Daten vom Shopfloor nicht aus, um OEE oder OPE auszuführen. Die Anbindung von ERP oder weiterer Systeme wie Wartungs-Software oder MES kann via Meltano (open Source ELT) oder im Microsoft Umfeld via Azure Data Factory geschehen. Letzteres bietet auch eine einfache Analytics-Plattform inkl. Daten-Orchestrierung an.

Lakehouse

Im Kapitel 2 wurde schon ein kurzer Abriss über die von uns verwendeten Lakehouse Anbieter gemacht. Die heutigen Lakehouse Lösungen sind cloudbasierte Gesamtlösungen, die automatisiert skalieren. Das bedeutet, man muss sich um die unterliegende Hardware nicht kümmern und zusätzliche Ressourcen werden nach Belastung der Plattform zu- oder weggeschaltet. Die Cloud-Lösungen bieten auch Hand bei der Automatisierungen der Datenprozesse und in der Erstellung von Datapipelines sowie Datenstrukturen. Die Lösungen weisen darüber hinaus auch Schnittstellen zu Datenvisualisierungstools auf und sind mit diesen fliessend integriert. Die Auswahl der richtigen und einer nachhaltig kostengünstigen Plattformlösung geschieht idealerweise in Workshops auf der Basis zukünftiger Use Cases.

Orchestrierung

Lösungen wie Github können nicht nur die Versionierung der Codes steuern, die «Continous Integration» sicherstellen, sondern können heutzutage auch den gesamten Datenaufbereitungs-Prozess steuern. Von uns genutzte Orchestrierungslösung sind beispielsweise die Open Source-Lösung Apache Airflow oder Azure Data Factory, welches innerhalb der Microsoft-Welt integriert ist.

Data Transformation und Künstliche Intelligenz

LLMs (Large Language Models) haben die Welt im Jahr 2023 euphorisiert, die Künstliche Intelligenz ist von neuem aufgewacht. Natürlich wird es in den kommenden Jahren auf LLM

viele Anwendungen geben, im Bereich Industrie 4.0 werden LLM-Lösungen aber nicht ausreichen.

Neben den LLM bieten Software-Anbieter im Bereich Industrie 4.0 Modellumsetzungen an, die meistens eher generisch aufgesetzt sind. Unsere eigenen Machine Learning Modelle von LeanPredict wenden wir seit bald 10 Jahren an und sind in diesem Bereich projekterprobt. Unsere Modelle sind klein und flexibel und können direkt in die Umsetzung der Industrie 4.0 Anwendungen einbezogen werden. Wir sehen es als Kundenvorteil an, dass wir LeanPredict nicht als Produkt verkaufen, sondern es als Softwarecode in den Projekten für eine schnellere Umsetzung nutzen. Es entsteht damit für Kunden keine Abhängigkeit zu einem weiteren Provider, also auch nicht zur Substring selbst. Die Modelle können sowohl in der Cloud als auch vor Ort als Edge Lösung in der Produktion laufen. Wie Abb. 6 zeigt, haben wir LeanPredict nach Anwendungsfällen in der Produktion strukturiert. Damit können Unternehmen bspw. Qualitätseinbussen am Produkt erkennen oder ein Alarming auf Schäden oder auch Energieproblemen einrichten. LeanPredict ist damit ein Beschleuniger unserer Industrie 4.0 Projekte.

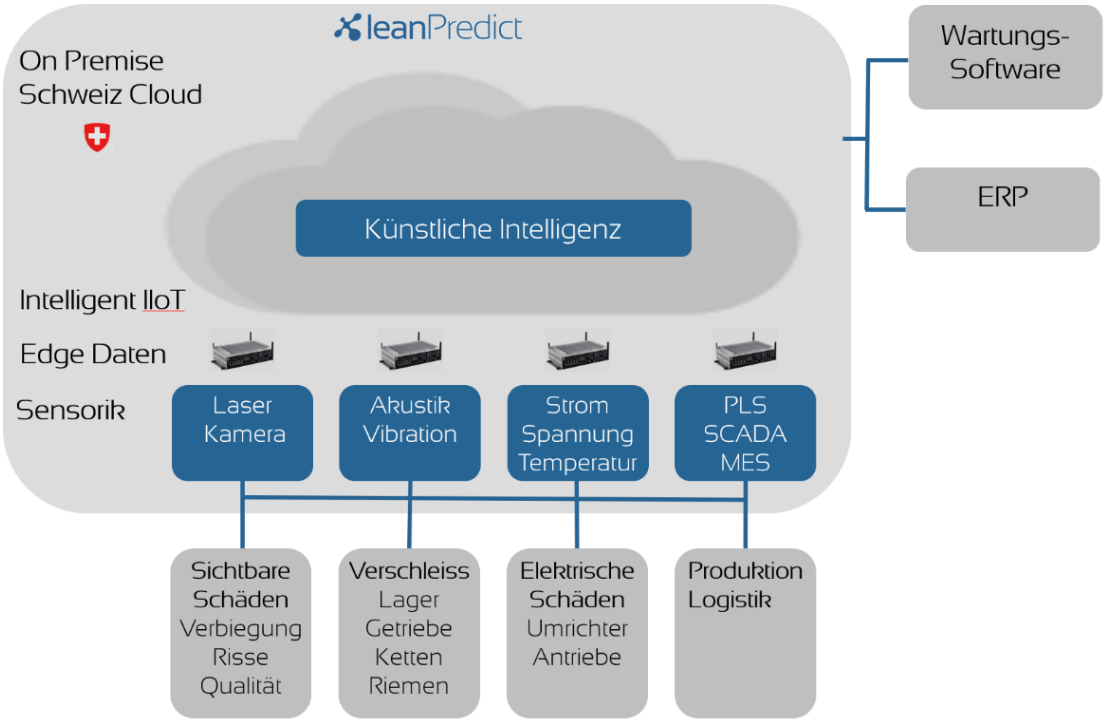

Abb.6: LeanPredict als Komponente der Daten-Architektur

Data Visualization

Die Visualisierung und die Datenanalytik sind Schlüsselfaktoren der Daten-Architektur. Damit gute Entscheidungen getroffen werden können, müssen die Daten verständlich und klar visualisiert werden. Ein von uns genutztes Visualisierungs-Tool ist das Open Source Produkt Apache Superset für einfache und schnelle Visualisierungen. Es gibt sehr viele BI Tools am

Markt, aber im KMU-Bereich wird in der Schweiz häufig Power BI von Microsoft und Tableau eingesetzt. Deshalb und weil wir in Projekten damit erfolgreich waren, haben wir bei Substring auf diese beiden BI-Tools Kompetenzen aufgebaut.

Somit beenden wir hier unsere Ausführungen zum Thema Daten-Architektur im Umfeld Industrie 4.0, ohne näher auf das wichtige Thema Security eingegangen zu sein. Sehr gerne besprechen wir auch das Thema Security oder all die obigen Themen direkt. Kommen Sie dazu gerne auf uns zu.

Dr. Marc Tesch
Substring AG

Senior Consultant und Strategic Business Developer
Dr. sc. techn., Dipl. Masch. -Ing. ETH,
Betriebswissenschafter NDS ETH
Phone: +41 79 247 99 59

m.tesch@substring.ch
https://substring.ch/