

Wie aus Daten Wissen wird

Mit der lernenden Organisation zu Big Data

Autor Wolfgang Pleus

Überblick

TIMOCOM hat in neun Monaten eine vollständig neue Big Data Plattform etabliert. Die Initiative umfasste den technischen und organisatorischen Aufbau, agiles Vorgehen und agiles Engineering, sowie Entwicklung und Betrieb der bis dahin unbekannten Technologien.



Ausgangssituation

Im Laufe der Unternehmensgeschichte hat TIMOCOM eine stark relational geprägte Softwarelösung aufgebaut. Ständig wachsende Datenbestände weckten den Wunsch nach Veränderung. IT-Trends im Bereich Künstliche Intelligenz und Maschinelles Lernen verstärkten das Streben nach Unterstützung in diesem Bereich. Zudem war mit der Zeit eine Archivierungslösung entstanden, die das Wachstum sowie die Flexibilität eingeschränkt hat. Weitere datengetriebene Use Cases wie Betrugserkennung sollten ermöglicht werden.

Diese Aspekte sollten durch eine moderne, technische Lösung abgedeckt werden.

Um diese Ziele zu erreichen, hat TIMOCOM auf die externe Beratung durch die Firma IKS GmbH in Zusammenarbeit mit PLEUS Consulting gesetzt.

Ergebnisse

- Verarbeitung großer Datenbestände
- Neue Big Data Plattform
- Massive Skalierbarkeit
- Ablösung von Altsystemen
- Agiles Vorgehen, agile Engineering Praktiken und agiles Team
- Neue Anwendungsfälle und maschinelles Lernen
- Verbessertes Reporting
- Integration in bestehende Systemlandschaft



„Ich bin total begeistert. Nun haben wir eine Lösung, die mit dem Unternehmen weiterwachsen kann.“

-Ansgar, Direktor IT



PLEUS CONSULTING

Ziele

Die neue Lösung muss in der Lage sein, sehr große Datenmengen zu speichern und zu verarbeiten. Sie soll dazu genutzt werden, aus den ständig wachsenden Datenmengen Wissen zu erlangen. Dieses Wissen dient dann der Verbesserung von Prozessen, Dienstleistungen für die Kunden, sowie der Entwicklung neuer Geschäftsmodelle. Die Lösung soll umfassend skalierbar sein, um mit den Anforderungen von TIMOCOM wachsen zu können. Das Ergebnis wird verwendet, um wichtige Use Cases wie Betrugserkennung, Archivierung oder auch Reporting zentraler KPIs zu unterstützen (siehe Abbildung 1).

Herausforderungen

TIMOCOM benötigte zu Beginn der Initiative Unterstützung und personelle Ressourcen, um den Aufbau und Betrieb der neuen Lösung zu implementieren (siehe Abbildung 4). Die angestrebte Lösung wies eine hohe Komplexität auf. Die erforderliche Expertise musste also im Rahmen des Projektes ausgebaut werden.



Abbildung 1: Typische Use Cases

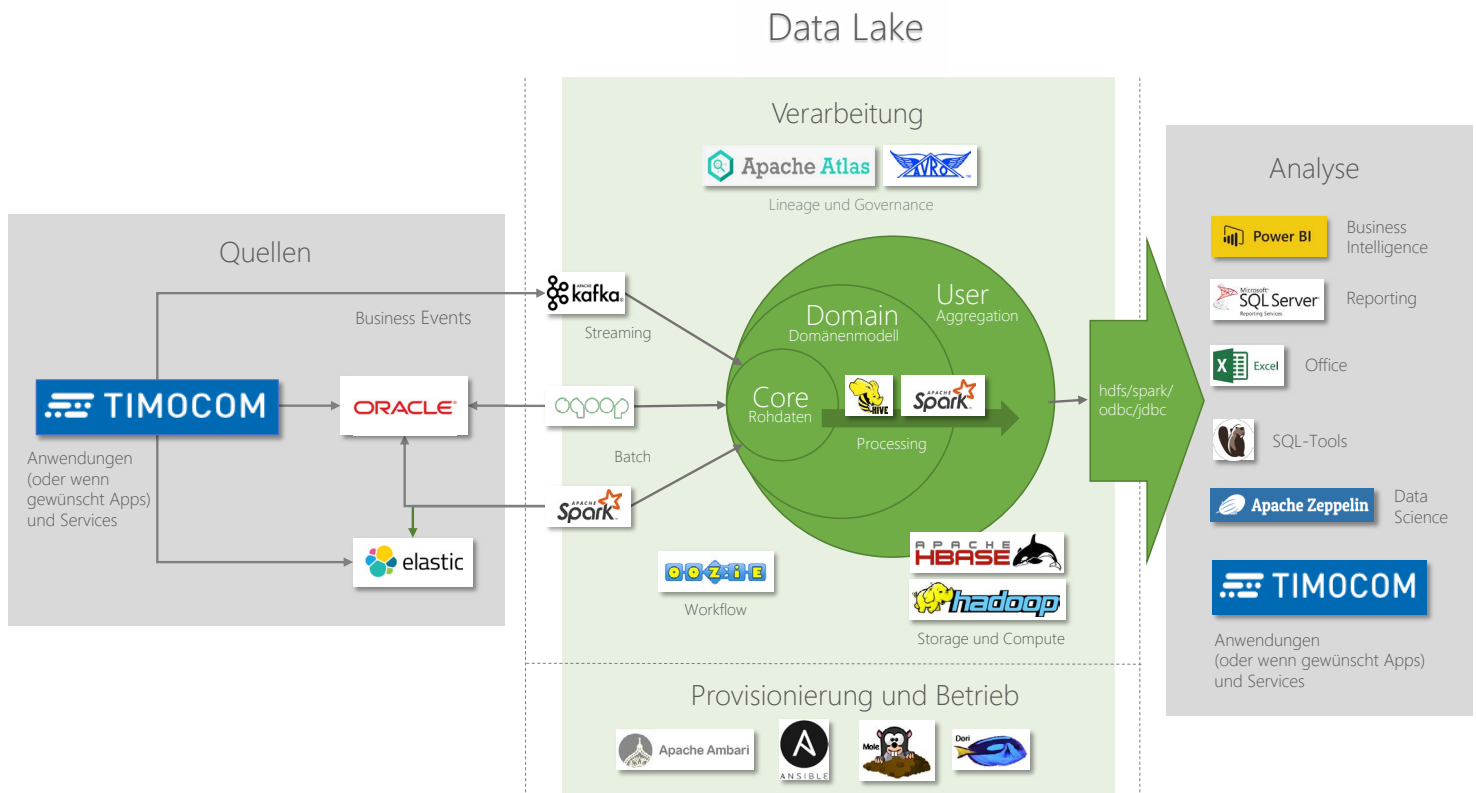


Abbildung 2: Architektur und Produktübersicht

Technische Lösung

Im Rahmen des Projektes wurde eine Data Lake Plattform geschaffen. Die Plattform besteht aus einem Cluster verbundener Maschinen, die massive Ressourcen im Bereich Storage, CPU und RAM bereitstellen. Der Data Lake ist damit in der Lage, große Datenmengen zu speichern und für Aufbereitungen und Analysen bereitzustellen. Gleichzeitig können umfassende Transformationen und Berechnungen im Data Lake durchgeführt werden. Somit werden Datenverarbeitungen möglich, die mit einzelnen Maschinen nicht umsetzbar wären.

Organisiert ist der Data Lake in logische Bereiche zur Speicherung von Rohdaten, aufbereiteter Domänendaten sowie Archivdaten. Alle Daten sind katalogisiert und vollständig durch Schemas in einem Repository beschrieben. Zusätzlich wurden automatische Prozesse implementiert, um nicht dokumentierte Daten zu detektieren und einer nachträglichen Dokumentation zu unterziehen. Dies ist essenziell um Governance zu ermöglichen und Anforderungen, wie sie beispielsweise durch die Datenschutzgrundverordnung (DSGVO) gestellt werden, zu erfüllen.

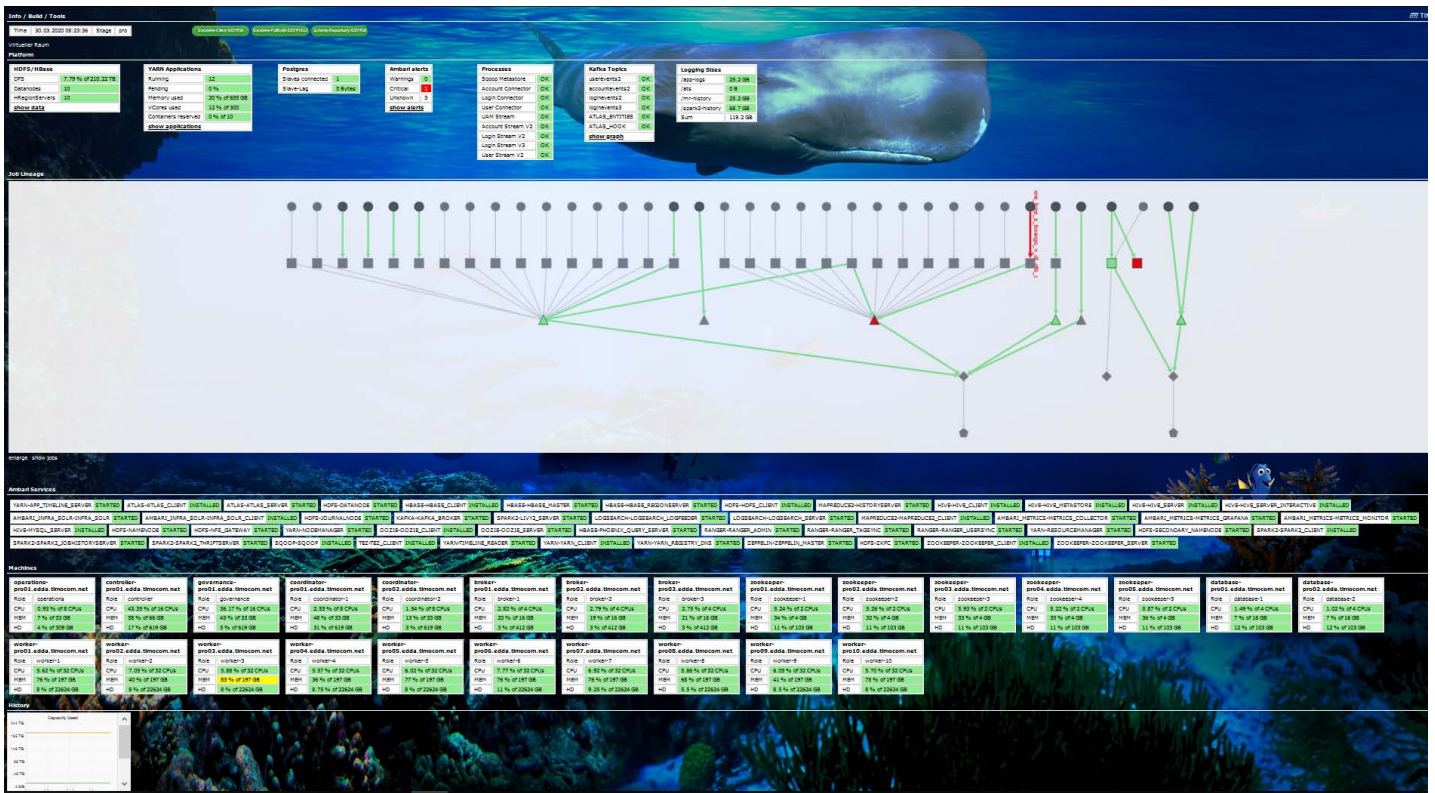


Abbildung 3: Mole Monitoring

Konnektoren versorgen den Data Lake mit Daten der Quellsysteme. Zu den Daten gehören beispielsweise Transport- und Frachtangebote sowie Interaktionsdaten der Anwender. Wichtige Quellen sind relationale Oracle Datenbanken sowie Datenströme aus den Livesystemen. Dadurch werden Echtzeitanalysen, wie beispielsweise zur Betrugserkennung, ermöglicht.

Der Data Lake basiert vollständig auf Open Source Produkten aus dem Big Data Umfeld. Dazu gehören Hadoop mit HDFS und YARN für die Verwaltung von Storage, CPU und RAM. Die strukturierte Ablage der Daten erfolgt über Hive, HBase und Avro. Für Import, Streaming und Transformation von Daten dienen SQOOP, Kafka und Spark. Zur Jobsteuerung wird Oozie eingesetzt. Daneben sorgen Produkte wie Atlas, Zookeeper, Ambari und Ansible für den ausfallsicheren Betrieb der Plattform. Apache Zeppelin ermöglicht, es Data Scientists explorative Analysen durchzuführen. Bestehende Systeme werden mittels ODBC/JDBC angebunden.

Die Technologieauswahl ermöglicht prinzipiell unbegrenzte technische und wirtschaftliche Skalierbarkeit, die den Wachstumsanforderungen der TIMOCOM entspricht.

„Die Leistungsfähigkeit der Plattform ist faszinierend. Und das ohne Lizenzkosten.“

-Markus, Entwickler Data Lake

All diese Produkte arbeiten im Verbund als integrierte Gesamtlösung und sind angebunden an die bestehende Business Intelligence Lösung. Die Visualisierung erfolgt auf der Basis von Power BI (siehe Abbildung 2).

Betrieben wird der Data Lake On Premise in den Rechenzentren der TIMOCOM, in einer ausfallsicheren Topologie. Sie erfüllt damit die hohen Anforderungen von TIMOCOM an Datensicherheit, Verfügbarkeit und hohe Kosteneffizienz.

Ergänzt wird die Lösung um ein leichtgewichtiges, selbst entwickeltes Monitoring (Mole), dass die Live-Überwachung aller betriebswichtigen Ressourcen ermöglicht (siehe Abbildung 3).

Organisatorische Lösung

TIMOCOM verfügte zu Beginn über wenig Expertise im Bereich der Big Data Technologien. Der Markt stellte ebenfalls kaum Experten bereit. Zudem sollte internes Know-how aufgebaut werden, um langfristig handlungsfähig zu sein.

Um die Plattform effizient entwickeln und betreiben zu können, wurde ein neues Team etabliert. Geeignete Teammitglieder fand man durch die Entwicklung eines funktionalen Prototypen, der bereits typische Anwendungsfälle abgebildet hat. Der Prototyp wurde innerhalb weniger Wochen von einem externen Lead Developer entwickelt und den Mitarbeitern im Rahmen einer mehrwöchigen Roadshow vorgestellt. Dadurch konnte das Verständnis für die Data Lake Plattform in der Organisation gestärkt werden. Gleichzeitig nutzte man die Gelegenheit, so Mitarbeiter zu identifizieren, die sich für diese Technologien interessieren. Im Ergebnis sind Menschen zusammengekommen mit hohem Interesse und großer Motivation für das Thema.

Die vorhandenen Fähigkeiten wurden über Skillmatrizen ermittelt (Abbildung 4). Nach der GAP-Analyse hat sich das Team im Rahmen des Projektes aktiv mit Skillentwicklung beschäftigt, immer mit starkem Bezug zur Problemstellung. Dabei wurden moderne Formen der Zusammenarbeit, Scrum und Pair Programming genutzt. Gelerntes wurde unmittelbar präsentiert und an die Gruppe weitergegeben. Der Lead Developer hat das Team begleitet und zum Wissensaufbau beigetragen.

Nach einiger Zeit war das Team in der Lage, die Plattform eigenständig zu betreiben. Dazu hat es agile Vorgehensweisen und agile Engineering Praktiken immer weiter eingeübt. In Reviews und Retrospektiven wurde die eigene Arbeit im Sinne von Inspect & Adapt kontinuierlich verbessert. Hohe Transparenz und Prognosen ermöglichen die Steuerbarkeit des Vorhabens. Von Anfang an wurde darauf geachtet, dass keine Wissensmonopole entstehen, die die Resilienz des Teams beeinträchtigen würden. Ausgeprägte

Lernkultur und iteratives Vorgehen nach Scrum sind zentralen Erfolgsfaktoren um die komplexe Technologie erfolgreich einzusetzen.

Fachlich / Technisch	MA1	MA2	MA3	MA4	MA5	MA6	MA7	MA9	Ø
HDFS	1	1	7	0	0	0	1	1	1,38
Hive	1	1	7	0	0	2	1	1	1,63
Spark	1	1	6	1	1	0	1	1	1,50
Kafka	1	1	6	0	0	0	1	1	1,25
Sqoop	1	1	7	0	0	0	1	1	1,38
Fume	0	1	3	0	0	0	1	1	0,75
Oozie	0	0	4	0	0	0	1	1	0,75
HBase	0	0	3	0	0	0	1	1	0,63
Atlas	0	0	3	0	0	0	1	1	0,63
Ranger	0	0	2	0	0	0	1	1	0,50
Oracle	7	10	6	5	5	8	2	6	6,13
SQL	7	10	8	7	8	8	2	5	6,88
Java	0	4	8	2	3	2	1	2	2,75
Maven	0	5	9	0	0	0	0	1	1,88
Power BI	0	0	3	7	4	2	6	1	2,88
Domänenwissen	6	1	1	3	3	5	7	2	3,50
PostgreSQL	5	3	9	2	3	3	0	2	3,38
Unix	0	7	7	2	2	0	1	5	3,00
Git	5	4	8	5	5	3	1	2	4,13
Docker	2	5	7	1	2	0	1	2	2,50
Scrum / Agile	5	8	9	8	7	7	8	1	6,63
Datenschutz / DSGVO	1	1	2	0	0	0	0	0	0,50
Aufgaben									
Konzeption der Plattform	5	6	8	2	1	2	0	6	3,75
Einrichtung und Betrieb der Plattform	0	6	8	0	0	0	0	7	2,63
Entwicklung von Pipelines (Data Lineage)	2	6	9	2	1	2	0	0	2,75
Visualisierungen, Data Science	4	6	7	8	6	3	4	0	4,75
Analyse fachlicher Anforderungen	8	6	9	7	2	7	8	2	6,13
Design des Domänenmodells (Governance)	8	6	8	4	2	7	4	0	4,88
Dokumentation bestehender Datenstrukturen	0	0	0	6	3	4	0	2	1,88
Entwickeln von Team und Prozess	4	6	9	6	5	6	7	0	5,38
Teampplay (Lencioni)									
Humble	7	4	7	5	6	6	7	7	6,13
Hungry	7	4	8	9	7	7	8	4	6,75
Smart	3	5	8	5	5	6	8	7	5,88
Voraussetzungen									
Motivation	6	8	10	8	5	4	8	7	7,00

„Mit einem Team, das lernen will und darf, lässt sich technisches Wissen sehr gut aufbauen.“

-Wolfgang, Lead Developer, Agile Coach

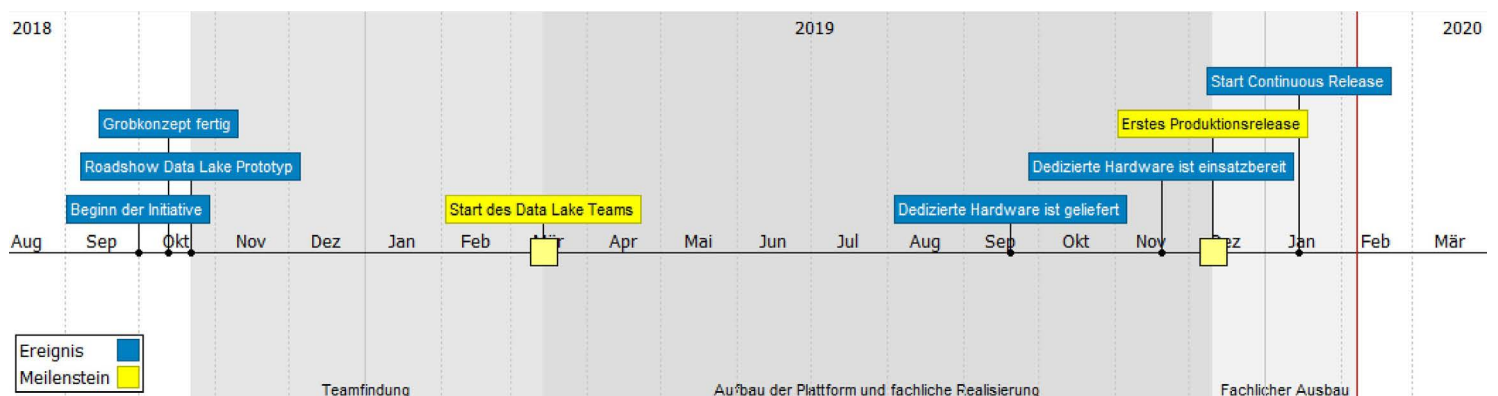
Abbildung 4: Skillmatrix

Was ist ein Data Lake?

Ein Data Lake ist eine Metapher für ein technisches System, das in der Lage ist, sehr große Datenmengen zu speichern und zu verarbeiten. Dies geschieht durch den Zusammenschluss vieler Rechner zu einem Clusterverbund. Die Datenhaltung- und -verarbeitung erfolgen verteilt in diesem Cluster, wodurch die Verarbeitung sehr großer Datenmengen möglich wird. Wie Wasser in einen See fließen Daten in unterschiedlicher Form in den Data Lake. Data Lakes sind aufgrund ihrer Architektur leicht erweiterbar.

Der Prozess im Überblick

- Start der Initiative 1.10.2018
- Roadshow Data Lake Prototyp 22.10.2018
- Initiierung des Data Lake Teams 14.3.2019
- Aufbau der Plattform und fachliche Implementierung
- Dedizierte Hardware ist geliefert 20.9.2019
- Dedizierte Hardware ist einsatzbereit 20.11.2019
- Erstes Produktionsrelease 11.12.2019
- Aufbau Prozesse für Qualitätssicherung und Release
- Start Continuous Release alle 3 Wochen 15.01.2020
- Fachlicher Ausbau



„Die neue Lösung ermöglicht uns ganz neue Anwendungsfälle, die das Unternehmen bestimmt weiterbringen.“

-Alexandra, Product Owner

Fazit

TIMOCOM ist es gelungen, in nur neun Monaten (654 Personentagen) mit einem 7-köpfigen Team eine vollständig neue Big Data Plattform in Produktion zu bringen und zu betreiben. Die Plattform stellt die Basis für neue, anspruchsvolle Anwendungsfälle im Bereich Advanced Analytics

und Machine Learning dar. Das neue Team dient als Lernraum und Leuchtturm für die weitere agile Ausrichtung des Unternehmens. Damit macht sich TIMOCOM fit für die Zukunft im Bereich Big Data, maschinellem Lernen sowie beim Einsatz moderner, agiler Methoden.

„Es lohnt sich, eingetretene Pfade zu verlassen, um Innovation zu erreichen.“

- Michael, Data Lake Team



Über TIMOCOM

Die TIMOCOM GmbH ist ein mittelständisches FreightTech-Unternehmen mit mehr als 500 Mitarbeitern aus über 30 Nationen und Repräsentanzen in Polen, Tschechien und Ungarn. Der IT- und Datenspezialist unterstützt seine Kunden mit smarten, sicheren und einfachen Lösungen dabei, ihre logistischen Ziele zu erreichen. Über das Smart Logistics System von TIMOCOM vernetzen sich mehr als 43.000 geprüfte Unternehmen aus ganz Europa, die täglich bis zu 750.000 internationale Fracht- und Laderaumangebote einstellen.

www.timocom.de

PLEUS CONSULTING

Über PLEUS Consulting

PLEUS Consulting unterstützt Unternehmen bei der Realisierung digitaler Lösungen. Der Fokus liegt dabei gleichermaßen auf den Bereichen Technologien, Prozesse und Teams, um optimale Ergebnisse zu erreichen. Die Unterstützung in diesem Projekt lag im Bereich der Rollen Lead Developer und Agile Coach.

www.pleus.net



Individuelle Softwarelösungen

Über IKS

IKS ist ein IT-Beratungsunternehmen und Spezialist für individuelle Softwarelösungen. Als Projektpartner hat IKS gemeinsam mit PLEUS Consulting TIMOCOM zu agilem Vorgehen, Big Data und Softwarearchitektur beraten.

www.iks-gmbh.com