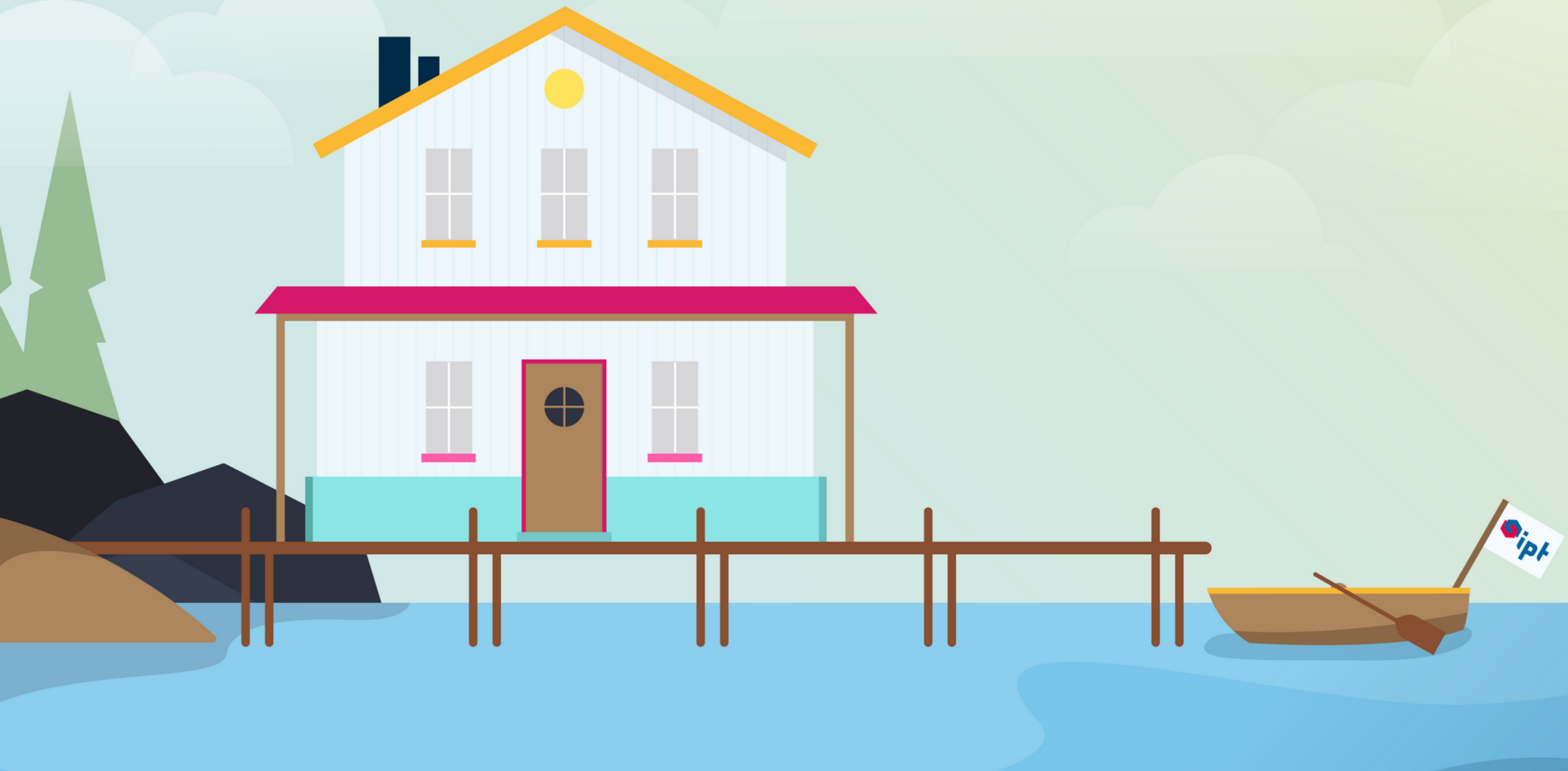


Bauplan für Dein Lakehouse

Alles, was Du beim Bau eines Lakehouse in der Cloud wissen musst –
zumindest die Kurzfassung: *Entstehung, Nutzen, Musterarchitektur & Vorgehen*



Wie ein Lakehouse entstanden ist & was es genau ist



1980
Data Warehouse (DWH)

2010
Data Lakes

2017
Lakehouse

Der bisher de-facto Standard

Seit ca. 1980 wurden DWHs gebaut. Sie eignen sich für strukturierte Daten und bestehen im Wesentlichen aus drei Elementen:

- Ein RDBMS (z.B. Oracle DB)
- Ein ETL Tool (z.B. SSIS)
- Ein Modellierungskonzept (z.B. Kimball)

Hauptnutzer sind Business Personen, welche datenbasierte Entscheidungen treffen wollen.

Herausforderungen:

DWH ist optimiert für BI und Reporting. Jedoch fehlt die Unterstützung von modernen Analytics Anforderungen wie:

- Unstrukturierte Daten, wie z.B. Video, Audio, PDF
- Data Science und Machine Learning (ML)
- Streaming Daten
- Entkopplung Speicher & Verarbeitung

Ein See aus Daten

Seit ca. 2010 werden Data Lakes gebaut. Sie eignen sich auch für semi-strukturierte und unstrukturierte Daten, weil sie keine Vorgaben über das eingehende Format machen (schema-on-read). Data Lakes bestehen aus:

- Günstiger Speicherlösung (z.B. HDFS, S3) &
- Separatem Processing Layer (z.B. Yarn, Spark)

Die *Hauptnutzer* sind Data Scientists, die experiment-getrieben arbeiten wollen.

Herausforderungen:

Ein Data Lake erfüllt alle Anforderungen für Data Science und ML. Jedoch hat sich herausgestellt, dass im Versicherungs-, Banken- und Bundesumfeld die Anwendungsfälle heute stark strukturiert sind und sich dadurch folgende Herausforderungen ergeben:

- Anbindung von BI
- Komplexer Aufbau
- Spezifische Performance Anforderungen
- Stringentes Datenmanagement

Das beste aus zwei Welten

Der Begriff Lakehouse wird seit 2017 verwendet und wurde durch verschiedene Softwareanbieter geprägt. Lakehouse beinhaltet das Beste aus 2 Welten, nämlich:

- Data-Lake-first Ansatz - Daten werden im offenen Format im Objektspeicher abgespeichert
- Wichtige Eigenschaften von DWHs, wie z.B. Transaktionen oder Delta-Beladung

Damit ermöglichst Du Deinem Unternehmen, eine Datenplattform für alle Bedürfnisse zu schaffen.



Der Nutzen & die Vorteile eines Lakehouse



1

Mit dem Lakehouse holst Du Dir Real-time.

Steht die Cloud Foundation sowie die Lakehouse Plattform, dann kannst Du das volle Potential aus deinem Lakehouse schöpfen:

- Neue Möglichkeiten für ML & Data Science mit Open-source
- Real-time Reporting durch Streaming-Unterstützung
- Anreicherung von Reporting mit ML Modellen, wie z.B. Forecasting

Tipp: Investiere in diesem Schritt in diese Skills:

- Infrastruktur-Automatisierung
- CI/CD
- Verstehen von strukturierten Entwicklungsprozessen

2

Baue eine Plattform für all Deine analytischen Zwecke.

Eine Plattform reduziert Redundanzen, den Engineering- und Betriebsaufwand. Sowohl DWHs als auch ML Workloads können mit einem Lakehouse bedient werden.

Tipp: Schaue Dir den Data Mesh Ansatz an und vermeide so gut es geht zentrale Bottlenecks.

3

Profitiere von nachhaltiger Innovation durch die Cloud.

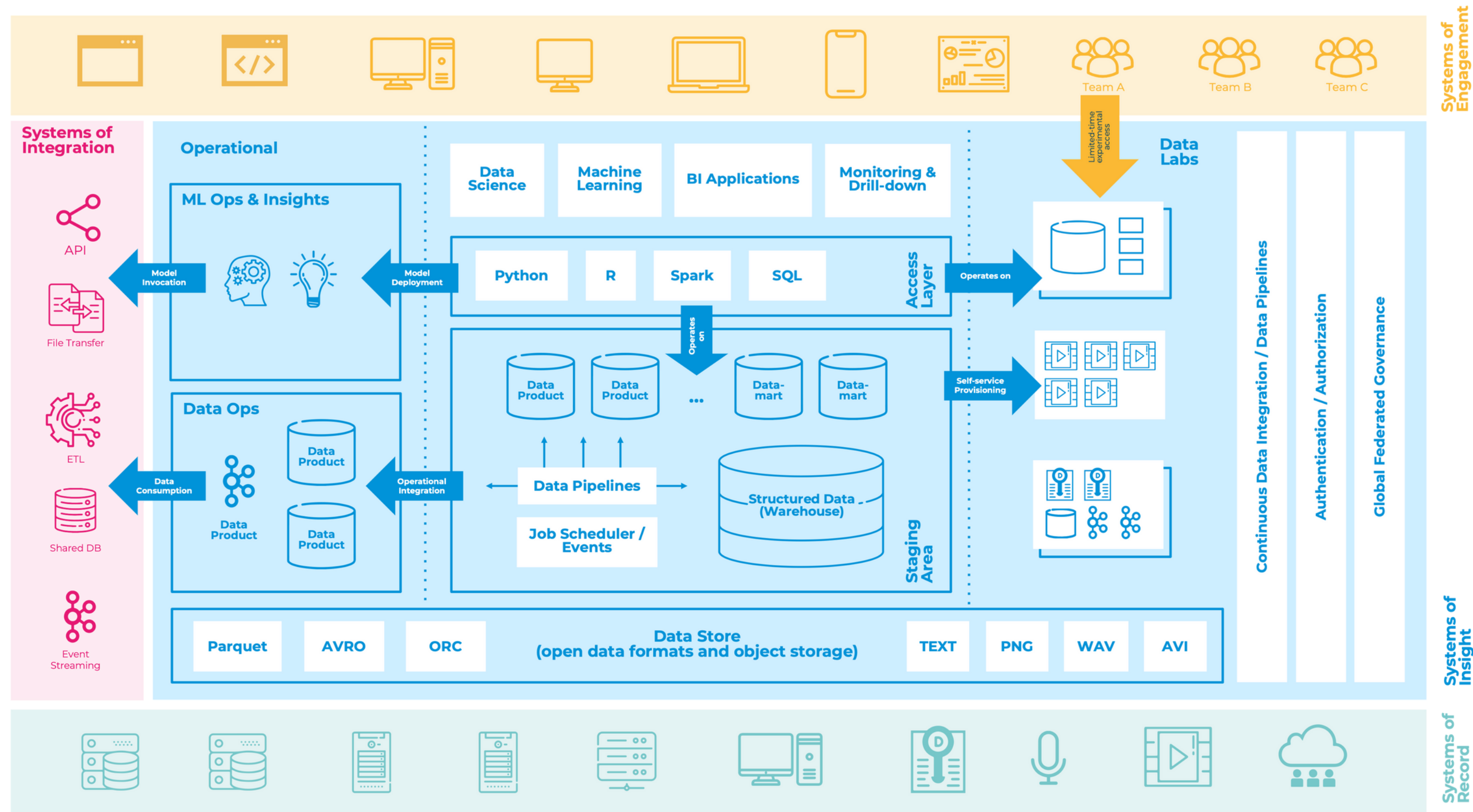
Die Vorteile von Cloud gelten auch für die Datenplattformen:

- Schnelle Provisionierung
- Kürzere Time to Market
- Technologische Innovation durch PaaS & SaaS
- Skalierung durch Serverless

Tipp: Setze auf Cloud, um die Skalierbarkeit deines Lakehouse zu gewährleisten. Zuerst solltest Du daher in Deine Cloud Foundation investieren.



Deine Lakehouse Musterarchitektur





Im Lakehouse gilt das Data-Lake-first Prinzip. Das zeigt sich in der Speicherschicht, indem die Daten in erster Linie im Objektspeicher und in offenen Dateiformaten (z.B. parquet) gespeichert werden. Darüber hinaus können je nach Bedarf Anforderungen in Richtung relationale Datenbanken, nicht-relationale Datenbanken und Streaming Daten abgedeckt werden.



Spark, SQL, Python und R sind die dominanten Sprachen resp. Frameworks für die Arbeit mit Big Data und anderen analytischen Daten. Mit diesen Werkzeugen werden Datenpipelines gebaut, die die Daten automatisiert über verschiedene Stufen aufbereiten, analysieren und zur Verfügung stellen können. Modellierung und Struktur werden den Bedürfnissen des Unternehmens angepasst.



Analytische Erkenntnisse müssen erarbeitet und geteilt werden. Dazu können den Data Scientisten zum einen transiente Arbeitsbereiche zur Verfügung gestellt werden (rechte Seite). Erfolgreiche Ergebnisse werden in den permanenten Bereich überführt, von wo aus sie als Datenprodukte auch in operative Systeme integriert werden können (linke Seite).

Wie Du beim Bau eines Lakehouse vorgehen kannst



Unternehmen, die
ipt darin vertrauen:

sanitas

IX

SWICA

Swiss Re

Hast Du weitere Fragen zum Lakehouse?

[Hier](#) findest Du weitere Infos oder kontaktiere uns gleich persönlich:



Andy Langner
andy.langner@ipt.ch
+41 76 639 31 11



Gerald Reif
gerald.reif@ipt.ch
+41 79 226 20 14



Lukas Bieri
lukas.bieri@ipt.ch
+41 79 128 30 53



Yves Brise
yves.brise@ipt.ch
+41 79 792 62 18



Yu Li
yu.li@ipt.ch
+41 76 700 15 44