Einsatz von dbt im Data Lakehouse

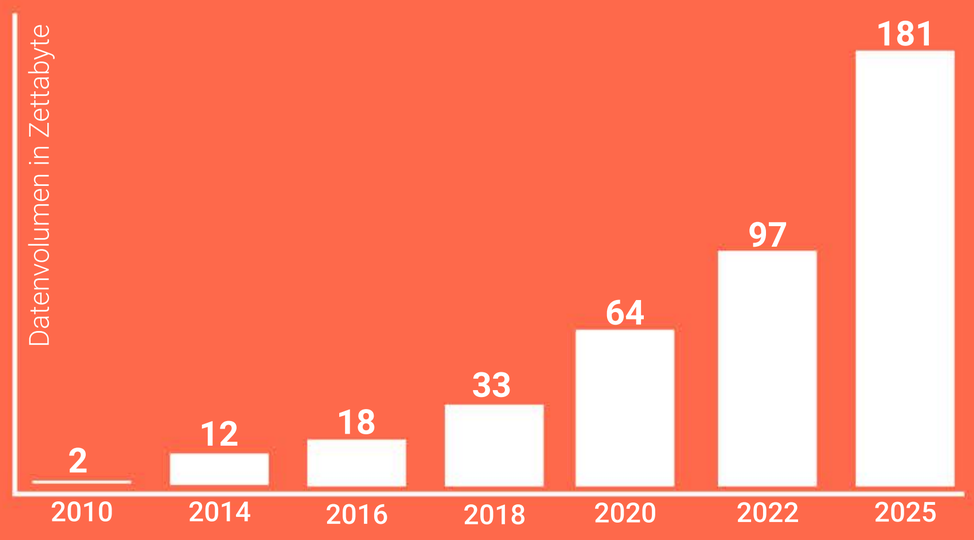
Technische Hochschule Rosenheim

x QUNIS

WARUM?

Unternehmen brauchen ***Daten***: Schnell, korrekt, effizient

- Bis 2025 wird es ***~181 Zettabyte*** an Daten geben
- 80% der Daten werden davon ***unstrukturiert*** sein
- Mit traditionellen Ansätzen und Tools ist das nicht zu bewältigen



WER?

- Wirtschaftsinformatik Bachelor TH Rosenheim
- 7. Semester Modul Praxisprojekt
- Zusammenarbeit mit QUNIS GmbH

QUNIS GmbH

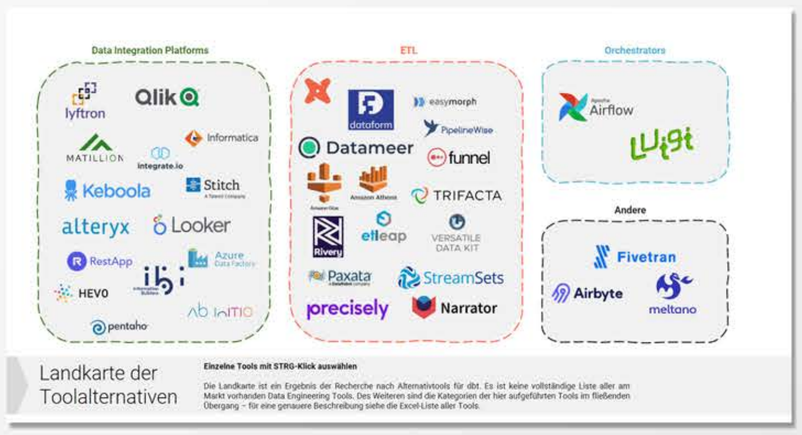
- IT Strategie-Beratung und Umsetzung
- BI, Advanced Analytics, Data Science und Daten-architekturen
- Vorreiter in Deutschland

"Ist dbt das richtige und beste Tool für unser Data Lakehouse?"

PROJEKTERGEBNISSE:

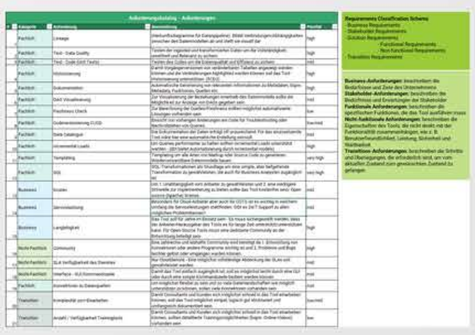
1. MARKTRECHERCHE

39 Tools identifiziert, klassifiziert und kommentiert



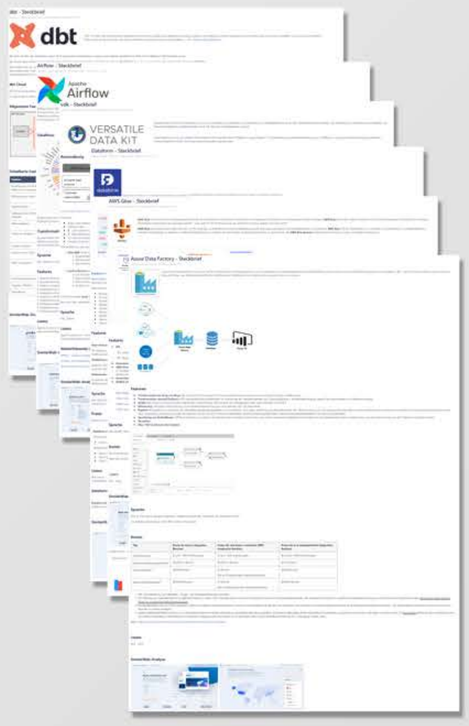
2. ANFORDERUNGSKATALOG

Aus Stakeholder Personas, Industrie Standards und Anwenderbedürfnissen wurde ein Katalog mit Fach- und Businessanforderungen als Bewertungs-grundlage erstellt.



3. STECKBRIEFE

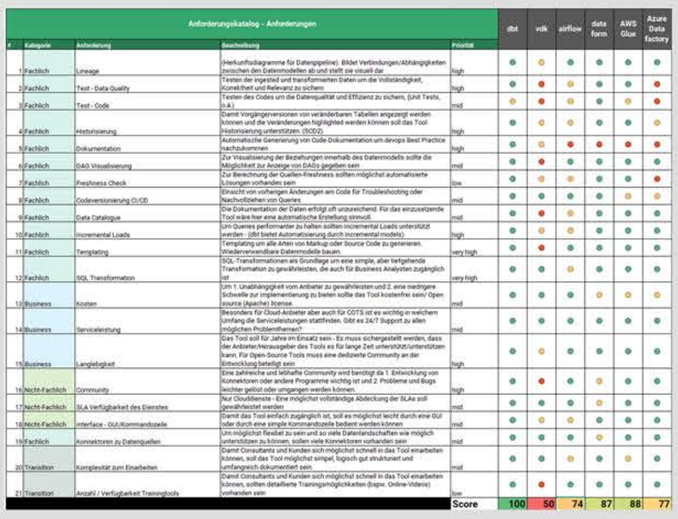
Für die 5 wichtigsten Tools und dbt wurden detaillierte Steckbriefe mit relevanten Informationen erstellt, um QUNIS Consultants bei Kundengesprächen zu unterstützen.



4. EVALUATION

Durch detaillierte und begründete Bewertung von über 120 Faktoren wurde ein zukunftsfähiges Scoring-Sheet erstellt, um die beste Toolalternative einzuschätzen.

Die Matrix kann zur automatischen Evaluierung spezifisch auf Kundenwünsche angepasst werden.



Anforderungskatalog - Anforderungen					dbt	vdk	airflow	data form	AWS Glue	Azure Data factory
1	Fachlich	Lineage	(Herkunftsdiagramme für Datenpipeline). Bildet Verbindungen/Abhängigkeiten zwischen den Datenmodellen ab und stellt sie visuell dar	high	●	○	●	●	●	●
2	Fachlich	Test - Data Quality	Testen der ingested und transformierten Daten um die Vollständigkeit, Korrektheit und Relevanz zu sichern	high	●	●	○	●	●	●
3	Fachlich	Test - Code	Testen des Codes um die Datenqualität und Effizienz zu sichern, (Unit Tests, o.ä.)	mid	○	●	○	●	○	●
4	Fachlich	Historisierung	Damit Vorgängerversionen von veränderbaren Tabellen angezeigt werden können und die Veränderungen highlighted werden können soll das Tool Historisierung unterstützen. (SCD2).	high	●	○	○	●	●	○
5	Fachlich	Dokumentation	Automatische Generierung von Code Dokumentation um devops Best Practice nachzukommen	high	●	○	●	●	●	●
6	Fachlich	DAG Visualisierung	Zur Visualisierung der Beziehungen innerhalb des Datenmodells sollte die Möglichkeit zur Anzeige von DAGs gegeben sein	mid	●	●	●	●	●	●
7	Fachlich	Freshness Check	Zur Berechnung der Quellen-Freshness sollten möglichst automatisierte Lösungen vorhanden sein	low	●	○	○	●	●	●
8	Fachlich	Codeversionierung CI/CD	Einsicht von vorherigen Änderungen am Code für Troubleshooting oder Nachvollziehen von Queries	mid	●	●	●	●	○	○
9	Fachlich	Data Catalogue	Die Dokumentation der Daten erfolgt ich unzureichend. Für das einzuselzende Tool wäre hier eine automatische Erstellung sinnvoll.	mid	●	●	○	●	●	●
10	Fachlich	Incremental Loads	Um Queries performanter zu halten sollten incremental Loads unterstützt werden - (dbt bietet Automatisierung durch incremental models)	high	●	○	○	●	●	●
11	Fachlich	Templating	Templating um alle Arten von Markup oder Source Code zu generieren. Wiederverwendbare Datenmodelle bauen.	very high	●	●	●	●	●	●
12	Fachlich	SQL Transformation	SQL-Transformationen als Grundlage um eine simple, aber fortgehende Transformatnes zu gewährleisten, die auch für Business Analysten zugänglich ist	very high	●	●	○	●	●	●
13	Business	Kosten	Um 1. Unabhängigkeit vom Anbieter zu gewährleisten und 2. eine niedrigere Schwelle zur Implementierung zu bieten sollte das Tool kostenfrei sein/ Open source (Apache) lizense.	mid	●	●	●	●	●	○
14	Business	Serviceleistung	Besonders für Cloud-Anbieter aber auch für COTS ist es wichtig in welchem Umfang die Serviceleistungen stattfinden. Gibt es 24/7 Support zu allen möglichen Problemthemen?	mid	●	●	●	●	●	●
15	Business	Langlebigkeit	Das Tool soll für Jahre im Einsatz sein - Es muss sichergestellt werden, dass der Anbieter/Herausgeber des Tools es für lange Zeit unterstützt/unterstützen kann. Für Open-Source Tools muss eine dedizierte Community an der Entwicklung beteiligt sein	high	●	○	●	●	●	●
16	Nicht-Fachlich	Community	Eine zahlreiche und lebhafte Community wird benötigt da 1. Entwicklung von Konnektoren oder andere Programme wichtig ist und 2. Probleme und Bugs leichter gelöst oder umgangen werden können.	high	●	●	●	●	●	●
17	Nicht-Fachlich	SLA Verfügbarkeit des Dienstes	Nur Cloudlösnste - Eine möglichst vollständige Abdeckung der SLAs soll gewährleistet werden	mid	●	●	●	○	●	●
18	Nicht-Fachlich	Interface - GUI/Kommandozeile	Damit das Tool einfach zugänglich ist, soll es möglichst leicht durch eine GUI oder durch eine simple Kommandozeile bedient werden können	mid	●	○	○	●	●	●
19	Fachlich	Konnektoren zu Datenquellen	Um möglichst flexibel zu sein und so viele Datenlandschaften wie möglich unterstützen zu können, sollen viele Konnektoren vorhanden sein	mid	●	●	●	○	●	●
20	Transition	Komplexität zum Einarbeiten	Damit Consultants und Kunden sich möglichst schnell in das Tool einarbeiten können, soll das Tool möglichst simpel, logisch gut strukturiert und umfangreich dokumentiert sein.	mid	●	●	●	●	●	●
21	Transition	Anzahl / Verfügbarkeit Trainingstools	Damit Consultants und Kunden sich möglichst schnell in das Tool einarbeiten können, sollten detaillierte Trainingsmöglichkeiten (bspw. Online-Videos) vorhanden sein	low	●	●	●	●	●	●
				Score	100	50	74	87	88	77