Der unverzichtbare Leitfaden für die Datenintegration

Endlose Datenmengen und wie man sie sich zunutze macht

Von Charles Wang

In diesem Leitfaden behandelte Themen:

- Die Datenintegration als Treibstoff für die Datenanalyse
- Die Entwicklung von ETL über ELT zur automatisierten Datenintegration
- Die Vorteile der automatisierten Datenintegration
- Bewertung von Datenintegrationsanbietern

Fivetran

fivetran.com →

Inhaltsverzeichnis

Zu diesem Buch

Ziele

In diesem Buch wird erläutert, wie hilfreich die Datenintegration für Ihr Unternehmen ist, welche unterschiedlichen Konzepte der Datenintegration gegenwärtig verfügbar sind, was diese bieten und wie Sie die Technologie implementieren können.
Sie müssen das Buch nicht von vorne bis hinten durchlesen, es ist allerdings so aufgebaut, dass sich sein Inhalt so am einfachsten erschließt.

Die **Datenintegration** umfasst die Prozesse, die dem Verwalten und zentralen Zusammenführen von Datenflüssen aus verschiedenen Quellen dienen – mit dem Ziel, diese Daten als Entscheidungshilfe zu nutzen. Die praktische Interpretation von Daten als Orientierungshilfe bei Entscheidungen wird häufig als **Analytics** (Datenanalyse) bezeichnet. Wie wir sehen werden, sind die Qualität Ihres Analyseprogramms und die Qualität Ihrer Datenintegrationstechnologie eng miteinander verknüpft.

Dank Datenintegration kann Ihr Unternehmen sämtliche Daten in einer einzigen Umgebung pflegen. Dadurch hat Ihr Team einen umfassenden Überblick über Geschäftsabläufe und Kundeninteraktionen. Die Zentralisierung von Daten und damit ihre Zugänglichkeit fördert eine weitreichende Datenkompetenz, die Unternehmen in die Lage versetzt, versteckte Chancen zu erkennen, ihre Performance zu steigern und ihre Innovationskraft zu erhöhen.

In diesem Leitfaden werden folgende Themen besprochen:

- Datenintegration und ihre Bedeutung
- Der traditionelle Ansatz der Datenintegration: ETL (Extrahieren, Transformieren, Laden)
- Der neuere Ansatz, der durch die Cloud möglich wird: ELT (Extrahieren, Laden, Transformieren)
- Die Vorteile der Automatisierung des Datenintegrationsprozesses
- Die Evaluierung und Einführung von Datenintegrationstools

Zielgruppe

Zur optimalen Nutzung dieses Leitfadens sollten Sie mit Data Engineering, Data Warehousing, Datenanalyse, Business Intelligence, Datenvisualisierung und verwandten Konzepten vertraut sein. Wir gehen davon aus, dass Sie in einem Unternehmen beschäftigt sind, das Betriebssysteme, Anwendungen und andere Tools nutzt, die digitale Daten erzeugen, und dass ein Teil Ihrer betrieblichen Abläufe bereits über die Cloud läuft. Zudem gehen wir davon aus, dass Sie aufgrund Ihrer Funktion – Analyst, Data Engineer, Datenwissenschaftler oder Leiter eines Datenanalyse-Teams – Einfluss darauf haben, mit welchen Tools Ihr Unternehmen arbeitet.

In diesem Leitfaden verwendete Symbole

Beim Lesen werden Sie Symbole für Tipps, Warnungen, wichtige Punkte und Fallstudien sehen.

TIPP: *Praktische Empfehlungen bezüglich Datenintegration und analyse*

ACHTUNG: *Flle von falschem Umgang mit Daten oder Technik*

MERKE: *Wichtige Punkte, die man sich merken sollte*

FALLSTUDIE: *Erfolgsgeschichten im Bereich Datenintegration aus der Praxis*

Über diesen Leitfaden hinaus

Wenn Sie diesen Leitfaden nützlich finden und mehr erfahren möchten, besuchen Sie fivetran.com/blog. Dort veröffentlichen wir neue Informationen zu Data Engineering und Datenanalyse. Eine weitere gute Informationsquelle ist unsere Dokumentation unter fivetran.com/docs. Sie vermittelt einen detaillierten Überblick über die Funktionsweise der automatisierten Datenintegration für bestimmte Datenquellen und -ziele.

Kapitel 1: Datenintegration und -analyse

INHALT DIESES KAPITELS:

- Die Geschichte der Datenanalyse
- Der Wert der Datenanalyse für Sie und Ihr Unternehmen
- Ein großes Hindernis für die Datenanalyse: die Datenintegration

Die Geschichte der Datenanalyse

Die Datenanalyse gab es schon lange vor der modernen Datenerfassung. Bereits Florence Nightingale analysierte mithilfe von Polar-Diagrammen die Ursachen der Krankenhaussterblichkeit während des Krimkrieges und trug damit zu ihrer Senkung bei (Abbildung 1.0). William Sealy Gosset, Chefbraumeister bei Guinness, entwickelte den Student-T-Test zur Gewährleistung der Bierqualität. Schon seit Langem ziehen die Menschen aus Zahlen wertvolle Lehren und Erkenntnisse.

Seit dieser Zeit hat die Statistik als Wissenschaft eine enorme Entwicklung durchlaufen – genau wie die Werkzeuge und Methoden zur Analyse von Daten. Vor allem das exponentielle Wachstum der Rechenleistung und das Entstehen des Internets ermöglichten die Erfassung und Analyse von Daten in viel größerem Maßstab, als dies noch mit Stift, Papier und Rechenmaschinen möglich war.

Die Ziele der Datenanalyse

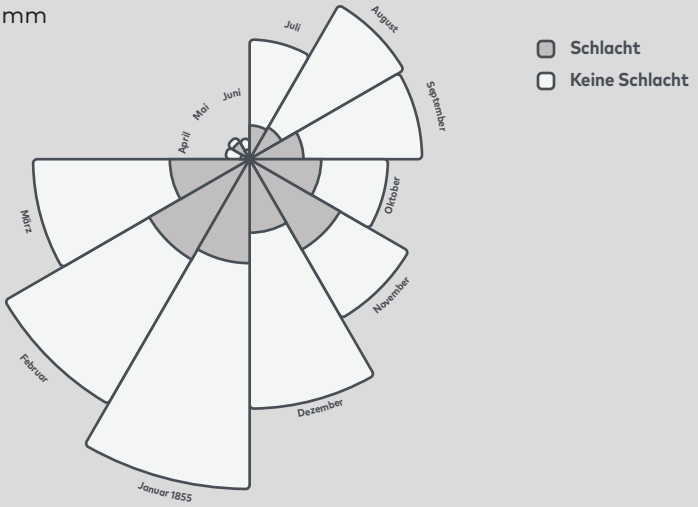
Die Datenanalyse bietet verschiedene Wettbewerbsvorteile. Sie kann zur Optimierung

von Kundengewinnung und -bindung, zur Ermittlung neuer möglicher Produkte und zur Optimierung bestehender Chancen eingesetzt werden. Durch die Verbesserung der Entscheidungsfindung im Unternehmen können Datenanalysen ein Vielfaches ihrer Kosten an ROI einspielen.

Grob gesagt gibt es folgende Anwendungsmöglichkeiten für Datenanalysen:

1. **Ad-hoc-Berichte:** Wichtige Stakeholder und Entscheidungsträger benötigen mitunter Antworten auf sehr spezifische Fragen, die einmalig oder hin und wieder beantwortet werden müssen.
2. **Business intelligence:** Business Intelligence (BI) wird häufig synonym mit „Datenanalyse" oder „Analytics" verwendet und bezeichnet die Verwendung von Visualisierungen und Datenmodellen zur Ermittlung von Chancen und zur Steuerung von Geschäftsentscheidungen und -strategien. Dies erfolgt normalerweise in Form von regelmäßigen, einheitlichen Berichten und aktuellen Dashboards.
3. **Daten als Produkt:** Daten, die Ihr Unternehmen sammelt oder erzeugt, können in Form von eingebetteten Dashboards, Datenströmen, Empfehlungen oder anderen Datenprodukten Dritten zur Verfügung gestellt werden.

Abbildung 1.0
Polardiagramm



Quelle: „Notes on Matters Affecting the Health, Eficiency and Hospital Administration of the British Army", von Florence Nightingale. London: Harrison & Sons, 1858.

4. **Künstliche Intelligenz / maschinelles Lernen:** In ihrer höchsten Entwicklungsstufe besteht die Datenanalyse darin, Produkte und Systeme zu entwickeln, die mithilfe von Vorhersagemodellen wichtige Entscheidungen und Prozesse automatisieren.

Auf unternehmerischer Ebene kann Ihnen die Datenanalyse auch bei Folgendem helfen:

1. **Demokratisierung des Zugangs zu Daten/Datenkompetenz:** Je mehr Mitarbeiter Daten als Entscheidungsgrundlage nutzen, desto klüger reagiert Ihr Unternehmen auf sich ändernde Umstände. Mit den richtigen BI-Tools können auch technisch nur wenig versierte Teammitglieder auf Daten gestützte Entscheidungen treffen. Dazu müssen Sie Ihrem Team natürlich viel Vertrauen und Spielraum gewähren – haben dadurch aber auch die Möglichkeit dazu.
2. **Verbesserung Ihrer Produkte und Dienstleistungen:** Die aus der Datenanalyse gewonnenen Erkenntnisse helfen Ihnen, Ihre Angebote zu verbessern und Ihren Kunden zusätzliche Transparenz- und Berichtsoptionen zu bieten.
3. **Sichern der Wettbewerbsfähigkeit Ihres Unternehmens:** Datenkompetenz ermöglicht es Ihnen, begrenzte Ressourcen optimal zu nutzen und Chancen zu erkennen, die Ihnen sonst verborgen bleiben würden.

Wissen ist Macht, und es ist immer von Vorteil, gegenüber der Konkurrenz einen Wissensvorsprung zu haben.

TIPP: *Es kann hilfreich sein, sich alle datenbezogenen Aktivitäten als Phasen einer Bedürfnispyramide vorzustellen, bei der die Befriedigung der Grundbedürfnisse die Voraussetzung für das Streben nach höheren Bedürfnissen bildet (Abbildung 1.1).*

Das grundlegendste Bedürfnis ist das Sammeln und Speichern von Rohdaten, d. h. die Datenintegration. Wenn dieses Bedürfnis erfüllt ist, lassen sich die nächsten Bedürfnisse – Datenanalyse und Prognosemodelle – einfacher erfüllen. Auf diese Weise kann Ihr Unternehmen eine datengesteuerte Kultur schaffen, bei der jeder Mitarbeiter Zugriff auf die Daten hat, die er benötigt, um fundiertere Entscheidungen zu treffen.

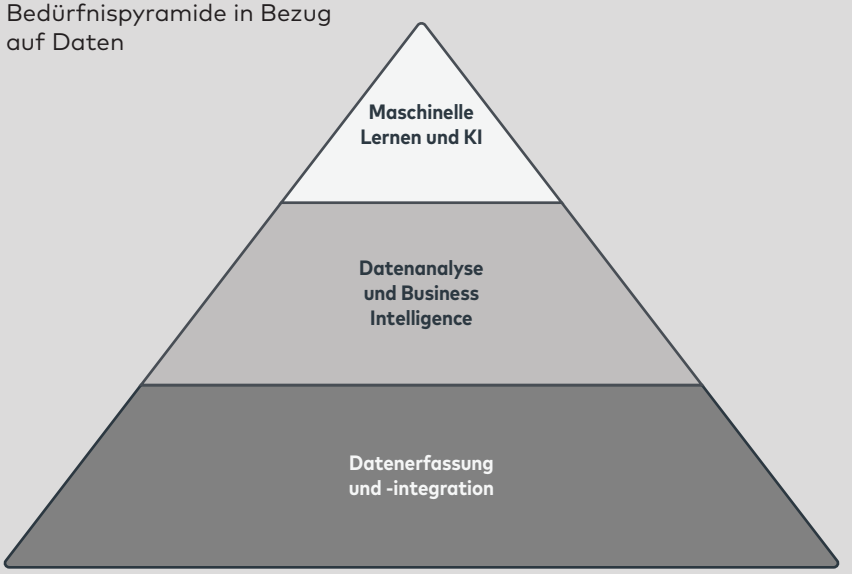
An der Spitze der Pyramide werden Daten für das Trainieren von Modellen für das maschinelle Lernen und die Entwicklung von künstlicher Intelligenz genutzt. Das ermöglicht seinerseits die Automatisierung von Workflows und Entscheidungsprozessen in Ihrem Unternehmen sowie die Entwicklung „intelligenter" verbraucherorientierter Produkte.

Ein großes Hindernis für die Datenanalyse: die Datenintegration

AEin zentrales Daten-Repository für Datensätze bietet Ihrem Unternehmen folgende Vorteile:

1. Sie erhalten einen Überblick über die Abläufe Ihres Unternehmens und sehen, wie die einzelnen Teile ineinandergreifen, statt sich einzelne, isolierte Bereiche anzuschauen.
2. Sie können Datensätze abgleichen und dieselben Entitäten (Kunden, Partner usw.) über verschiedene Phasen ihres Lebenszyklus hinweg verfolgen.
3. Sie können Analysen in einer von den operativen Systemen getrennten Umgebung durchführen, um zu verhindern, dass Ihre Abfragen Ihren Betrieb beeinträchtigen.
4. Sie üben gezielte Kontrolle über Zugriff und Berechtigungen aus. Damit stellen Sie sicher, dass Ihr Team die Informationen erhält, die es zur Ausführung seiner Aufgaben benötigt, ohne sensible Systeme zu gefährden.

Abbildung 1.1
Bedürfnispyramide in Bezug auf Daten



Das Erstellen dieses zentralen Daten-Repositorys kann eine Herkulesaufgabe sein. Jede Datenquelle erfordert andere Verfahren und Tools für die Erfassung, Bereinigung und Modellierung der zugehörigen Daten. Die Verbreitung von Cloud-basierten Anwendungen und Diensten ließ diese Herausforderung noch wachsen. Das Auftauchen von webfähigen Geräten und Sensoren (d. h. das Internet der Dinge) trug ebenfalls zum explosionsartigen Wachstum der Datenmengen bei (Abbildung 1.2). Seit 2013 gilt es als Gemeinplatz, dass 90 % der weltweiten Daten in den zwei zurückliegenden Jahren erzeugt wurden.[1]

Woher stammen die Daten?

Daten können aus folgenden Quellen stammen:

1. **Sensoreingänge**, z. B. Scans an einer Ladenkasse
2. **Manuelle Eingabe von Daten**, z. B. über Formulare, die vom statistischen Bundesamt gesammelt werden
3. **Digitale Dokumente und Inhalte**, z. B. Postings in sozialen Medien
4. **Digitale Aktivitäten** die von Software-Triggern aufgezeichnet werden, z. B. Klicks auf einer Webseite oder in einer App

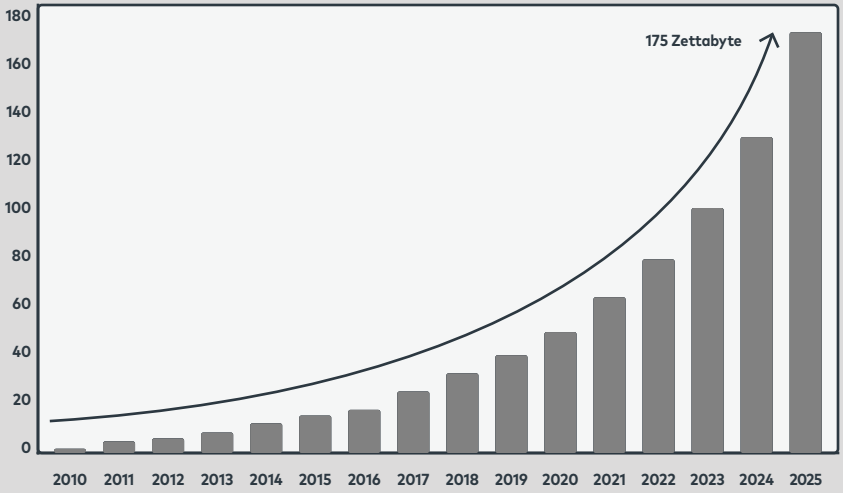
Daten aus den genannten Quellen werden in der Regel in Cloud-basierten digitalen Dateien und operativen Datenbanken gespeichert und einem Endbenutzer dann in folgender Form zugänglich gemacht:

1. API-feeds
2. Dateien
3. Datenbankprotokolle und Abfrageergebnisse
4. Ereignisverfolgung

API-feeds ermöglichen es Anwendungen, miteinander zu kommunizieren, häufig durch Datenaustausch in Formaten wie JSON oder XML. Die meisten Unternehmen verwenden eine breite Palette von Anwendungen, um Vorgänge wie Kundenbeziehungsmanagement, Fakturierung und Kundendienst abzuwickeln. Weitere APIs ermöglichen die Datenaufnahme und die Interoperabilität zwischen Softwareanwendungen.

1 sciencedaily.com/releases/2013/05/130522085217.htm

Abbildung 1.2
Jährliche Größe der globalen Datensphäre



Quelle: Data Age 2025, gefördert von Seagate mit Daten von IDC Global DataSphere, November 2018

Datendateien wie CSV, XLSX und TSV können aus mehreren Aktivitäten in einem Unternehmen stammen, von der manuellen Datenerfassungbiszu Ad-hoc-Berechnungen.

Datenbankprotokolle und Abfrageergebnisse werden von operativen Datenbanken generiert, die in Echtzeit aktualisiert werden. Sie unterstützen die täglichen Interaktionen für alle Datenquellen von Sensoren bis zu Software. Eine E-Commerce-Website kann eine operative Datenbank beispielsweise nutzen, um Einkäufe, Angebote und Kundenprofile aufzuzeichnen.

Die **Ereignisverfolgung** erfolgt über vom Benutzer ausgelöste Code-Snippets, die in Webseiten und Anwendungen eingebettet sind. Ein einfaches Tool kann Klicks auf Schaltflächen in einer App aufzeichnen; ein komplexeres Tool kann die Cursorposition verfolgen. Noch ausgefeiltere Tools greifen auf die Laptop-Kamera zu, um die Augenbewegungen des Benutzers zu verfolgen. Ereignisverfolgungsdaten erzeugen ein detailliertes Protokoll der Interaktion der Benutzer mit einer Website oder Anwendung. Besonders hilfreich sind sie für die UI/UX-Forschung. Eine der gängigsten Formen ist der **Webhook**. Er ist in Webanwendungen eingebettet und wird über HTML gesendet – statt formatiert in XML oder JSON.

Wie Sie sich vorstellen können, stellt die schiere Vielfalt an Datenquellen und -formaten die Data Engineers vor große Herausforderungen, was die Integration und Normalisierung von Datenströmen angeht.

TIPP: *Ihnen ist im Zusammenhang mit verschiedenen Datenvorgängen, einschließlich der Datenintegration, vielleicht schon mal der Begriff „Datenintegrität" begegnet. Datenintegrität bezeichnet die Vollständigkeit, Genauigkeit und Konsistenz von Daten in allen Phasen ihrer Verwendung. Eine Verletzung der Datenintegrität liegt u. a. dann vor, wenn Daten falsch eingegeben oder falsch formatiert wurden, bzw. bei Duplikaten, Auslassungen und falschen Beziehungen zwischen Tabellen.*

SaaS-Daten: Eine wachsende Herausforderung und Chance zugleich

Seit Beginn der Cloud-Ära sind SaaS-Anwendungen eine der wichtigsten Quellen für Geschäftsdaten. Sie decken eine Vielzahl von betrieblichen Vorgängen und Branchen ab: Marketing, Zahlungsabwicklung, Kundenpflege, E-Commerce, technisches Projektmanagement und Vieles mehr. Sie bieten ausgefeilte Dienste und Funktionen und machen die interne Entwicklung von Tools bzw. die manuelle Ausführung derselben Aufgaben mit massivem Arbeitsaufwand überflüssig.

SaaS-Anwendungen zeichnen in der Regel Aktionen von Benutzern auf und bieten Unternehmen ein sehr detailliertes Bild von ihren betrieblichen Vorgängen – aus denen sich dann Muster und Kausalzusammenhänge ableiten lassen. Je mehr Bereiche eines Unternehmens sich quantifizieren und analysieren lassen, desto wettbewerbsfähiger ist das Unternehmen im Allgemeinen.

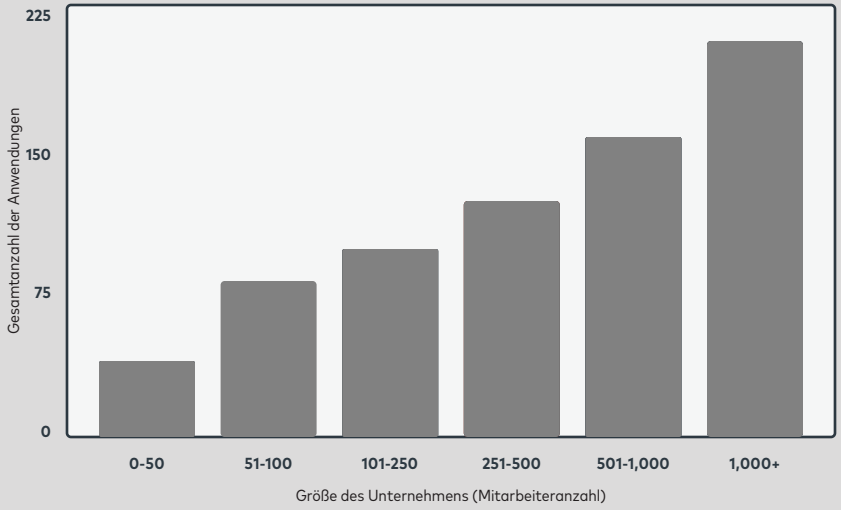
Riesige Datenmengen sind jedoch zugleich eine enorme Herausforderung für die Datenintegration. Unternehmen nutzen heute im Schnitt mehr als 100 Anwendungen (Abbildung 1.3). In diesem Maßstab ist die manuelle Datenintegration praktisch unmöglich. Wie wir sehen werden, entwickeln viele Unternehmen immer noch kundenspezifische Software und eine eigene Infrastruktur zur Integration von Daten. Wenn Daten aus Dutzenden von Quellen stammen, die einen kontinuierlichen Datendurchsatz mit hohem Datenvolumen generieren, ist der dafür zu betreibende Aufwand jedoch nicht mehr vertretbar.

Selbst in kleinerem Maßstab kann die Arbeitslast, die mit der Erstellung und Wartung komplexer Datenpipeline-Software einhergeht, die für die Datenanalyse verfügbaren Ressourcen beschneiden. Hoher Zeitaufwand für andere Aufgaben hält Analysten, Datenwissenschaftler und Ingenieure von anderen Aktivitäten ab.

Glücklicherweise bietet die Cloud-Technologie dafür eine Lösung. Moderne Datenpipeline-Tools, Data Warehouses und Business Intelligence-Plattformen sind eigenständige Cloud-basierte Anwendungen, die im Umfeld der Cloud-Technologie

entstanden sind. Sie machen die manuelle Eigenentwicklung kundenspezifischer Tools und Lösungen für die Datenintegration und -analyse überflüssig.

Abbildung 1.3
Anzahl der Anwendungen pro Unternehmen



Quelle: Annual SaaS Trends Report 2019 von Blissfully

MERKE: *Dank der Möglichkeiten der Datenanalyse können Unternehmen ihre Arbeit optimieren. Damit die Datenanalysen leistungsfähig und umfassend sind, müssen die Daten jedoch in einer zentralen Umgebung verfügbar sein. Mit einem zentralen Daten-Repository kann Ihr Unternehmen:*

- *sich ein umfassendes Gesamtbild von den betrieblichen Vorgängen verschaffen*
- *Datensätze abgleichen und dieselben Entitäten über verschiedene Datenquellen hinweg verfolgen*
- *die Datenanalyse von den operativen Systemen getrennt halten*
- *Zugang und Berechtigungen steuern*

Datenintegration als Prozess der Zentralisierung und Bereitstellung von Daten ist äußerst wichtig und zugleich äußerst schwierig durchzuführen. Diese Herausforderung sollten Sie nicht auf die leichte Schulter nehmen.

Kapitel 2: Ansätze der Datenintegration

INHALT DIESES KAPITELS:

- Was ist Datenintegration und wie wird sie durchgeführt?
- Der traditionelle Ansatz (ETL) und der moderne Ansatz (ELT) im Vergleich
- Wie automatisiertes ELT das Problem der Datenintegration löst

Der grundlegende Datenintegrationsprozess

Datenintegration besteht aus folgenden Schritten:

1. Daten werden aus Sensor-Feeds, manuellen Dateneingaben oder Software gesammelt und in Dateien oder Datenbanken gespeichert
2. Daten werden aus Dateien, Datenbanken und API-Endpunkten extrahiert und in einem Data Warehouse zentral gespeichert.
3. Daten werden bereinigt und gemäß den Analyseanforderungen verschiedener Geschäftsbereiche zu Modellen aufbereitet.
4. Daten werden als Grundlage für Produkte oder die Erzeugung von Business Intelligence verwendet.

Die Datenintegration kann manuell, ad hoc oder programmgesteuert mithilfe von Software erfolgen. Der Ad-hoc-Ansatz ist weder replizierbar noch skalierbar, und

der programmatische Ansatz erfordert einen **Data Stack** mit einem eigenen Satz ergänzender Tools. Im verbleibenden Teil dieses Kapitels werden wir diese Konzepte sowie die Geschichte und Zukunft der Datenintegration und Data Stacks näher beleuchten.

Nicht skalierbare Ansätze der Datenintegration

Viele Unternehmen setzen bei der Datenintegration auf einen manuellen Ad-hoc-Ansatz. 62 % nutzen in der Tat Tabellenkalkulationsprogramme wie Excel und Google Sheets, um Elemente aus Datendateien zusammenzufügen und Daten zu visualisieren.[2] Das umfasst das Herunterladen von Dateien, das manuelle Ändern oder Bereinigen von Werten, das Erstellen von Zwischendateien und ähnliche Aktionen. Die Ad-hoc-Datenintegration hat eine Reihe von Nachteilen, insbesondere:

- nur für sehr kleine Datenmengen geeignet
- langsam
- anfällig für menschliche Fehler
- zu unsicher für vertrauliche Informationen
- häufig nicht reproduzierbar

Ein nachhaltigerer Ansatz besteht darin, die Silos zwischen separaten Datenquellen zu pflegen und die Lücken zwischen ihnen mit „verbindenden" Abfragen zu schließen, die mehrere Quellsysteme direkt abfragen und Daten im laufenden Betrieb zusammenführen. Dazu stehen Unternehmen SQL-Abfrage-Engines wie Presto zur Verfügung. Dieser Verbundansatz hat den Nachteil, dass er viele dynamische Teile involviert und die Leistung mit steigender Datenmenge sinkt.

Ein skalierbarer, nachhaltiger Ansatz für die Datenanalyse erfordert daher einen systematischen, replizierbaren Ansatz für die Datenintegration – einen Data Stack.

Datenintegration mit einem Data Stack

Ein **Data Stack** besteht aus Tools und Technologien, die Daten aus verschiedenen Quellen gemeinsam integrieren und analysieren. Ein Data Stack umfasst Folgendes:

2 zdnet.com/article/spreadsheets-still-dominate-business-analytics/

1. **Datenquellen:**
    a. Anwendungen
    b. Datenbanken
    c. Dateien
    d. digitale Ereignisse

2. **Datenpipeline** und **Datenkonnektoren**. Software zum Extrahieren von Daten aus einer Datenquelle und zum Laden in ein Data Warehouse. Das macht den Großteil der Datenintegration aus.

3. **Data Warehouse** und/oder **Data Lake**. Ein Daten-Repository für Datensätze, das für die dauerhafte Aufnahme großer Datenmengen ausgelegt ist. Data Warehouses sind fast immer spaltenbasiert und enthalten Daten in einer relationalen Struktur. Data Lakes hingegen sind Objektspeicher, die sowohl strukturierte Daten als auch unstrukturierte Rohdaten enthalten können.

4. **Datenmodellierung** und/oder -**transformationen**. Häufig müssen Daten durch Anwenden einer kundenspezifischen Geschäftslogik aufbereitet werden, z. B. durch Ändern von Spaltennamen oder Durchführen von Aggregationen, um sie für einen spezifischen Anwendungsfall der Datenanalyse vorzubereiten.

5. **Business-Intelligence-Tool.** Software für das Zusammenfassen, Visualisieren und Modellieren von Daten als Grundlage für das Treffen von Geschäftsentscheidungen.

ACHTUNG: *In Data Lakes und Data Warehouses wurden traditionell unterschiedliche Datentypen gespeichert, um unterschiedliche Anwendungsfälle zu ermöglichen. Data Lakes enthalten normalerweise unstrukturierte Rohdaten und sind daher „trübe". Diese unstrukturierten Daten werden nicht bereinigt, normalisiert oder transformiert, bevor sie im Zielsystem landen. Datenwissenschaftler müssen die Daten also zunächst in einen verwertbaren Zustand versetzen. Data Warehouses enthalten aus Spalten zusammengesetzte Tabellen und basieren im Allgemeinen auf traditionellen relationalen Datenbanken, die mit SQL abgefragt werden können.*

Erst in jüngster Zeit liefen Data Lakes und Data Warehouses in ihrer Entwicklung stärker zusammen. Data Lakes übernahmen schrittweise Funktionen wie ACID-Transaktionen (ACID = Atomicity, Consistency, Isolation, Durability, zu Deutsch: Atomizität, Konsistenz, Isolation, Dauerhaftigkeit) und die Durchsetzung von Schemen. Im Zuge dessen ging die „Trübheit" ihrer Daten zurück. Analog dazu unterstützen Data Warehouses, die bereits ACID-Transaktionen durchführen können, jetzt

Abgesehen von der konvergierenden Entwicklung eignen sich Data Lakes besser für Anwendungsfälle, in denen die Unterstützung von maschinellem Lernen, künstlicher Intelligenz und einem offenen Ökosystem datenwissenschaftlicher Tools höhere Priorität als die Zugänglichkeit hat. Die Endbenutzer von Data Lakes sind in der Regel hochgradig spezialisierte Datenwissenschaftler mit Erfahrung in der Nutzung von Spark, Python, Pandas und ähnlichen Tools, mit denen sich eine Vielzahl von Datentypen in großem Maßstab handhaben lassen. Data Warehouses eignen sich besser für operative Datenanalysen und Business Intelligence-Anwendungsfälle, bei denen sich die Endbenutzer hauptsächlich auf SQL- und BI-Dashboards stützen.

Der Fluss von Daten durch den Stack

Die Grundeinheit in der Datenpipeline ist eine Software, die als **Datenkonnektor** bezeichnet wird. Eine Datenpipeline kann einen oder mehrere Konnektoren enthalten. Jeder von ihnen extrahiert Daten aus einer Quelle – einer Anwendung, einem Event-Tracker, einer Datei oder einer Datenbank – und führt in der Regel eine Normalisierung und grundlegende Bereinigung durch.

Anschließend werden die Daten an ein **Data Warehouse**. **Transformationen** können entweder vor Eintreffen der Daten in einem Data Warehouse oder nach ihrem Eintreffen im Data Warehouse durchgeführt werden. Das unterscheidet ETL von ELT – dazu aber später mehr. In beiden Fällen können Transformationen orchestriert werden – das heißt, arrangiert in einer Sequenz mit automatisierter Logik, um Sequenzierung, Timing und Fehler zu koordinieren. Idealerweise dienen Data Warehouses für das gesamte Unternehmen als Datensatz-Repository. Als Data Warehouse kann jede Art von relationaler Datenbank verwendet werden, Data Warehouses sind jedoch in der Regel spaltenorientiert – im Gegensatz zu Transaktions- oder Produktionsdatenbanken, die meist zeilenorientiert und daher für Datenanalyseabfragen weniger effizient sind.

Abschließend werden die Daten mithilfe eines **Business-Intelligence-Tools**. analysiert. Business-Intelligence-Tools zeigen in der Regel Trends, Verhältnisse und andere Erkenntnisse in Dashboards und periodischen Berichten an.

Die einzelnen Komponenten eines Data Stacks können lokal oder in der Cloud gehostet werden. Früher nutzten Unternehmen Data Stacks im eigenen Haus (On-Premise). Inzwischen nutzen viele die Cloud, einige bleiben im Interesse der Einhaltung gesetzlicher Vorschriften oder hochspezifischer Leistungsanforderungen beim On-Premise-Ansatz. Sie entwickeln intern zentrale Komponenten ihrer Dateninfrastruktur, um externe Abhängigkeiten oder die Bindung an einen Anbieter zu vermeiden.

TIPP: *Eine technische Erörterung des Unterschieds zwischen zeilen- und spaltenorientierten Datenbanken würde den Rahmen dieses Leitfadens sprengen. Hier jedoch eine kurze Einführung: Zeilenorientierte Datenbanken – auch als OLTP-Datenbanken (Online Transaction Processing) bezeichnet – werden in der Produktion in der Regel zur Abwicklung einzelner Transaktionen verwendet. Spaltenorientierte Datenbanken können die bei der Datenanalyse verwendeten Spaltenoperationen (MIN, MAX, SUM, COUNT, AVG) im Allgemeinen besser verarbeiten. Wenn Sie sich mit der Technik auskennen, könnten Sie versucht sein, einfach eine Kopie Ihrer Produktionsdatenbank zu erstellen und diese für Datenanalysen zu verwenden. Machen Sie das nicht! Verwenden Sie für die Datenanalyse stets eine spaltenorientierte Datenbank oder ein Data Warehouse. Das ist effizienter und spart Ihnen viel Zeit.*

Herausforderungen, die ein Data Stack lösen kann

AEin Data Stack muss bei der Weiterleitung von Daten von Konnektoren an Data Warehouses sicherstellen, dass die Daten in einer einzigen Umgebung zentral zusammengeführt werden und so aktuell und quellengetreu wie möglich bleiben. Der Prozess sollte kontinuierlich und mit minimalem Eingriff durch den Menschen vonstatten gehen.

Fragmentierung

Häufig kommen Daten in fragmentiertem Zustand von Apps, Tools und Datenbanken an. Es gibt zwei Arten von Fragmentierung: Die erste tritt ein, weil API-Endpunkte und operative Datenbanken nicht für Analyseabfragen ausgelegt sind. Das heißt, dass den von ihnen generierten Daten häufig der wichtige Kontext fehlt. Zudem sind sie nicht organisiert, was ihre Analyse erschwert. Es bedarf häufig einer umfassenden Datenmodellierung, um die Daten mit Sinn zu erfüllen.

Der zweite Fall tritt ein, weil die meisten Apps, Tools und Datenbanken nicht speziell auf die Interoperabilität mit Daten aus anderen Systemen ausgelegt sind. Der Aufwand für die Erstellung des nötigen Kontexts durch das Zusammenführen von Datensätzen über mehrere Quellen hinweg kann zu sehr langen Bearbeitungszeiten für Berichte führen.

Daher auch das Phänomen der „dunklen Daten", das auftritt, wenn ein großer Teil der von einem typischen Unternehmen erfassten Informationsgüter ungenutzt bleibt. Die Zentralisierung von Daten in einer einzigen Umgebung ermöglicht eine schnellere Erstellung von Berichten und versetzt Unternehmen in die Lage, Datensätze

zusammenzuführen und kohärente Berichte über ihre betrieblichen Vorgänge und Kunden zu erstellen. Der Immobilienmakler Zoopla kombinierte beispielsweise ERP- und CRM-Daten zu einem wöchentlichen Dashboard mit mehr als 40 einzelnen KPIs.

FALLSTUDIE: DiscoverOrg stellte Nutzung von OLTP für die Datenanalyse ein

DiscoverOrg ist eine B2B-Plattform für die Lead-Generierung, die Profile von Einzelpersonen und Unternehmen erstellt, auf deren Grundlage sich zielgerichtetere Vertriebs- und Marketingkampagnen durchführen lassen. Vor dem Umstieg auf einen Data Stack extrahierte DiscoverOrg seine Analysedaten aus einer Kopie seiner OLTP-Produktionsdatenbank. Daten aus Anwendungen von Drittanbietern wurden dabei ausgeklammert. Abfragen konnten bis zu 36 Stunden dauern – wenn sie das System nicht komplett zum Absturz brachten.

Die Einführung eines automatisierten Datenintegrationstools brachte DiscoverOrg folgende Vorteile: Das Unternehmen konnte seine Produktionsdaten mit Daten aus Quellen von Drittanbietern in einem Data Warehouse kombinieren, Berichte in wenigen Minuten statt Tagen erstellen und einen Lead-Routing-Algorithmus für das Gewinnen von Aufträgen mit einem um 80 bis 90 % höheren Durchschnittswert entwickeln. Zudem sparte man sich die Arbeit von zwei oder drei Data Engineers.

Seit Kurzem bettet DiscoverOrg zudem Datenanalyse-Dashboards in seine Plattform ein, wovon die Kunden von DiscoverOrg profitieren.[3]

Genauigkeit

Daten können auf zwei Arten ungenau sein: eine davon ist auf eine fehlerhafte Messung oder Aufzeichnung zurückzuführen, vor allem wenn die Daten von Hand eingegeben oder von nicht digitalen Medien transkribiert wurden. Umfragen und Formulare haben zwangsläufig Rechtschreibfehler, Buchstaben- und Zahlendreher sowie andere Schreibfehler zur Folge. Eine zweite – eher systemische – Fehlerquelle ist auf Berechnungen oder Transformationen von Rohdaten zurückzuführen. Es gibt viele Möglichkeiten, einen Datensatz zu verarbeiten, und jede Berechnung führt einen Schritt weiter von den ursprünglichen Werten weg. So kann es vorkommen, dass verschiedene Mitarbeiter und Teams innerhalb eines Unternehmens zu völlig unterschiedlichen Erkenntnissen gelangen.

Veraltete Daten

3 Die komplette Fallstudie finden Sie unter: fivetran.com/blog/case-study-discoverorg

Äußere Bedingungen ändern sich schnell. Wenn es Sie Wochen oder Monate kostet, einen Bericht zusammenzustellen, dann treffen Sie möglicherweise schwerwiegende Fehlentscheidungen, weil Sie mit veralteten Daten arbeiten. Wer sich mit Entscheidungsmodellen wie PDCA (Plan-Do-Check-Act) oder OODA (Observe-Orient-Decide-Act) auskennt, weiß, wie wichtig es ist, fundierte Entscheidungen schneller als die Konkurrenz zu treffen. Diese Modelle haben für alle wettbewerbsorientierten, dynamischen Umgebungen ihre Geltung, einschließlich Kriegsführung, Gaming, Leistungssport und natürlich die Wirtschaft.

FALLSTUDIE: Zoopla nutzt eine Datenpipeline zur Vereinheitlichung der Datenintegration

Zoopla ist ein britischer Online-Immobilienmarkt für den Kauf, Verkauf und die Vermietung von Wohn- und Gewerbeimmobilien.

Bevor Zoopla auf einen modernen Data Stack umstieg, erfolgten Datenanalysen äußerst fragmentiert. Analysten und Engineers schrieben auf Ad-hoc-Basis eine Reihe eigener Skripts für das Extrahieren und Analysieren der Unternehmensdaten. Diese Skripts wurden nicht dokumentiert und waren oft in verschiedenen Sprachen geschrieben. Zudem nutzten die Analysten im BI-Tool von Zoopla native Datenkonnektoren für quellenübergreifende Abfragen.

Das BI-Team von Zoopla erkannte, dass dieses System nicht nachhaltig ist und mit dem Wachstum des Unternehmens und dem Hinzukommen weiterer Datenquellen schnell an seine Grenzen stoßen würde, was die Quantifizierung der Fortschritte angeht. Nach Einführung eines modernen Data Stacks war Zoopla in der Lage, Daten aus seiner ERP- und CRM-Software in einem Dashboard zusammenzuführen, das im wöchentlichen Rhythmus automatisch aktualisiert wird und unternehmensübergreifend mehr als 40 verschiedene KPIs bietet. Diese KPIs werden kontinuierlich angezeigt und dienen der Geschäftsführung sowie den einfachen Mitarbeitern gleichermaßen als Grundlage für ihre Entscheidungen.[4]

Opportunitätskosten

Daten zu haben, die nicht dem Erkenntnisgewinn dienen, ergibt keinen Sinn. Früher verbrachten Analysten und Data Engineers jedoch weniger Zeit mit der Analyse von Daten. Sie waren zum Großteil der Zeit damit beschäftigt, komplexe Software für die Verarbeitung der Daten zu entwickeln und zu pflegen. Datenwissenschaft wird meist mit modernsten Prognosemodellen und maschinellem Lernen assoziiert. Der normale Datenwissenschaftler bringt jedoch etwa 80 % seiner Zeit damit zu, Daten

4 Die komplette Fallstudie finden Sie unter: fivetran.com/blog/case-study-zoopla

zu suchen und zu integrieren, statt sie zu analysieren.[5]

ACHTUNG: *Das Simpson-Paradoxon (Abbildung 2.0) bietet ein hervorragendes Beispiel dafür, wie dieselben Daten bei unterschiedlicher Transformation extrem unterschiedliche, ja mitunter gegensätzliche Schlussfolgerungen ergeben.*

Kurz gesagt beschreibt das Simpson-Paradoxon das Phänomen, dass sich je nach Aufteilung oder Kombination von Daten äußerst unterschiedliche Trends und Muster zeigen.

Ein ähnliches Konzept beschreibt das Anscombe-Quartett (Abbildung 2.1): vier verschiedene Datensätze mit identischem Mittel-, Varianz-, Korrelations- und R2-Wert. Das Simpson-Paradoxon und das Anscombe-Quartett führen uns eindrucksvoll vor Augen, dass es bestenfalls naiv und schlimmstenfalls äußerst irreführend ist, sich bei der Datenanalyse auf grundlegende zusammenfassende statistische Werte zu beschränken. Man muss sich schon die Mühe machen, die Daten zu visualisieren, und sich überlegen, wie diese kategorisiert werden und inwieweit versteckte Variablen die Antwort erschweren können.

Der traditionelle Ansatz der Datenintegration: ETL

Der traditionelle Ansatz der Datenintegration – ETL (Extrahieren, Transformieren, Laden) – war ab den 1970ern der bestimmende Ansatz. Unter etablierten Unternehmen gilt ETL als Branchenstandard. Häufig werden mit diesem Akronym umgangssprachlich alle Datenintegrationsaktivitäten beschrieben. ETL entstand zu einer Zeit, als Rechenleistung, Speicherplatz und Bandbreite noch rar und teuer waren. Die aus der Not heraus geborenen technischen Defizite von ETL muten in der heutigen Ära der Cloud-Technologie zunehmend anachronistisch an.

5 infoworld.com/article/3228245/the-80-20-data-science-dilemma.html

Abbildung 2.0
Simpson-Paradoxon

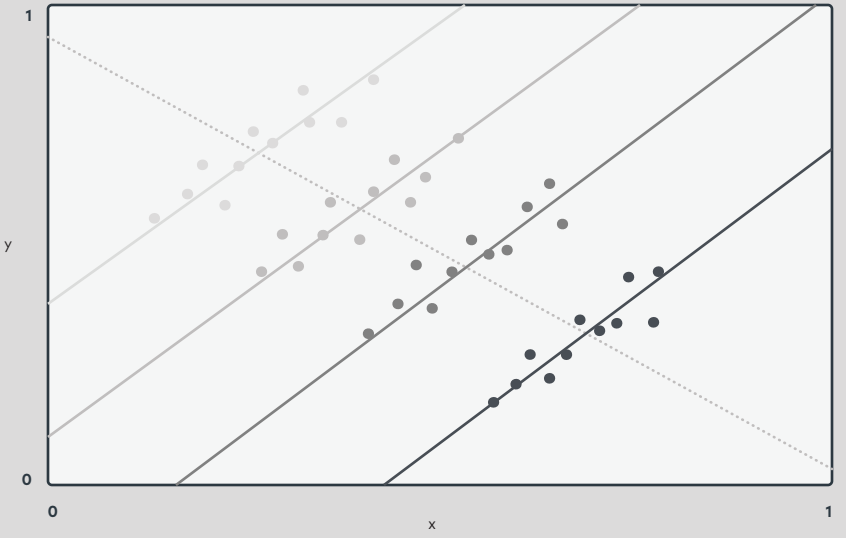
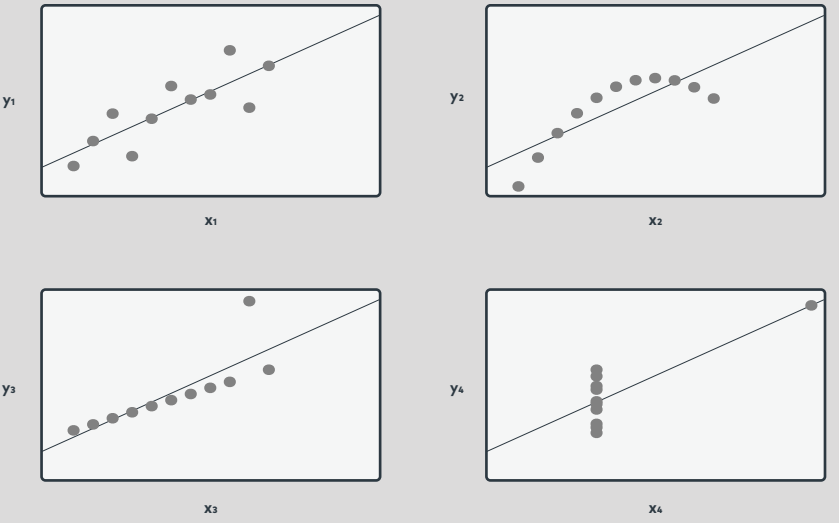


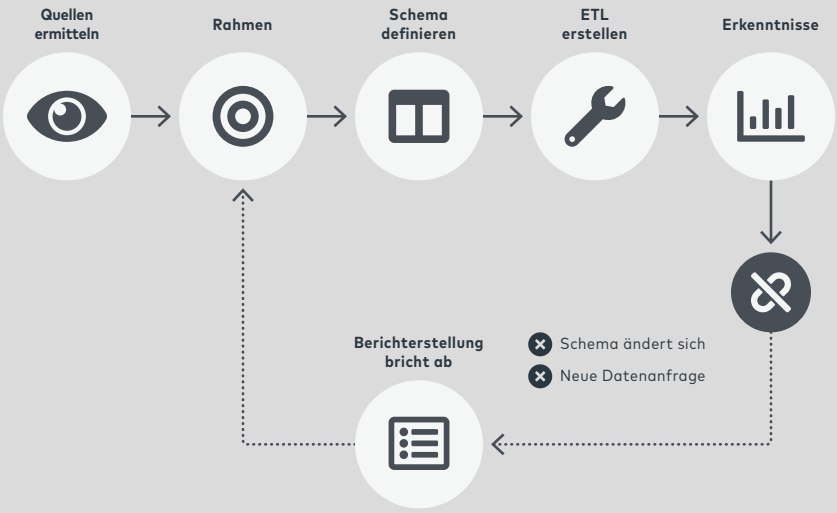
Abbildung 2.1
Anscombe-Quartett



ETL-Workflow

Der Workflow, den Data Engineers und Analysten ausführen müssen, um eine ETL-Pipeline zu erstellen, sieht wie folgt aus:

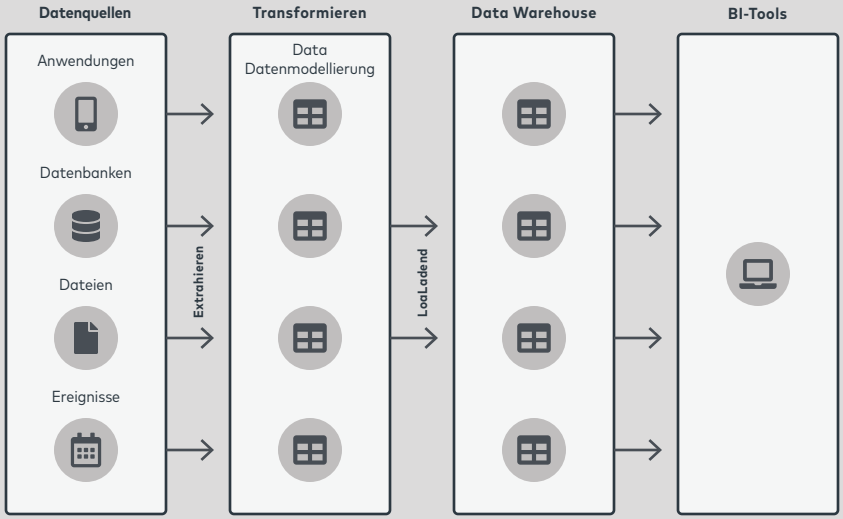
Abbildung 2.2
Workflow der Datenintegration und -analyse



1. **Quellen ermitteln** – Anwendungen, Ereignis-Tracker oder Datenbanken
2. **Rahmen** – Rahmen und Geschäftsziele des Berichts ermitteln
3. **Schemen definieren** – Daten modellieren und nötige Transformationen ermitteln
4. **ETL erstellen** – die Software schreiben, die Details der aufzurufenden API-Endpunkte, die Art der Normalisierung der Daten und das Laden in den Zielort definieren
5. **Erkenntnisse im Überblick** – Berichte erstellen, die für die wichtigen Entscheidungsträger verständlich sind
6. **Berichterstellung bricht ab** – bei Abbrüchen erhalten die Endbenutzer keine zeitgerechten Daten und es kommt zu Ausfällen infolge von:
    a. vorgelagerten Schemaänderungen
    b. neuen Datenanfragen, weil sich die Erfordernisse an die Datenanalyse ändern
7. **Rahmen des Projekts neu definieren**

Das ETL-System führt folgende Schritte aus:

Abbildung 2.3
ETL



1. **Extrahieren** – Daten werden von Konnektoren extrahiert.
2. **Transformieren** – Mittels einer Reihe von Transformationen werden die Daten in Modellen so neu strukturiert, wie sie von Analysten und Endbenutzern benötigt werden.
3. **Laden** – Daten werden in ein Data Warehouse geladen.
4. **Visualisieren** – Die Daten werden von einem Business-Intelligence-Tool zusammengefasst und visuell aufbereitet.

Die Orchestrierung und die Transformation vor dem Laden stellen eine kritische Sicherheitslücke im ETL-Prozess dar. Transformationen müssen speziell auf die einzigartigen Konfigurationen der Ursprungs- und auch der Zieldaten zugeschnitten sein. Das bedeutet, dass vorgelagerte Änderungen an Datenschemen sowie nachgelagerte Änderungen an Geschäftsanforderungen und Datenmodellen die Software beschädigen können, die die Transformationen durchführt.

Weil ETL nicht direkt Daten aus jeder Quelle in das Data Warehouse repliziert, gibt es in keiner Phase des Prozesses ein umfassendes Datensatz-Repository für die Datenanalyse. Fehler in einer der Phasen des Prozesses machen die Daten für Analysten unzugänglich und müssen mit einigem technischen Aufwand behoben werden.

Beschränkungen von ETL

Der traditionelle ETL-Prozess hat insgesamt drei große inhärente Nachteile:

1. **Komplexität**. Datenpipelines werden mit benutzerdefiniertem Code ausgeführt, der von den spezifischen Anforderungen bestimmter Transformationen bestimmt wird. Das heißt, dass das Data Engineering-Team hochspezialisierte, mitunter nicht übertragbare Kompetenzen bezüglich der Verwaltung seiner Codebasis entwickelt.
2. **Fragilität**. Aus den oben genannten Gründen macht eine Kombination aus Fragilität und Komplexität schnelle Anpassungen kostspielig oder unmöglich. Teile der Codebasis können mit geringer Vorwarnzeit ausfallen, und neue Geschäftsanforderungen und Anwendungsfälle erfordern eine umfangreiche Überarbeitung des Codes.
3. **Unzugänglichkeit**. Wichtiger noch: Für kleinere Unternehmen ohne dedizierte Data Engineers ist ETL so gut wie nicht zugänglich. On-Premise-ETL verursacht weitere Infrastrukturkosten. Kleinere Unternehmen könnten gezwungen sein, Daten stichprobenartig zu analysieren oder manuelle Ad-hoc-Berichte zu erstellen.

Die Entstehung von Cloud-Technologie

Auch wer Techniktrends nur am Rande verfolgt, hat bemerkt, dass Rechenleistung, Speicherung und Bandbreite inzwischen kein kostbares und seltenes Gut mehr sind. Mit fortschreitender Entwicklung der Datenverarbeitung sanken die Rechenkosten (Abbildung 2.4).

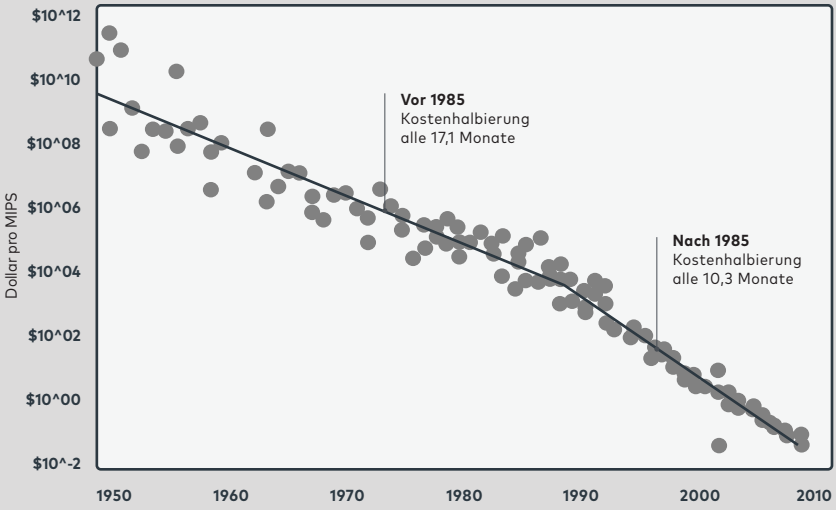
Innerhalb von etwa 35 Jahren sanken die Kosten für ein Gigabyte von fast 1 Million US-Dollar auf einen Cent-Betrag (Abbildung 2.5).

Eine Folge dieses radikalen Kostenrückgangs ist, dass Data Warehouses viel größere Datenmengen aufnehmen können. Unternehmen müssen heute nicht mehr viele Quelldaten voraggregieren und im Laufe des Prozesses einen Großteil dieser Daten wieder verwerfen. Dadurch lassen sich tiefgreifendere und umfassendere Analysen durchführen als je zuvor.

Das World Wide Web gibt es zwar erst seit 1991, dennoch sanken auch die Kosten für die Datenübertragung über das Internet drastisch: in nicht einmal zwanzig Jahren von circa 1.200 USD pro Mbit/s auf wenige Cent (Abbildung 2.6).

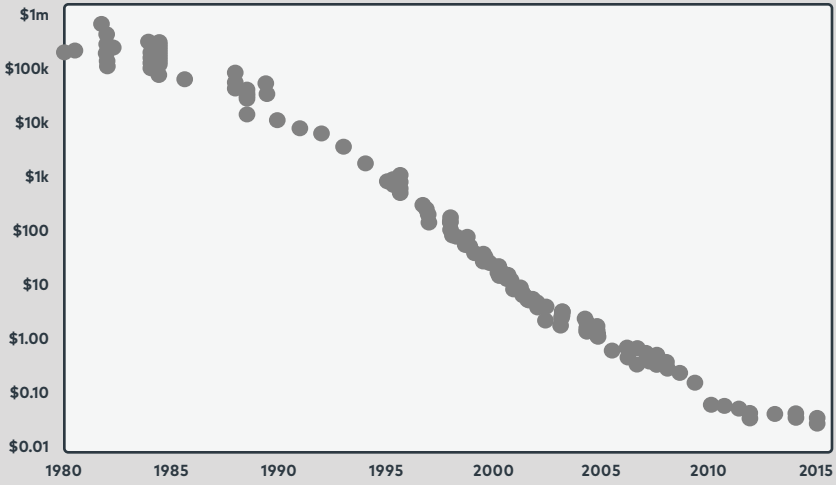
Diese drei rückläufigen Kostenentwicklungen mündeten letztlich in der Entstehung der Cloud – die Nutzung standortferner, dezentraler, webfähiger Rechenressourcen. Die Cloud-Technologie brachte ihrerseits eine Vielzahl Cloud-nativer Anwendungen und Dienste hervor.

Abbildung 2.4
Rechenkosten



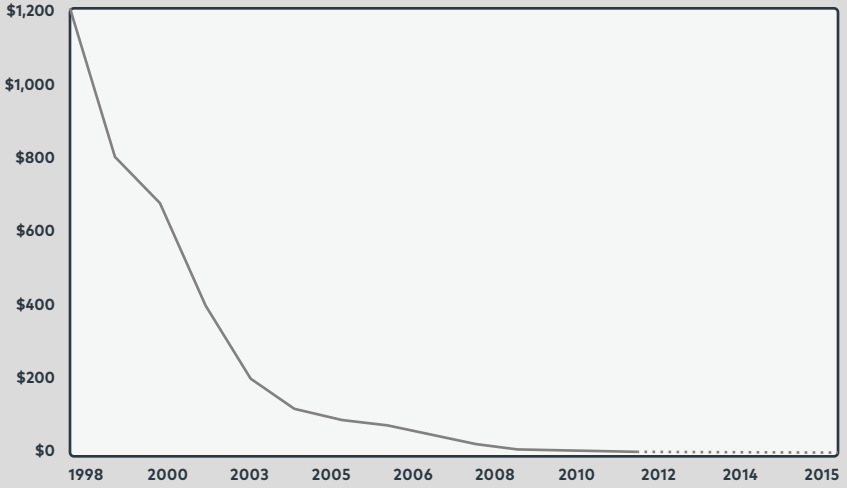
Quelle: https://frc.ri.cmu.edu/~hpm/book97/ch3/processor.list.txt

Abbildung 2.5
Festplattenkosten pro Gigabyte



Quelle: mkomo.com

Abbildung 2.6
Kosten von Internetverkehr (Bandbreite)



Quelle: dr1peering.net

TIPP: *Cloud-native Apps und Dienste decken geschäftliche Aktivitäten jeglicher Art ab: Kundenbeziehungsmanagement, Rechnungswesen und Bezahlung, E-Commerce, E-Mail-Marketing, Verwaltung von Arbeitgeberleistungen, Projektmanagement, Kundendienst und Einiges mehr. Die Chance ist hoch, dass Ihr Unternehmen bereits mehrere solcher Dienste nutzt.*

Einer der großen Vorteile der Cloud besteht darin, dass Analysten und Endbenutzer von Daten nicht mehr an physische Infrastruktur gebunden sind. Stattdessen können sie Dienste im Web hosten. Fragen der Skalierung und Zugänglichkeit lassen sich so viel einfacher lösen. Unternehmen können die Rechen- und Speicherressourcen im laufenden Betrieb erweitern oder verkleinern, und Benutzer können über jedes webfähige Gerät auf Dashboards und Berichte zugreifen.

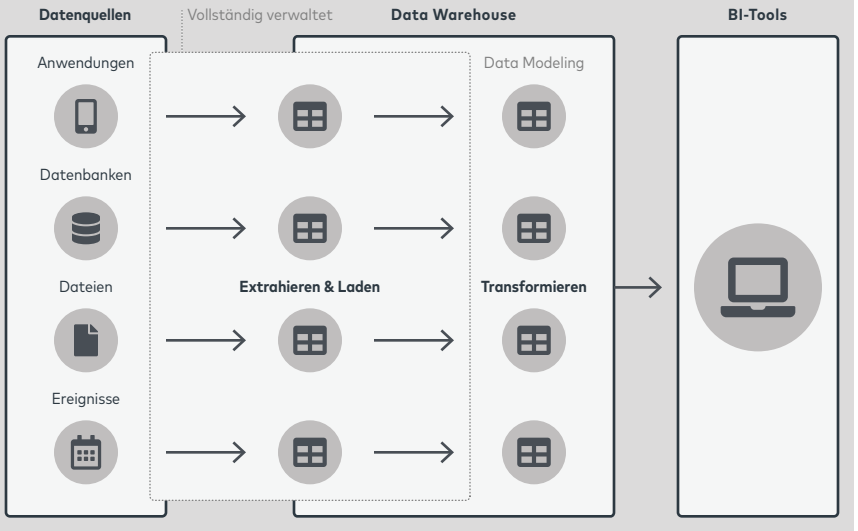
ACHTUNG: *Wir sollten die Fähigkeiten des modernen Web aber auch nicht überbewerten. Wir verfügen (noch) nicht über die Technologie, um Terabytes an Daten auf der ganzen Welt einfach hoch- und herunterzuladen. In sehr großem Maßstab betrachtet kann der physische Versand eines Containers mit Festplatten schneller gehen als das Verschicken derselben Datenmenge über das Internet. Nach wie vor müssen Daten komprimiert werden, und die von der Quelle zum Ziel übertragene Menge muss minimiert werden.*

Der moderne Ansatz der Datenintegration: ELT

Dieselben Entwicklungen, die schon das Wachstum der Cloud ermöglichten – die rapide sinkenden Kosten für Rechenleistung, Speicherplatz und Bandbreite – ermöglichen es Unternehmen, die mit ETL einhergehenden Probleme zu umgehen. Insbesondere können die Daten dadurch vor ihrer Transformation gestreamt und geladen werden. Diese umgekehrte Reihenfolge wird als ELT (Extrahieren, Laden, Transformieren) bezeichnet und gilt als moderner Nachfolger von ETL (Abbildung 2.7).

Was wir als den „modernen Data Stack" bezeichnen, basiert auf ELT. An die Stelle von On-Premise-Technologien treten dabei Cloud-native SaaS-Technologien. Dieses Setup erleichtert die Automatisierung, Zusammenarbeit und Skalierbarkeit. Viele der Kosten eines On-Premise-Stacks lassen sich damit vermeiden. Bei sachgemäßer Implementierung bietet der moderne Data Stack eine kontinuierliche Datenintegration und organisationsweite Zugänglichkeit – bei einem Minimum an manuellen Eingriffen und proprietärem Code.

Abbildung 2.7
ELT



Durch die Umkehr der Reihenfolge der Lade- und der Transformationsphase können drei der größten Nachteile von ETL beseitigt werden:

1. **Komplexität**. Die Pipeline wird vereinfacht: Weil zunächst Standardschemen ohne individuelle Transformationen an ein Warehouse geliefert werden, wandert ein Großteil der Pipeline-bezogenen Arbeit von den Data Engineers zu den Analysten.
2. **Fragilität**. Die Pipeline ist weniger fragil und risikobehaftet – weil Transformationen nach der Speicherung der Daten im Warehouse angewendet werden, kommen Brüche durch Änderungen in den Quellsystemen vorrangig auf der Datenanalyseschicht zum Tragen. Analysten können diese Probleme in der Regel ohne Hilfe von Data Engineers lösen.
3. **Zugänglichkeit**. Die Pipeline ist einfacher zugänglich, weil ihre Pflege weniger arbeitsintensiv ist. Da die Pipeline viel einfacher und dadurch weniger fragil ist, können Drittanbieter ein standardisiertes Tool für mehrere Kunden sowie daraus abgeleitete Produkte entwickeln, um die Analysearbeit zu optimieren. Durch den Kauf eines solchen standardisierten Tools lassen sich die Extraktions- und Ladephase zum Großteil auslagern und automatisieren.

Transformationen innerhalb des Warehouse ermöglichen die Erstellung abgeleiteter Tabellen, die als „Ansichten" bezeichnet werden, ohne dass dazu die Quelldaten geändert werden. Auf diese Weise können Unternehmen ein Datensatz-Repository erstellen, das gegen sich ändernde Geschäftsanforderungen oder vorgelagerte Schemaänderungen immun ist. Dieselben Daten können für mehrere Anwendungsfälle genutzt werden.

Zudem verringert ELT den Arbeitsaufwand der Engineers. Sobald sich die Daten im Warehouse befinden, können Analysten mittels SQL beliebige Transformationen durchführen. Es mögen zwar nach wie vor komplexe Transformationen nötig sein, die sorgfältig koordiniert und geplant werden müssen, aber Unterbrechungen und Fehler wirken sich nicht mehr auf die gesamte Datenpipeline aus oder binden erhebliche Engineering-Ressourcen.

Der nächste Quantensprung: automatisiertes ELT

Durch seine vereinfachte, Cloud-basierte Auslegung eignet sich der ELT-Data Stack gut für die Automatisierung und das Outsourcing.

Zu den spezifischen Aktivitäten beim automatisierten ELT gehören das Erkennen und Replizieren von Datenänderungen, das leichte Bereinigen und Normalisieren von Daten sowie das Aktualisieren und Erstellen von Tabellen. Diese Aktivitäten erfordern umfassende Kenntnisse über Datenquellen, die Datenmodellierung und

Datenanalyse sowie das erforderliche technische Wissen darüber, wie man robuste Softwaresysteme entwickelt. Ohne ein automatisiertes Datenintegrationstool muss Ihr Team diese Aktivitäten ausführen und die erforderlichen Funktionen entwickeln.

Wenn Sie ELT automatisieren und auslagern, können Sie hingegen auf das Fachwissen externer Anbieter zurückgreifen, die alle Eigenarten der zugrunde liegenden Datenquellen bestens kennen – und deren Konnektoren auf viel mehr Ausnahmefälle getestet haben, als Sie es wahrscheinlich jemals tun werden. Die größten Vorteile von automatisiertem ELT sind – wie bei fast jeder Automatisierung – die Einsparung von Zeit, Aufwand und Kosten. Statt sich mit routinemäßigen, vorgelagerten Arbeiten zu befassen, bei denen es um bereits erkannte und gelöste Probleme geht, kann sich Ihr Daten- oder Business-Intelligence-Team darauf konzentrieren, Erkenntnisse als Handlungsgrundlage zu liefern.

Data Engineers können die durch Automatisierung von ELT gesparte Zeit nutzen, um sich Problemen zu widmen, die externe Kunden betreffen, oder sich mit anspruchsvolleren Aktivitäten wie maschinellem Lernen und künstlicher Intelligenz zu befassen. Automatisiertes ELT stellt man sich am besten als „Kraft-Multiplikator" statt als Ersatz für menschliche Fähigkeiten vor.

Welche radikale Zugänglichkeit automatisiertes ELT bietet, wird im typischen Workflow der automatisierten Datenintegration veranschaulicht:

1. **Anmelden** – Sie aktivieren Ihr Konto.
2. **Auswählen** – Sie wählen Ihre Quellen und Ihr Data Warehouse aus.
3. **Authentifizieren** – Sie aktivieren Ihre Konnektoren mit Ihren bestehenden Anmeldedaten.
4. **Automatisieren** – Das System kümmert sich um die Synchronisierung der Bestandsdaten und laufende Änderungen.

Eine vollständige Synchronisierung der Bestandsdaten kann zwar je nach der Datenmenge, die von der Datenquelle gehostet wird, Stunden oder gar Tage dauern, für die Schritte, bei denen tatsächlich menschliches Eingreifen erforderlich ist, werden jedoch höchstens einige Minuten benötigt.

Ein automatisiertes Datenintegrationstool bietet zudem den Vorteil, dass jedes Unternehmen, das eine bestimmte Datenquelle verwendet, genau dasselbe Problem lösen muss, und der Anbieter des Tools jedem Kunden eine standardisierte Lösung mit genau denselben Schemen anbieten kann. Diese standardisierten Schemen ermöglichen die Entwicklung abgeleiteter Produkte zur Unterstützung der Datenanalyse. Benutzer desselben Integrationstools haben Zugriff auf dieselben SQL-basierten Transformationen, eingebetteten Datenanalyseprodukte und BI-

Tool-Module. Dank ELT profitiert die Datenanalyse von den Vorteilen modularer, austauschbarer Komponenten und ihrem möglichen Einsatz im großen Maßstab.

Und nicht zuletzt bietet das automatisierte ELT den Vorteil, dass Sie die Cybersicherheit und die Einhaltung gesetzlicher Vorschriften auslagern können. Die Richtlinien, Verfahren und Technologien, durch die der böswillige oder gesetzwidrige Zugriff auf Ihre Daten verhindert wird, erfordern tiefgreifendes Fachwissen, und Sie sind besser damit beraten, dies einem vertrauenswürdigen externen Anbieter zu überlassen, als zu versuchen, eine eigene Lösung zu entwickeln.

Durch die Vielzahl an neuen Möglichkeiten, die Engineers, Analysten und Endbenutzern durch Automatisierung und Self-Service erhalten, steigt jedoch auch die Bedeutung der Data Governance. Zugriff und Transparenz können äußerst wertvoll sein, müssen jedoch durch strenge Prüfung, Dokumentation und Zuweisung von Berechtigungen gesteuert werden. Im Verlauf der Entwicklung eines Unternehmens sehen sich Analysten, die normalerweise Dashboards und Berichte erstellen, u. U. damit konfrontiert, dass sich ihre Aufgaben in Richtung Data Governance verschieben, während die Berichte und Dashboards zunehmend von den Endbenutzern selbst erstellt werden.

FALLSTUDIE: DocuSign nutzt automatisierte Datenintegration zur Verdreifachung der Anzahl von Datenquellen

DocuSign ist weltweit führend bei Technologie für elektronische Unterschriften. Die angebotenen Leistungen umfassen das automatische Erstellen, Unterschreiben, Umsetzen und Verwalten von Verträgen auf individueller und gewerblicher Basis.

Früher verwendete DocuSign SQL Server als Data Warehouse für sechs Datenquellen, die von einem Engineer verwaltet wurden. Die Entwicklung dieser

Konnektoren in Eigenregie dauerte drei bis sechs Monate und ihre Wartung bis zu 20 Stunden pro Woche. Dieser Aufwand war bei zunehmendem Wachstum des Unternehmens nicht mehr zu leisten, vor allem auch, weil die Spezialisten für wichtige Projekte benötigt wurden und die Geschäftsteams Daten aus Anwendungen modellieren und katalogisieren mussten.

Mithilfe einer automatisierten Datenintegrationslösung und eines flexibleren Cloud-Data-Warehouse konnte DocuSign die kompletten 20 Stunden Engineering-Zeit einsparen und die Anzahl der genutzten Datenquellen von sechs auf 18 verdreifachen. Diese Ausweitung sowie die Zeit- und Aufwandsersparnis schlugen sich in einer weiteren äußerst positiven Entwicklung nieder: Mitarbeiter aller Teams im gesamten Unternehmen nutzen jetzt über 100 aktive Dashboards in ihrem BI-Tool.[6]

MERKE: *Die Cloud-Technologie hat eine Fülle wertvoller Daten und darüber hinaus die Tools hervorbracht, die für ihre adäquate Handhabung benötigt werden. ELT beseitigt viele der Nachteile von ETL und macht Daten und Analysen deutlich besser zugänglich und skalierbar als je zuvor.*

6 Die komplette Fallstudie finden Sie unter: fivetran.com/blog/case-study-docusign

Kapitel 3: Warum Sie keine eigene Datenpipeline aufbauen sollten

INHALT DIESES KAPITELS:

- Schätzung der monetären und nicht-monetären Kosten der Entwicklung einer eigenen Datenpipeline
- Überzeugen Ihres Unternehmens von der Einführung einer handelsüblichen Lösung

Wichtige Überlegungen

Wenn der moderne Data Stack die Datenintegration radikal vereinfacht, lohnt es sich dann für Ihr Unternehmen, auch in der Cloud eine eigene ELT-Pipeline zu entwickeln?

ACHTUNG: *Rufen Sie sich in Erinnerung, was in Kapitel 2 erläutert wurde: Die manuelle Datenintegration ist nicht skalierbar und ETL ist aus technischer Sicht inzwischen überholt. Hier besprechen wir hauptsächlich den Aufbau einer maßgeschneiderten ELT-Pipeline. Die nachstehenden Argumente gelten jedoch auch, wenn ein Unternehmen versucht, einen benutzerdefinierten ETL-Workflow zu entwickeln.*

Zeitlicher und monetärer Aufwand

Wie bereits in Kapitel 2 erläutert, verbringen Datenanalysten durchschnittlich 80% ihrer

Zeit damit, Datenpipelines zu entwickeln – eine Aufgabe, für die den meisten Datenwissenschaftlern die Fähigkeiten, das Interesse oder die entsprechende Ausbildung fehlen dürfte (Abbildung 3.0 und 3.1). Was jedoch am stärksten gegen die Entwicklung und Pflege einer eigenen ELT-Pipeline spricht, sind die damit einhergehenden Kosten in Bezug auf Zeit, Geld, Moral und vertane Chancen.

Angenommen, Ihre Organisation benötigt fünf Konnektoren für das Kundenbeziehungsmanagement, das Ticketing für den Kundensupport, die Automatisierung von Werbung, das Projektmanagement und die Abrechnung von Abonnements.

Für jeden der fünf Konnektoren benötigt ein Engineer rund fünf Wochen; das sind fünf Mannwochen (MW):

```
(5 Konnektoren) * (5 MW)
```

Für jeden Konnektor wird wahrscheinlich eine Woche pro Quartal für Wartungsarbeiten benötigt. Das sind weitere vier Wochen pro Jahr:

```
(5 Konnektoren) * (5 MW + 4 MW)
(5 Konnektoren) * (9 MW) = 45 MW
```

Das sind 45 von 52 Wochen im Jahr. Geht man von einer etwas großzügigen Urlaubs- oder Krankheitstageregelung aus, ist das rund ein Jahr Arbeit für einen Software-Engineer, der rund 120.000 US-Dollar plus Vorsorgeleistungen kostet (Abbildung 3.2)

In den Folgejahren wird Ihr Engineer die Konnektoren jedes Quartals (vier Wochen) aktualisieren sowie auftretende Fehler und Grenzfälle behandeln (eine Woche). Dies entspricht insgesamt fünf MW pro Konnektor.

```
(5 Konnektoren) * (5 MW) = 25 MW
```

Das sind 25 von 52 Wochen im Jahr für die laufende Wartung. Legen wir diese Kosten grob überschlagen auf die Hälfte des Jahresgehalts des Engineers bzw. 60.000 US-Dollar fest.

Der Kauf oder das Outsourcing von fünf Konnektoren wird wahrscheinlich deutlich unter den beiden oben genannten Beträgen liegen. Diese Kosten wachsen natürlich direkt proportional zur Anzahl der von Ihnen verwendeten Datenquellen.

Abbildung 3.0
Womit verbringen Datenwissenschaftler die meiste Zeit?

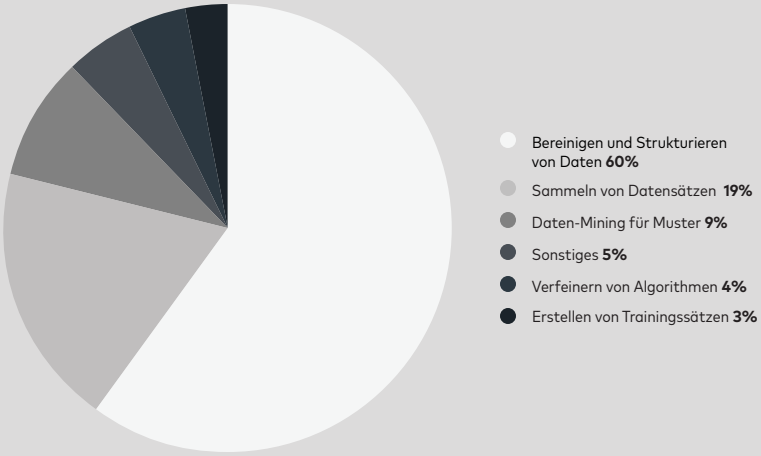
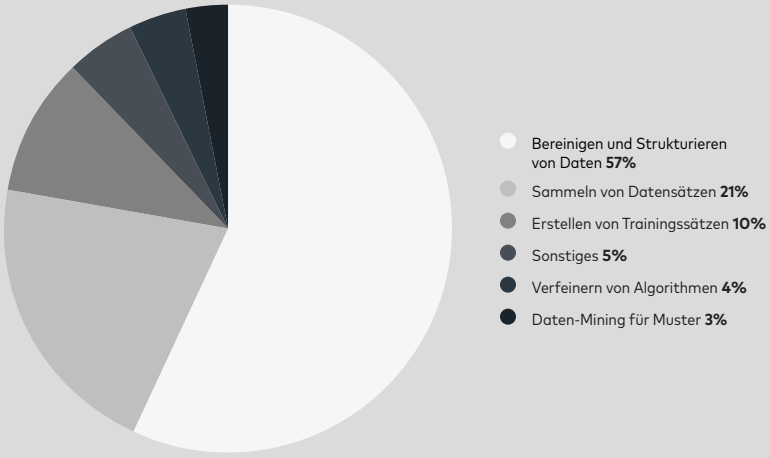
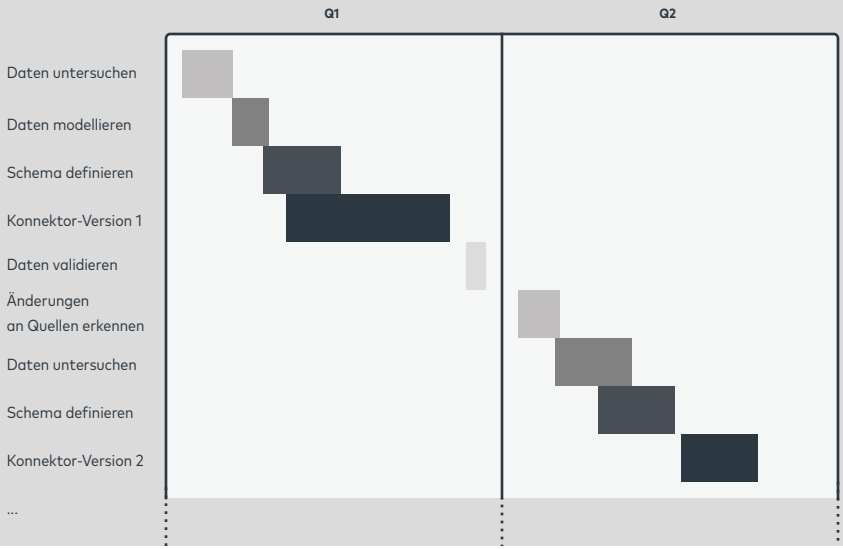


Abbildung 3.1
Was ist der unerfreulichste Teil der Datenwissenschaft?



Quelle: https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/

Abbildung 3.2
Wie viel Aufwand fließt in die Erstellung eines eigenen Konnektors?



Das obige Gantt-Beispieldiagramm veranschaulicht das zyklische, repetitive Wesen der Data-Engineering-Arbeit bei fortdauernder Änderung vorgelagerter Schemen – selbst in einem ELT-Rahmen.

Moral

Wenn Sie Ihre Analysten, Engineers und Manager bei Laune halten möchten, sollten Sie die folgenden Probleme beim Erstellen eigener Konnektoren oder beim manuellen Erstellen von Berichten berücksichtigen:

1. Ablenkung von anderen Aufgaben im Bereich Software Engineering, Datenwissenschaft oder Datenanalyse – ein sehr häufiges Ärgernis bei neuen Datenwissenschaftlern in unterbesetzten Unternehmen, was zu hoher Personalfluktuation führen kann
2. Frust und Überarbeitung – ausgelöst durch die schwierige Aufrechterhaltung der Datenintegrität, insbesondere bei Personen, die nicht entsprechend geschult sind
3. Ausfallzeiten bedingt durch die ständig zunehmende Komplexität, weil zwangsläufig zusätzliche Datenquellen hinzukommen
4. ive Fehlgeleitete Entscheidungen bedingt durch Verzögerungen zwischen der Anforderung von BI-Daten und der Bereitstellung entsprechender Erkenntnisse – die zum Zeitpunkt ihres Eintreffens möglicherweise veraltet sind

Für die meisten Datenprofis ist die Datenbankpflege eher lästige Pflicht als gern gemachte Arbeit.

Lernkurven

Die oben veranschlagten fünf Wochen gelten für APIs, die relativ einfach sind. Das trifft aber nicht auf alle APIs zu. Bei manchen wurden Best Practices ignoriert, andere sind schlecht dokumentiert und wieder andere sind einfach nur sehr komplex.

Daten aus einem ERP-Tool (Enterprise Resource Planning) können beispielsweise alle möglichen Geschäftsaktivitäten umfassen – abgebildet in Dutzenden oder Hunderten von Einzeltabellen mit komplexen Zusammenhängen. Es kann viele Iterationen erfordern, um eine ausgereifte Software für eine solche Datenquelle zu entwickeln. Dadurch vervielfachen sich die oben genannten Kosten.

Komplexität in großem Maßstab

Es ist sehr unwahrscheinlich, dass der Datenbedarf Ihres Unternehmens bei fünf Konnektoren endet. Wie eingangs erwähnt, nutzt ein typisches Unternehmen heute mehr als 100 Anwendunge.[7] Und diese Zahl wird eher noch steigen. Es ist schwer zu rechtfertigen, Ihrem Team mehr Arbeit aufzubürden, wenn sich das Pipeline-Engineering kostengünstig auslagern lässt.

Standardisierung

TDas letzte Argument gegen die Entwicklung einer eigenen Datenpipeline ist, dass die von einem externen Anbieter entwickelten Konnektoren durch Tests hinsichtlich Dutzenden von Sonderfällen bei vielen Kunden robust werden. Diese Konnektoren erzeugen standardisierte Datensätze mit standardisierten Schemen (Abbildung 3.3).

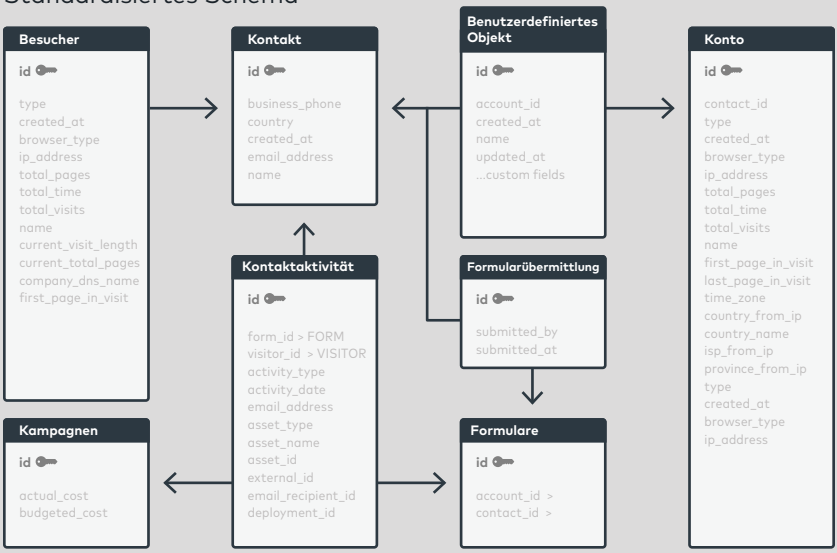
In der Praxis ist es eher unwahrscheinlich, dass Sie normalisierte Tabellen verwenden, um Ihre Dashboards direkt mit Daten zu versorgen. Wahrscheinlich werden Sie sie in Modelle umwandeln, die sich für ihre Endbenutzer besser eignen.

Im Prinzip ermöglicht dies auch jedem Unternehmen, das die gleichen Konnektoren verwendet, die gleichen Transformationen zu nutzen, weil die Daten alle identisch strukturiert sind. Plug-and-Play-Rezepte oder -Vorlagen dieser Art können in SQL oder in Sprachen geschrieben werden, die für eine BI-Plattform wie beispielsweise LookML spezifisch sind.

7 https://www.wsj.com/articles/employees-are-accessing-more-and-more-business-apps-study-finds-11549580017

Normalerweise entstehen durch Standardisierung Skaleneffekte, weil der Anbieter jede Eigenart der zugrunde liegenden Datenquelle kennt und diesen Vorteil an den Kunden weitergibt.

Abbildung 3.3
Standardisiertes Schema



TIPP: *Es gibt so viele Möglichkeiten, aus einem Datensatz ein Schema zu erstellen, wie es Meinungen und Anwendungsfälle gibt. Eine Methode, wie sich ein replizierbarer, leicht verständlicher Standard durchsetzen lässt, ist die Normalisierung. Eine vollständige Erörterung der Normalformen würde den Rahmen dieses Handbuchs sprengen. Die Kernidee ist jedoch die, dass durch die Normalisierung eine relationale Struktur entsteht, die Redundanz und uneinheitliche Beziehungen zwischen Datenelementen beseitigt.*

MERKE: *Das Konzept des Extrahierens und Ladens von Daten mutet simpel an, Datenpipelines sind jedoch hochentwickelte Technologien, deren erfolgreiche Entwicklung viel Geschick, Arbeit und Liebe zum Detail voraussetzt. Mit ziemlicher Sicherheit ist es besser, eine Lösung einzukaufen, als zu versuchen, eine eigene zu entwickeln und zu pflegen.*

Argumente für den Kauf

Wenn Sie den Kauf eines Datenpipeline-Tools vorschlagen, könnten Sie in Ihrem Unternehmen auf Widerstand stoßen. Data Engineers empfinden es u. U. als bedrohlich, dass ein Teil ihrer Arbeit automatisiert werden könnte, und Führungskräfte mit geringem Bezug zur Thematik erkennen die Vorteile möglicherweise nicht sofort.

FALLSTUDIE: Papier nutzt ELT zur Verdoppelung seiner Datenquellen und zur Entwicklung eines Kundenattributionsmodells

Papier ist ein im Bereich Design und Personalisierung tätiges Unternehmen, das Schreibwaren, Einladungen, Karten und Fotobücher verkauft. Das 2015 gegründete Unternehmen war in seinem Mutterland Großbritannien recht erfolgreich und ist derzeit auf Expansionskurs – sowohl bezüglich seiner Geschäftstätigkeit als auch seines Produktportfolios. Die Komplexität der Datenintegration wuchs mit der Ausweitung des Geschäfts.

Ursprünglich zentralisierte Papier seine Anzeigen-, Clickstream- und Transaktionsdaten mithilfe selbst entwickelter ETL-Skripte und -Tools. Dieser Ansatz stieß schon schnell an seine Grenzen – in Bezug auf die reine Arbeitslast und den Korrekturbedarf bei Ungenauigkeiten und Inkonsistenzen, die häufig eine erneute Synchronisierung erforderten.

Nach Einführung einer ELT-Datenpipeline konnte Papier seinen Workflow für die Datenerfassung automatisieren und die Anzahl der für die Analyse verwendeten Datenquellen verdoppeln. Synchronisierungen, die früher einmal am Tag stattfanden, erfolgen jetzt stündlich. Wichtiger noch: Durch Zentralisierung der Daten und Verdoppelung der Anzahl der Datenquellen konnte Papier ein robusteres Kundenattributionsmodell entwickeln. Damit ermittelt Papier, welche Anzeigen und anderen Marketingaktivitäten den stärksten Umsatzeffekt haben.[8]

Überzeugen der Data Engineers

Data Engineers bevorzugen mitunter maßgeschneiderte Anpassungen, eine differenzierte Konfigurierbarkeit und Kontrolle über die Zugänglichkeit. Mit folgenden Argumenten bringen Sie sie auf Ihre Seite:

8 Die komplette Fallstudie finden Sie unter: fivetran.com/blog/case-study-papier

Pipelines zu entwickeln, ist Knochenarbeit

1. Komplexität und lange Bearbeitungszeiten führen zu ärgerlichen Engpässen.
2. Ohne die Fähigkeit, schnell fundierte Entscheidungen zu treffen, sind Unternehmen nicht anpassungsfähig und geraten gegenüber der Konkurrenz ins Hintertreffen.
3. Reine Fleiß- und Routinearbeit macht keinen Spaß.

Die Vorteile des Outsourcings

1. Sie bilden keinen Engpass mehr.
2. Externe Anbieter, die sich auf Datenkonnektoren spezialisiert haben, sind sachkundiger und erfahrener, was die Lösung dieser Probleme angeht. Sie haben wahrscheinlich weit mehr Sonderfälle untersucht als Sie und erledigen die Aufgabe besser.
3. LT ist ein „Kraft-Multiplikator". Sie können jetzt mit viel geringerem Aufwand eine viel größere Auswahl an Konnektoren verwalten. Das macht es einfacher, die Erwartungen Ihres Unternehmens zu erfüllen.
4. Vollständig verwaltetes ELT ist unglaublich intuitiv und erfordert fast keine Schulung oder Anleitung. So haben Sie Zeit, andere Dinge zu tun, als hochgradig Tool-spezifische Fähigkeiten zu erwerben.
5. You won't have to keep hiring and managing more people for data engineering.
6. Sie haben die Möglichkeit, strategische Aktivitäten zu verfolgen, z. B.:
    a. maßgeschneiderte Tools und Infrastruktur für Analysten
    b. Infrastruktur zur Unterstützung von KI und maschinellem Lernen
    c. neue Softwareprodukte

Dieses Problem lösen Sie am besten, indem Sie Ihren Engineers Aufgaben geben, die sie ausfüllen. In der Praxis haben nur sehr wenige Engineers Interesse daran, den Großteil ihrer Zeit mit dem Schreiben von Datenkonnektoren zu verbringen. Statt mit dem Aufbau und der Wartung von Infrastruktur befassen Sie sich lieber mit anspruchsvolleren Dingen.

Überzeugen Ihres Chefs

Führungskräfte, die nicht so nahe an der Thematik dran sind, müssen möglicherweise davon überzeugt werden, dass der Wert eines neuen Tools sowohl seinen hohen Preis als auch die voraussichtliche Umstrukturierung des Personals rechtfertigt.

Hervorheben des Erfolgs anderer Unternehmen

Glücklicherweise liefern die Erfahrungen anderer Unternehmen viele positive Beispiele. Vielleicht sollten Sie das Feld von hinten aufrollen: Erläutern Sie zunächst die Vorteile einer verbesserten Business Intelligence und der gewonnenen Erkenntnisse und erst dann, wie sich diese Vorteile realisieren lassen würden.

Ansprechen von Problemen und Anbieten von Lösungen

Häufige Probleme:

1. Die manuelle Berichterstellung dauert lange und ist nur geringfügig besser als ein Blindflug.
2. Die verschiedenen Geschäftsbereiche haben separate Datensilos und es fällt ihnen schwer, relevante Informationen auszutauschen.
3. Interne Tools verlieren ihre Nachhaltigkeit, wenn neue Datenquellen hinzugefügt, das Datenvolumen erhöht und die Leistungsanforderungen strenger werden.
4. Ältere Datenbanken und lokale Data Warehouses stoßen an ihre Leistungs- und Brauchbarkeitsgrenzen.
5. Engineers haben Besseres zu tun, als Datenbanken zu pflegen.

Allgemeine Vorteile:

1. Zeiteinsparungen – drastische Verkürzung der Zeit zwischen den Berichten
2. Vorteile bezüglich der Daten – massive Erhöhung der Zugänglichkeit und Aktualität von Daten
3. Qualitätsgewinne – Daten sind umfassender und aktueller
4. Vorteile bezüglich der Unternehmenskultur – Datenzugang und datenbasierte Entscheidungen werden im gesamten Unternehmen demokratisiert
5. Arbeitszeitersparnis – weniger Entwicklungszeit für maßgeschneiderte Integrationstools und die Datenbankwartung; Analysten müssen Berichte nicht manuell zusammenstellen
6. Neue Erkenntnisse und Produkte – Wegfall von Entwicklungszeit für die Datenintegration bedeutet mehr Spielraum für die Sondierung von Möglichkeiten und die Entwicklung von Produkten

Bedenken Sie, dass Führungskräfte den Datenintegrationsprozess im Allgemeinen aus der Ergebnisperspektive sehen (es sei denn, sie sind CTOs). Daher ist es wichtig, zunächst die Vorteile einer guten BI hervorzuheben und erst dann die technischen Aspekte von Data Warehousing und Datenintegration zu erläutern.

FALLSTUDIE: MVF konnte seinen Monatsumsatz dank Verwendung eines modernen Data Stacks mehr als verdoppeln

MVF ist eine Plattform zur Kundengewinnung, die Sales Leads auf Pay-per-Lead-Basis liefert. Ohne modernen Data Stack hatte MVF keine alleinige Informationsquelle (Single Source of Truth), weil separate Datenbanken und Ad-hoc-Data Stacks über Teams verteilt waren. Das Erzeugen von Berichten dauerte zwei bis drei Wochen.

Mit Fivetran konnte MVF seine Daten zentralisieren und nicht verkaufte Leads einfach ermitteln. MVF verdoppelte seinen monatlichen Umsatz mit nicht verkauften Leads von 300.000 GBP pro Monat auf rund 700.000 GBP. Die Berichte, deren Erstellung früher zwei oder drei Wochen dauerte, werden jetzt automatisch und kontinuierlich generiert.

Darüber hinaus blieb den Datenspezialisten der Anfangsaufwand erspart, zunächst vier bis acht Wochen einen Konnektor zu entwickeln, und es entfielen die laufenden Kosten für Wartung und Debugging. Durch die Automatisierung der Datenintegration konnte MVF acht weitere Datenquellen anbinden. Die Data Engineers ihrerseits widmen sich jetzt anspruchsvolleren, strategischeren Projekten.[9]

MERKE: *Der langfristige Erfolg Ihres Unternehmens hängt davon ab, ob es mit der technischen Entwicklung Schritt hält. Datenkompetenz hat für die Wettbewerbsfähigkeit am Markt mittlerweile eine entscheidende Bedeutung. Wenn sich Führungskräfte oder Datenspezialisten dem verweigern, kann das letztlich das gesamte Unternehmen gefährden.*

9 Die komplette Fallstudie finden Sie unter: fivetran.com/blog/case-study-mvf

Kapitel 4: Geschäftliche Überlegungen bei der Wahl eines Datenintegrationstools

INHALT DIESES KAPITELS:

- Wie funktioniert die Preisgestaltung?
- Passt ein Tool zu den Erfordernissen Ihres Unternehmens?
- Zukunftsfähigkeit

Ob eine automatisierte Datenintegrationslösung Zeit, Geld und Aufwand spart, hängt von der Größe und Reife Ihres Unternehmens sowie von den besonderen Merkmalen des Datenpipeline-Anbieters ab.

In sehr kleinem Maßstab benötigt Ihr Unternehmen u. U. keine Datenpipeline, vor allem, wenn Sie ein Startup in der frühen Phase sind, das nur maximal zwei Datenquellen nutzt, oder wenn Sie bei dem Versuch, Ihr Produkt markttauglich zu machen, nur qualitative Recherche betreiben. Möglicherweise hat Ihr Unternehmen auch einen Nischen-Anwendungsfall mit extrem strikten Leistungs-, Sicherheits- oder behördlichen Vorgaben. Für bestimmte datenwissenschaftliche Anwendungen können schon Latenzen im Nanosekundenbereich ein Problem darstellen.

Mit Ausnahme der oben geschilderten Szenarien hat Ihr Unternehmen wahrscheinlich mit den hohen Entwicklungskosten für den Aufbau und die Wartung von Datenkonnektoren zu kämpfen. Oder es muss aufgrund der Pflege von Konnektoren und der manuellen Berichterstellung lange Berichtserstellungszeiten in Kauf nehmen. In diesem Fall sollten Sie prüfen, ob sich der Kauf einer Datenintegrationslösung lohnt.

Wie funktionieren Preisfindung und Kosten?

Machen Sie sich mit den Preisstrukturen potenzieller Tools vertraut. Hier einige gängige Preismodelle:

- **Eine pauschale Abonnementgebühr**, die u. U. höhere Fixkosten verursacht, dafür aber die Prognostizierbarkeit der Kosten garantiert.
- **Preiskalkulation nach Datenvolumen** in Gigabyte oder Zeilen. Ein volumenbasiertes Preismodell kann sehr vorteilhaft sein, wenn Sie zurzeit mit einer eher kleinen Datenmenge arbeiten, aber über einen längeren Zeitraum ein neues Tool testen möchten, oder wenn Sie beabsichtigen, Ihren Workflow schrittweise auf das neue System umzustellen.
- **Preiskalkulation pro Platzlizenz oder Pauschalpreis** für Ihr Unternehmen. Preismodelle auf Platzlizenzbasis sind bei einer kleineren Mitarbeiteranzahl in der Regel günstiger, aber mit einem höheren Verwaltungsaufwand verbunden. Pauschalpreise für ein gesamtes Unternehmen können mit geringerem Aufwand verbunden sein und im größeren Maßstab weniger kosten.

Möglicherweise begegnen Ihnen auch Kombinationen von Preismodellen. So wird u. U. für einen Service eine Plattform-Pauschalgebühr und für jeden einzelnen Datenkonnektor eine zusätzliche Gebühr erhoben. Volumenbasierte Sätze können je nach Konnektor variieren. Möglicherweise wird ein Freemium-Modell bis zu einem bestimmten Datenvolumen oder bei unbegrenztem Datenvolumen mit einem eingeschränkten Funktionsumfang angeboten.

Passt ein Tool zu den Fähigkeiten und Zukunftsplänen Ihres Teams?

Weitere zu berücksichtigende Faktoren sind der Kompromiss zwischen Benutzerfreundlichkeit und Konfigurierbarkeit sowie die Kompatibilität mit den Fähigkeiten Ihres Teams.

Technisch wenig versierte Benutzer kennen sich vielleicht mit SQL aus. Das muss aber nicht unbedingt so sein. Mit ziemlicher Sicherheit finden sie sich aber in einem BI-Tool zurecht. Analysten beherrschen in der Regel SQL, Statistiken und möglicherweise eine Skriptsprache wie Python. Datenwissenschaftler verfügen wahrscheinlich über tiefergehende technische Kenntnisse wie erweitertes Wissen über Statistik und weitere Sprachen wie Java sowie Big-Data-Technologien wie Hadoop oder Spark. Engineers werden wahrscheinlich

mit einer Reihe von High- und Low-Level-Computersprachen sowie einigen Technologieplattformen vertraut sein.

Die verschiedenen Datenintegrationstools bieten unterschiedliche Grade an Komplexität und Zugänglichkeit. Einige basieren stark auf maßgeschneiderten Skripts und stellen lediglich das Grundgerüst für die Erstellung Ihrer eigenen Datenpipeline. Andere wiederum bieten Drag & Drop-GUIs, mit denen auch technisch eher wenig versierte Benutzer die Datenreplikation und -transformation steuern können. Diese haben jedoch zwei klare Nachteile: eine steile, stark plattformspezifische Lernkurve und die automatische Generierung von Spaghetti-Code. Wieder andere kombinieren die vollautomatische Datenreplikation mit versionskontrollierten SQL-basierten Transformationen.

Unter dem Strich müssen Sie zwischen Zugänglichkeit und Konfigurierbarkeit abwägen. Wenn Ihr Ziel darin besteht, die Datenkompetenz im gesamten Unternehmen zu fördern, sollten Sie ein Tool mit sehr niedrigen Zugangshürden und einer breiten Anwendbarkeit auf verschiedene Anwendungsfälle suchen.

Für spezialisiertere Anwendungsfälle eignen sich hingegen konfigurierbare Tools mit Optimierung für bestimmte Nischen, die weniger zugänglich, aber dafür leistungsstärker sind.

Anbieterbindung und sich ändernde Erfordernisse

Bevor Sie sich vertraglich binden, sollten Sie prüfen, ob das Tool auch zukünftig Ihre Erfordernisse erfüllt. Dazu sollten Sie sich folgende Fragen stellen:

- Enthält das Tool die Konnektoren, die Sie gegenwärtig verwenden oder voraussichtlich verwenden werden?
- Können Sie bei Ihrem Konto bei Bedarf problemlos zusätzliche Funktionen aktivieren oder Datenkonnektoren hinzufügen?
- Werden Konnektoren konsequent aktualisiert, damit sie bei Änderungen an den vorgelagerten APIs aktuell bleiben, und werden diese Aktualisierungen von Änderungsprotokollen begleitet, in denen die Änderungen verzeichnet sind?
- Wird das Tool regelmäßig um neue Konnektoren erweitert?
- Ist das Support-Team gut erreichbar und in der Lage, Produktänderungen und sich ändernden Erfordernissen Rechnung zu tragen?
- Können Sie Datenmodelle und Transformationen von einer Plattform auf eine

andere exportieren, oder müssen Sie sie zurückentwickeln und neu generieren, wenn Sie auf ein neues Tool umsteigen?

Zukunftssicherheit ist ein wichtiger Aspekt, weil das Wechseln von Plattformen sehr kostspielig und mit Unterbrechungen verbunden sein kann.

TIPP: *SQL gibt es zwar in einigen Versionen, in der Datenanalyse ist es aber Branchenstandard. In SQL geschriebene Datenmodelle und Transformationen müssten grundsätzlich einfach von einem System auf ein anderes zu portieren sein. Anders ist das bei Prozeduren, Datenmodellen und Transformationen, die in proprietären Dateisystemen oder Sprachen gespeichert sind. Sie lassen sich nicht einfach portieren und bergen das hohe Risiko der Bindung an einen Anbieter.*

MERKE: *Zur Bewertung der geschäftlichen Eignung eines bestimmten Tools schauen Sie sich an, wie hoch die Gesamtbetriebskosten sind und ob das Tool sie vor möglichen Komplikationen im Unternehmen schützt – sich ändernde Erfordernisse, Anfälligkeit für Unfälle und Ausfälle sowie die Einhaltung rechtlicher Bestimmungen.*

Kapitel 5: Technische Überlegungen bei der Wahl eines Datenintegrationstools

INHALT DIESES KAPITELS::

- ETL und ELT im Vergleich
- Evaluieren der Qualität von Datenkonnektoren
- Wie funktioniert Automatisierung in Ihrem Stack?

Sobald Sie festgestellt haben, dass Ihr Unternehmen eine Datenintegrationslösung benötigt, sollten Sie sich die technischen Merkmale der einzelnen Tools anschauen.

Qualität der Datenkonnektoren

Den Grundbaustein jeder ELT-Datenpipeline bildet der Datenkonnektor. Ein Datenkonnektor nimmt Daten von einer API oder einem Datenbankprotokoll entgegen, führt eine einfache Bereinigung und Normalisierung durch und lädt sie dann in ein Data Warehouse. Bei der Bewertung der Qualität der Datenkonnektoren sollten Sie Folgendes berücksichtigen:

- **Open Source oder proprietär?** Wie bei anderer Software gibt es auch hier zwei Ansätze: Software, die von Freiwilligen per Crowd-Sourcing betreut wird, und rein kommerzielle Software. Insgesamt gibt es mehr Open-Source-Datenkonnektoren für eine breitere Palette von Datenquellen. Proprietäre Konnektoren sind jedoch

meist von höherer Qualität und lassen sich nahtloser in andere Komponenten eines Data Stacks integrieren. Insbesondere Anbieter proprietärer Technik haben einen größeren Anreiz, strenge QA-, Wartungs- und Entwicklungsprinzipien anzuwenden.

- **Standardisierte Schemen und Normalisierung.** Daten aus API-Feeds werden normalerweise nicht in normalisierter Form bereitgestellt. Die Normalisierung fördert die Datenintegrität, weil Redundanzen beseitigt und eindeutige, einheitliche Beziehungen zwischen Tabellen hergestellt werden. Für einen gegebenen Datensatz gibt es so viele Meinungen wie mögliche Schemen, aber nur eine Handvoll möglicher normalisierter Schemen. Weil es nur wenige Möglichkeiten gibt, einen Datensatz zu normalisieren, eignet sich die Normalisierung auch für die Standardisierung von Schemen. Dadurch entstehen Skaleneffekte, von denen alle Benutzer profitieren.

TIPP: *Sie sollten sich unbedingt die Entitäten-Beziehungsdiagramme (Entity Relationship Diagrams: ERDs) in der Dokumentation des Anbieters anschauen. Diese Diagramme veranschaulichen die Schemen. Aus den ERDs sollte klar hervorgehen, welche Felder in den einzelnen Tabellen verfügbar sind und welche Beziehungen zwischen den Tabellen herrschen. Ihre Analysten müssten daraus schließen können, ob das Schema hilfreiche Felder enthält und normalisiert ist.*

- **Inkrementelle oder vollständige Updates?** Wie sieht die Replikationsstrategie des Konnektors aus? Bei der Erstsynchronisierung muss entweder der gesamte Datensatz oder ein großer Teil von ihm abgefragt werden. Bei späteren Aktualisierungen sollte das jedoch nicht erneut erfolgen. Wird der Konnektor inkrementell mithilfe von Protokollen oder anderen Formen der Änderungserkennung aktualisiert oder werden bei jeder Synchronisierung die vollständigen Datensätze abgefragt? Die inkrementelle Methode ermöglicht häufigere Aktualisierungen mit geringerem Volumen. Häufige vollständige Replikationen von operativen Datenbanken bergen zudem die Gefahr, dass kritische Geschäftsvorgänge beeinträchtigt werden.

Unterstützung von Quellen und Zielen

Die verschiedenen Datenpipeline-Tools unterstützen unterschiedliche Datenquellen und Data Warehouses. Vergewissern Sie sich, dass das von Ihnen evaluierte Tool diejenigen unterstützt, die für Sie wichtig sind. Wenn nicht, bietet der Anbieter den Kunden dann die Möglichkeit, neue Quellen und Ziele vorzuschlagen? Fügen sie routinemäßig neue Quellen hinzu?

Unterstützt das Tool zu diesem Zweck mehrere Quellen und Ziele? Ihr Unternehmen nutzt möglicherweise Dutzende oder Hunderte von Konnektoren für dieselbe Art von Datenquellen, wenn Sie beispielsweise viele Werbekonten im Auftrag Ihrer Kunden verwalten oder Ihr Unternehmen mit einem anderen fusioniert oder es übernimmt und Daten von mehreren Konten und Plattformen kombinieren muss. Auch aus Redundanzgründen können Sie die Synchronisierung mit mehreren Data Warehouses wählen.

Schauen Sie sich abschließend an, ob und in welchem Umfang das Tool kundenspezifische Datenintegrationen unterstützt. Möglicherweise müssen Sie Daten aus intransparenten Datenquellen integrieren, die nicht von einem Standard-Konnektor von der Stange unterstützt werden..

Unterstützt das evaluierte Tool Cloud-basierte Funktionen, mit denen Sie die eigenen Konnektoren Ihrer Engineers mit dem Rest Ihrer Infrastruktur kombinieren können? Unterstützt das Tool zur Not das Ad-hoc-Laden und -Warehousing von Daten aus CSV oder JSON?

Konfiguration oder „Zero-Touch“?

Hochgradig anpassbare und konfigurierbare Tools ermöglichen es Benutzern, noch den letzten Parameter zu optimieren und genau den gewünschten Workflow zu definieren. Das setzt Spezialisten voraus, die sich mit Skriptsprachen auskennen, über viel Erfahrung mit der Orchestrierung verfügen und gut im Schreiben robuster Software sind. Außerdem müssen sie jede einzelne Datenquelle genau verstehen oder eng mit den Analysten zusammenarbeiten, um die Daten zu untersuchen, zu verstehen und zu modellieren. Aus Schemen müssen letztlich nutzbare Datenmodelle werden, und das Entwerfen eines guten Schemas sowie der Übergang von Rohdaten zu aufbereiteten Daten ist schwierig und kann Kunst und Wissenschaft zugleich sein.

Bei einem hochgradig konfigurierbaren Ansatz müssen Benutzer die Datenintegrationssoftware richtig konfigurieren und warten. Dies umfasst die Neukonfiguration von Pipelines, wenn sich nachgelagerte Geschäftsanforderungen und vorgelagerte Datenquellen ändern. Dieser Ansatz eignet sich am besten für Unternehmen mit vielen Fachleuten auf diesem Gebiet, die sich dieser Herausforderungen aktiv stellen möchten und davon überzeugt sind, bessere und zuverlässigere Ergebnisse als ein Produkt von der Stange liefern zu können.

ACHTUNG: *Es gibt auch GUI-basierte Datenintegrationstools, mit denen auch keine ausgewiesenen Spezialisten Orchestrierungen und Transformationen visuell programmieren können. Anstelle eines hochqualifizierten Engineer-Teams benötigen Sie dann Analysten oder Endbenutzer, die mit einer proprietären visuellen Programmiersprache vertraut sind. Dies kann zu Problemen wie einer zu starken Spezialisierung oder Bindung an einen Anbieter führen.*

Anders sieht es bei vollständig verwalteten Zero-Touch-Tools aus. Aufgrund ihres wartungsfreien Charakters (Set and Forget) weisen sie eine hohe Zugänglichkeit auf. Aus Kundensicht sind die Konnektoren standardisiert, unter Belastung getestet und wartungsfrei. Die Wartung und künftige Iterationen der Konnektoren werden zur Leistungspflicht von absoluten Spezialisten, die jede Eigenart der zugrunde liegenden Daten kennen und ihre Konnektoren an einer Vielzahl von Grenzfällen getestet haben.

Anstatt die Daten vor dem Laden zu orchestrieren und zu transformieren, können Transformationen von Analysten mithilfe von SQL geplant und durchgeführt werden. Dadurch eignet sich der Zero-Touch-Ansatz sehr viel besser für Unternehmen, die für die Entwicklung und Wartung von Pipelines nicht auf viele erstklassige Spezialisten zurückgreifen können und deren Fähigkeiten lieber für andere anspruchsvolle Projekte einsetzen möchten.

TIPP: *Self-Service ist ein weiterer wichtiger Gesichtspunkt. Ermitteln Sie, ob und inwieweit das von Ihnen evaluierte Tool die Möglichkeit bietet, ein Konto ohne Hilfe eines Kundenbetreuers oder Kundendienstmitarbeiters zu eröffnen. Self-Service beschert Ihrem Team vielleicht einen etwas größeren Aufwand; dafür können Sie ein Abonnement aber auch einfacher starten und kündigen.*

Automatisierung

Moderne Datenintegrationstools sollen den Prozess von so viel manuellem Aufwand wie möglich befreien. In diesem Sinne sollten Sie die folgenden arbeitssparenden Automatisierungstools und -funktionen in Betracht ziehen:

- **API**. Es kann äußerst hilfreich sein, das Tool per Programm so steuern zu lassen, dass Verwaltungsfunktionen und andere Routineaufgaben automatisch statt von Hand ausgeführt werden können. Besonders hilfreich kann das sein, wenn eine große Anzahl von Personen unterschiedliche Grade der Kontrolle über das Tool

benötigt oder wenn Sie auf der Basis der Datenintegration Produkte erstellen.

- **Umgang mit Datentyp-Änderungen**. Vorgelagerte Schemaänderungen können den Typ eines bestimmten Werts ändern, z. B. von Integer zu Float. Ein automatisiertes Tool muss in der Lage sein, alte und neue Datentypen ohne menschliches Eingreifen abzugleichen.
- **Planung einer kontinuierlichen Synchronisierung**. Daten von diesen Konnektoren sollten entweder kontinuierlich in Ihr Data Warehouse fließen oder in kurzen, regelmäßigen Abständen synchronisiert werden. Legen Sie einmalig fest, wie oft die Daten aktualisiert werden müssen.
- **Automatische Migration von Schemen**. Schemen werden sich zwangsläufig ändern, wenn Datensätze um weitere Datenelemente erweitert werden. Akzeptiert der Konnektor diese Änderungen automatisch mit einem Minimum an Unterbrechungen für die nachgelagerten Elemente, ohne dabei Tabellen oder Felder zu löschen? Vermeidet der Konnektor nach Möglichkeit eine komplette Neusynchronisierung?
- **Allgemeine Performance**. Abschließend sollten Sie eine Reihe von Fragen betrachten, von denen mögliche Ausfallzeiten Ihres Systems abhängen:
    - Wie lange dauert die Erstsynchronisierung?
    - Werden die Daten inkrementell aktualisiert oder ist jedes Mal eine vollständige Synchronisierung erforderlich?
    - Welche Bedingungen lösen eine vollständige Synchronisierung aus?
    - Wie oft werden die Daten aktualisiert und deckt sich dies mit Ihren Erfordernissen?
        - Transformieren im bzw. vor dem Data Warehouse Transformieren im bzw. vor dem Data Warehouse?

Die Antworten auf diese Fragen können Einfluss auf die Kosten haben, die durch Ausfallzeiten und Auslastung der Infrastruktur entstehen – Kosten, die u. U. im formalen Preisgefüge eines Pipeline-Tools nicht berücksichtigt werden, aber eine erhebliche Belastung für Ihr Unternehmen darstellen können.

Transformieren im bzw. vor dem Data Warehouse

Bei der ELT-Methode erfolgen Transformationen in einem elastischen, Cloud-basierten Data Warehouse. Die Elastizität – sowie die Trennung von Rechenarbeit und Speicherung – ermöglicht die bedarfsgerechte Skalierung von Ressourcen. Das macht es überflüssig, Hardwarebedarf zu prognostizieren und eventuell

überschüssige Kapazitäten zu kaufen.

ETL hingegen – ganz gleich, ob Cloud-basiert oder On-Premise – erfordert eine Datenarchitektur mit einer zusätzlichen Phase im Data Stack, um Transformationen vor dem Laden zu verarbeiten. Bei einem lokalen Data Stack kann das Data Warehouse selbst die Menge an geladenen Daten so einschränken, dass Transformationen erforderlich sind, um das Datenvolumen und den Datenfluss zu begrenzen.

ELT und im Data Warehouse durchgeführte Transformationen haben den grundlegenden Vorteil, dass sie nicht destruktiv sind. Das heißt, die zugrunde liegenden Daten bleiben komplett unberührt, wenn zusätzliche Tabellen mit den gewünschten Modellen erstellt werden. Das bedeutet, dass fehlgeschlagene Transformationen keine dauerhaften Folgen haben und wiederholt durchgeführt werden können. Außerdem können Analysten Modelle an sich ändernde Geschäftsanforderungen anpassen, ohne dass dabei Daten verloren gehen.

Ein letzter Vorteil der Durchführung von Transformationen im Data Warehouse besteht darin, dass die Transformationen in SQL geschrieben werden können und dadurch für Analysten zugänglich sind. Normalerweise erstellen Analysten Ansichten in einem Data Warehouse, um Tabellen zu konsolidieren oder zu ändern. Bei ELT-Tools, die Transformationen im Warehouse unterstützen, können Analysten systematisch Ansichten erstellen.

Wiederherstellung nach einem Ausfall

Fehler sind im Verlauf der Datenintegration unvermeidbar, und Datenintegrationen schlagen unweigerlich fehl. Schlimm wäre das nur, wenn dabei versehentlich und dauerhaft Daten verloren gehen würden.

Ein wichtiges Merkmal des Datenintegrationstools ist die Idempotenz. Das ist die Fähigkeit, ein und denselben Prozess wiederholt durchzuführen und jedes Mal dasselbe Ergebnis zu erhalten. Besonders nützlich ist das bei komplizierten, mehrstufigen Prozessen, bei denen der genaue Ausfallpunkt nicht offenkundig ist.

Ein weiteres wichtiges Prinzip ist die netto-additive Integration. Wenn ein Wert in den Quelldaten gelöscht oder eine Tabelle verworfen wird, wir er/sie dann im Data Warehouse beibehalten (aber entsprechend gekennzeichnet)? Bei Beibehaltung, aber Kennzeichnung eines nicht mehr gültigen Werts bleiben historische Datensätze erhalten. Das ist hilfreich für Audits, die Wiederherstellung nach einem Ausfall sowie die Analyse von längerfristigen Trends und Datenschwund.

TIPP: *Lesen Sie sich unbedingt die Service Level Agreements (SLAs) aller Anbieter durch, die Sie in Betracht ziehen, und nehmen Sie sie in die Verantwortung! Stellen Sie insbesondere sicher, dass der gewählte Anbieter dieselben SLA-Leistungen erbringt, die Sie von Ihrem Team verlangen würden. Ein SLA formuliert klare Erwartungen hinsichtlich Verfügbarkeit, Ausfallzeit, Geschwindigkeit und Volumen des Datentransfers sowie weiterer Leistungskennziffern.*

Sicherheit und Einhaltung rechtlicher Vorgaben

Cybersicherheit und Datenschutz sind sowohl in rechtlicher Hinsicht als auch im Hinblick auf die öffentliche Wahrnehmung eine heikle Angelegenheit.

Folgende Punkte sind dabei zu bedenken:

- **Erfüllung gesetzlicher Auflagen**. Ihr Datenintegrationsanbieter sollte mindestens mit Standards wie der DSGVO, SOC 2, HIPPA und anderen relevanten Bestimmungen vertraut sein. Ein gutes Tool unterstützt die Möglichkeit, personenbezogene Daten gar nicht zu erheben, zu löschen oder zu verschlüsseln.
- **Rechte an Ihren Daten.** Ihr Datenintegrationsanbieter darf nur so lange auf die Daten zugreifen bzw. diese behalten, solange ihre Replikation dauert..
- **Rollen mit variierenden Zugangsstufen**. Nicht jeder, der das Tool verwendet, darf uneingeschränkt befugt sein, Warehouses, Konnektoren oder Transformationen zu erstellen, zu löschen oder zu ändern bzw. andere datenschutzrelevante Aktionen auszuführen. Das Tool muss eine Reihe von Rollen bieten – von schreibgeschütztem Zugriff bis Administrator.
- **Spalten sperren und verschleiern**. Aus Gründen der Sicherheit und der Erfüllung gesetzlicher Auflagen sollten Sie in der Lage sein, personenbezogene Daten in jeder von Ihnen synchronisierten Tabelle zu verschleiern oder auszulassen.

MERKE: *Bei der Auswahl eines Datenintegrationstools kommt es im Wesentlichen darauf an, inwieweit es die Arbeit von Analysten und Engineers erleichtert.*

Folgendes sollten Sie dabei berücksichtigen:

- *Qualität der Datenkonnektoren*
- *Ob das Tool Datenquellen und Ziele unterstützt, die Sie derzeit verwenden oder verwenden möchten*
- *Wie viel manuelle Konfiguration erforderlich ist*
- *Wie gut es ohne manuellen Eingriff oder Überwachung läuft*
- *Wann das Tool Transformationen durchführt*
- *Wie ausfallsicher das Tool ist*
- *Sicherheit und Einhaltung rechtlicher Vorgaben*

Probieren Sie mehrere Tools aus! Im nächsten Abschnitt geht es um die ersten Schritte.

Kapitel 6: Sieben Schritte für den Einstieg

INHALT DIESES KAPITELS:

- Was sind Ihre Erfordernisse und Ziele?
- Wie definieren Sie Erfolg für sich?
- Prüfen Sie, bevor Sie kaufen, und prüfen Sie gründlich!

Auch wenn ein Cloud-basierter, vollständig verwalteter Data Stack viel verspricht, eignet er sich nicht für jedes Unternehmen.

Um den richtigen Kurs für Ihr Unternehmen zu wählen, müssen Sie:

1. Ihre Erfordernisse gründlich analysieren
2. Entscheiden, ob Sie eine Migration vornehmen oder von Grund auf neu beginnen möchten
3. Cloud-Data-Warehouse- und Business-Intelligence-Tools evaluieren
4. Datenintegrationstools evaluieren
5. Die Gesamtbetriebskosten berechnen
6. Erfolgskriterien definieren
7. Ein Machbarkeitskonzept aufstellen

Analyse Ihrer Erfordernisse

Es gibt u. U. einige Gründe, die dafür sprechen, Ihre Datenvorgänge nicht an Dritte oder eine Cloud auszulagern.

Der erste und offenkundigste wäre, dass Ihr Unternehmen sehr klein ist oder nur mit Daten in geringer Menge oder Komplexität operiert. Wenn Sie als Vier-Personen-Startup immer noch dabei sind, Ihren Platz im Markt zu finden, haben Sie u. U. gar keine Daten, die sich analysieren ließen. Das kann dann der Fall sein, wenn Sie nur eine oder zwei Anwendungen nutzen, wahrscheinlich auch keine neuen Anwendungen hinzukommen und Ihre in die Anwendungen integrierten Datenanalysetools bereits ausreichen.

Ein zweiter Grund, keinen modernen Data Stack zu kaufen, könnte sein, dass er u. U. bestimmte Leistungsvorgaben oder rechtliche Vorgaben nicht erfüllt. Wenn Sie ein Unternehmen sind, das Hochfrequenzhandel betreibt, und Nanosekunden über den Erfolg oder Misserfolg Ihrer Aktivitäten entscheiden, ist es vielleicht ratsam, keine Cloud-Infrastruktur eines Drittanbieters zu nutzen, sondern dafür eigene Hardware zu erstellen.

Sollte Ihr Unternehmen eine ausreichende Größe oder Reife aufweisen, um von Datenanalysen zu profitieren, und sollten Datenaktualisierungszyklen von einigen Minuten oder Stunden akzeptabel sein, fahren Sie fort.

Migration oder Neubeginn

Datenintegrationsanbieter sollten in der Lage sein, Daten von der alten Infrastruktur in Ihren neuen Data Stack zu migrieren. Aufgrund der Komplexität und Vielfalt der Daten ist diese Aufgabe jedoch als lästig verschrien. Ob sich Ihr Unternehmen für eine Migration entscheidet oder einfach bei Null beginnt, hängt stark davon ab, für wie wichtig historische Daten erachtet werden.

Wenn Ihr Unternehmen bereits Produkte oder Dienstleistungen gekauft oder Verträge abgeschlossen hat, kann es teuer werden, diese Verträge zu kündigen. Abgesehen vom Geld kann die Vertrautheit mit bestimmten Tools und Technologien und deren Bevorzugung ein wichtiger Gesichtspunkt sein.

Achten Sie darauf, dass potenzielle Lösungen mit allen Produkten und Dienstleistungen kompatibel sind, mit denen Sie weiterarbeiten möchten.

TIPP: *Ein schrittweises Vorgehen ist völlig normal. Viele Unternehmen richten ein neues Data Warehouse ein, behalten aber zunächst ihr altes, demnächst nicht mehr benötigtes Data Warehouse bei, bis alle Daten und Prozesse in die neue Umgebung überführt wurden.*

Evaluieren von Cloud-Data-Warehouse- und Business-Intelligence-Tools

Sie müssen Lösungen für jeden Teil des Data Stacks vergleichen. Bevor Sie sich ein Datenintegrationstool besorgen, sollten Sie etwas früher ansetzen und überlegen, welche Funktionen Sie in einem Cloud-Data-Warehouse- und Business-Intelligence-Tool benötigen.

Folgende Cloud-Data-Warehouse-Funktionen sollten in die Überlegungen einfließen:

1. Zentrale oder dezentrale Datenspeicherung
2. Elastizität – kann das Data Warehouse die Ressourcen schnell nach oben oder unten skalieren? Sind Rechen- und Speicherressourcen voneinander getrennt oder eng miteinander verzahnt?
3. Gleichzeitigkeit – kann das Data Warehouse mehrere Aufgaben gleichzeitig ausführen?
4. Lade- und Abfrage-Performance
5. Data Governance und Metadatenverwaltung
6. SQL-Dialekt
7. Unterstützung von Backup und Wiederherstellung
8. Ausfallsicherheit und Verfügbarkeit
9. Sicherheit

Folgende BI-Tool-Funktionen sollten in die Überlegungen einfließen:

1. Nahtlose Integration in Cloud-Data-Warehouses
2. Bedienfreundliche Drag & Drop-Benutzeroberflächen – besonders hilfreich, wenn Sie im gesamten Unternehmen eine datenorientierte Kultur schaffen möchten
3. Automatisierte Berichterstellung und Benachrichtigungen
4. Möglichkeit der Ausführung von Ad-hoc-Berechnungen und Berichten durch Import und Export von Datendateien
5. Geschwindigkeit, Performance und Ansprechverhalten
6. Modellierungsschicht mit Versionskontrolle und Entwicklungsmodus
7. Umfangreiche Bibliothek mit Visualisierungen

Vergewissern Sie sich, dass alle von Ihnen evaluierten Data Warehouses und BI-Tools miteinander kompatibel sind. Es lohnt sich auch, verschiedene Tools gründlich aus unterschiedlichen Perspektiven zu prüfen.
Publikationen wie Gartner stellen solche Informationen häufig zusammen. Auch hier gilt: Drum prüfe, wer sich ewig bindet!

Evaluieren von Datenintegrationstools

Wie an früherer Stelle erwähnt, sind im Hinblick auf Datenintegrationstools viele wichtige Punkte zu berücksichtigen.

Hier eine kurze Liste:

1. CAnpassung und Konfigurierbarkeit kontra Benutzerfreundlichkeit und Zugänglichkeit
2. Ausfallsicherheit und Performance der Software
3. Qualität und Reaktionsvermögen der Kundenserviceteams
4. Anzahl und Art der abgedeckten Datenquellen
5. Kosten und Zahlungspläne

Viele Publikationen bieten aggregierte Rezensionen und Bewertungen von Datenintegrationstools, so auch für Data Warehouses und Business-Intelligence-Tools. Sie sollten die Tools also unbedingt vergleichen.

Vergewissern Sie sich, dass die von Ihnen in Betracht gezogenen Datenintegrationstools mit den Data Warehouses und BI-Tools kompatibel sind, die Sie bereits nutzen oder nutzen möchten.

Kalkulieren von Gesamtbetriebskosten und Investitionsrendite

Der moderne Data Stack verspricht erhebliche Zeit-, Geld- und Arbeitsersparnis. Vergleichen Sie Ihren vorhandenen Datenintegrations-Workflow mit einer Reihe möglicher Kandidaten.

Berechnen Sie die Kosten Ihrer gegenwärtigen Datenpipeline. Das erfordert u. U. eine sorgfältige Prüfung der bisherigen Ausgaben für Datenintegrationsaktivitäten. Zu berücksichtigen sind dabei der Anschaffungspreis, die Kosten für Konfiguration

und Wartung sowie sämtliche Opportunitätskosten, die durch Störungen, Unterbrechungen und Ausfallzeiten entstehen. Zudem müssen Sie die Kosten Ihres Data Warehouse und Ihres BI-Tools berücksichtigen.

Auf der Nutzenseite sind die Vorteile des potenziellen Ersatzes zu bewerten. Manche Nutzeffekte sind u. U. nicht einfach quantifizierbar oder berechenbar (z. B. eine bessere Arbeitsmoral der Analysten); andere wie z. B. Zeit- und Geldersparnisse lassen sich hingegen einfach quantifizieren.

Definieren von Erfolgskriterien

Wie sollte Ihre Datenanalyse in der Praxis aussehen, wenn Sie einen modernen Data Stack erfolgreich implementiert haben?

Das sind dabei die wichtigsten Kriterien:

1. Zeit-, Arbeits- und Geldersparnis gegenüber der vorherigen Lösung
2. Erweiterte Fähigkeiten des Datenteams
3. Erfolgreiche Umsetzung neuer Datenprojekte, z. B. Kundenattributionsmodelle
4. Geringere Erstellungszeit für Berichte
5. Geringere Ausfallzeit der Dateninfrastruktur
6. Stärkere Verwendung von Business-Intelligence-Tools in Ihrem Unternehmen
7. Neue Kennziffern, die verfügbar und als Handlungsgrundlage nutzbar sind

Aufstellen eines Machbarkeitskonzepts

Wenn Sie Ihre Suche auf ein paar Kandidaten eingegrenzt und die Erfolgskriterien definiert haben, testen Sie die Produkte mit möglichst geringem Kostenaufwand. Die meisten Produkte können für einige Wochen kostenlos zur Probe genutzt werden.

Richten Sie Konnektoren zwischen Ihren Datenquellen und Data Warehouses ein und erfassen Sie, wie viel Zeit und Aufwand die Synchronisierung Ihrer Daten erfordert. Führen Sie einige elementare Transformationen durch. Geben Sie Ihrem Team speziell für das Testen Zeit und weisen Sie es an, das System auf jede erdenkliche Weise unter Belastung zu testen.

Vergleichen Sie die Ergebnisse Ihres Tests mit den von Ihnen aufgestellten Erfolgskriterien.

MERKE: *Durch die automatisierte Datenintegration können die Fähigkeiten Ihrer Analysten und Data Engineers enorm erweitert werden. Bevor Sie loslegen, sollten Sie aber genau wissen, worin Ihre Erfordernisse bestehen und was Sie erreichen möchten. Überlegen Sie sich, wie ein Erfolg (oder Misserfolg) aussehen könnte, und unterziehen Sie einen Data Stack unbedingt einem Stresstest, bevor Sie ihn endgültig implementieren.*

Wettbewerbsfähige Datenanalyse beginnt mit automatisierter Datenintegration

Die Zunahme von Cloud-Daten eröffnete bis dato nicht gekannte Möglichkeiten, neue Produkte zu entwickeln und hochkomplexe Datenanalysen durchzuführen. Die Menge und Komplexität dieser Daten stellen jedoch eine Herausforderung bezüglich der Datenintegration dar, für die nur wenige Unternehmen gerüstet sind. Die automatisierte Datenintegration hilft Unternehmen, ihre Daten vollumfänglich zu nutzen, um strategische Entscheidungen zu beschleunigen und besser zu machen. In diesem Leitfaden untersuchen wir die Geschichte der Datenintegration, evaluieren aktuelle Lösungen und zeigen Ihnen, wie Sie das beste Integrationstool für Ihr Unternehmen finden. integration tool for your business.

Inhalt:

- Die geschichtliche Entwicklung des modernen Data Stack
- Warum Datenintegration heute noch schwieriger als früher ist
- Warum Automatisierung entscheidend für die moderne Datenintegration ist
- Eine Datenintegrationslösung selbst entwickeln oder kaufen?
- Geschäftliche und technische Kriterien für die Auswahl einer Datenintegrationslösung
- Implementierung einer Datenintegrationslösung

Charles Wang, Product Evangelist bei Fivetran, war zuvor als Datenanalyst, Datenwissenschaftler und Produktmanager tätig.

ISBN 978-1-7346299-0-3

9 781734 629903 90000>

Über Fivetran

Fivetran ist führend im Bereich automatisierte Datenintegration und liefert seinen Kunden einsatzbereite Datenkonnektoren, Transformationen und Muster für die Datenanalyse. Datenkonnektoren von Fivetran speisen Daten kontinuierlich in ein zentrales Repository ein und passen sich bei Änderung von Schemen und APIs an. Das garantiert einen einfachen und zuverlässigen Zugriff auf Daten und versetzt Sie in die Lage, so viele SaaS-Anwendungen wie nötig einzusetzen und die von ihnen erzeugten Daten mit großem Vertrauen in die Ergebnisse zu analysieren. Mehr über Zukunft der Datenintegration erfahren Sie auf fivetran.com.