Strategisches Patent-Intelligence-Dossier: Eine umfassende Analyse der Google Patents Public Datasets auf BigQuery

1. Einleitung: Der strukturelle Wandel der globalen Patentanalytik

Die Welt der Analyse von gewerblichem Rechtsschutz (Intellectual Property, IP) befindet sich in einer Phase tiefgreifender Transformation. Historisch betrachtet war die Patentrecherche eine Disziplin, die durch manuelle Kuration, physische Archive und später durch starre, schlüsselwortbasierte Datenbanken definiert wurde. Diese traditionellen Methoden, die oft auf proprietären Silos und teuren Lizenzmodellen basierten, stoßen im Zeitalter exponentiell wachsender Datenmengen an ihre physischen und kognitiven Grenzen. Mit weltweit über 100 Millionen Patentdokumenten und jährlich Millionen neuer Anmeldungen ist eine rein manuelle oder auf simplen Suchstrings basierende Sichtung nicht mehr skalierbar, um strategische Entscheidungen in Echtzeit zu treffen.

Der vorliegende Bericht untersucht die technologische Revolution, die durch die Integration von globalen Patentdaten in moderne Cloud-Data-Warehouses ausgelöst wurde, wobei der Fokus exklusiv auf der Infrastruktur von Google BigQuery und den damit verbundenen Public Datasets liegt. Google hat durch die Bereitstellung massiver Datensätze – von bibliografischen Metadaten über Volltexte bis hin zu maschinellen Lernvektoren (Embeddings) – unter offenen Lizenzen den Zugang zu High-End-Patentanalytik demokratisiert.[1]

Forschende, Unternehmen und Analysten stehen nun vor der Möglichkeit, Fragen zu beantworten, die weit über die reine „Prior Art"-Suche (Stand der Technik) hinausgehen. Es geht nicht mehr nur darum, ob eine Erfindung neu ist, sondern um die Modellierung makroökonomischer Trends, die Messung von Innovationsgeschwindigkeiten, die Kartierung komplexer Wettbewerber-Netzwerke und die Vorhersage technologischer Disruptionen durch semantische KI-Modelle.[2]

Dieses Dossier dient als erschöpfendes Handbuch und strategischer Leitfaden für die Nutzung der „Google Patents Public Datasets" auf BigQuery. Es richtet sich an Data Scientists, IP-Strategen und ökonomische Forschende, die das volle Potenzial dieser Daten ausschöpfen wollen. Die Analyse deckt die Datenarchitektur, die technische Implementierung mittels SQL, die Anwendung modernster KI-Verfahren wie Vektor-Suche sowie die Governance-Aspekte ab. Ziel ist es, ein tiefes Verständnis dafür zu entwickeln, wie diese Datenstrukturen genutzt werden können, um Wettbewerbsvorteile zu generieren und die Mechanismen des globalen Innovationssystems zu entschlüsseln.

2. Die Dateninfrastruktur: Das Ökosystem der Google Public Datasets

2.1 Das Paradigma des Cloud Data Warehousing

Grundlage der modernen Patentanalyse in der Google Cloud ist das „Public Dataset Program“. In diesem Modell hostet Google wertvolle Datensätze (wie Patentdaten, Wetterdaten oder Genomdaten) und übernimmt vollständig die Speicherkosten, was eine signifikante Abkehr von traditionellen Modellen darstellt, bei denen der Endnutzer für das Hosting der Daten verantwortlich war.[3] Der Nutzer zahlt in diesem Ökosystem lediglich für die Rechenleistung (Queries), die für die Analyse der Daten anfällt. Dies senkt die Eintrittsbarriere für komplexe Analysen drastisch, da keine eigene teure Infrastruktur für das Hosting von Terabytes an Patentdaten vorgehalten werden muss.[4]
Ein entscheidender Aspekt für die Verbreitung dieser Technologie ist das „Free Tier“: Jeder Nutzer erhält monatlich 1 TB an Abfragevolumen kostenlos. Dies ermöglicht es Forschern und Entwicklern, Prototypen zu entwickeln und explorative Datenanalysen durchzuführen, ohne sofortige Kosten zu verursachen.[1] Die Daten liegen in BigQuery, einem serverlosen, hochskalierbaren Data Warehouse, das SQL (Structured Query Language) unterstützt. Die Bedeutung dieser Architektur kann nicht hoch genug eingeschätzt werden: Da BigQuery Speicher und Rechenleistung entkoppelt, können selbst petabyte-große Abfragen in Sekundenbruchteilen ausgeführt werden, indem tausende von Slots (virtuelle CPUs) parallel arbeiten. Dies bedeutet, dass keine proprietären Abfragesprachen erlernt werden müssen; Standard-SQL (GoogleSQL) ist ausreichend, was die Integration in bestehende Business-Intelligence-Workflows massiv erleichtert.[3]

2.2 Die Fragmentierung und Taxonomie der Datensätze

Die Patentdaten in BigQuery sind nicht monolithisch in einer einzigen Tabelle gespeichert. Vielmehr verteilen sie sich auf verschiedene Projekte und Datasets, die unterschiedliche Schwerpunkte, Aktualisierungszyklen und Quellen haben. Eine genaue Kenntnis dieser Unterscheidungen ist für valide Analysen essenziell, da die Wahl der falschen Tabelle zu unvollständigen oder verzerrten Ergebnissen führen kann. Insgesamt stehen dem Analysten etwa 19 verschiedene Datensätze zur Verfügung.[4]
Die folgende Tabelle bietet eine strukturierte Übersicht über die primären Datensätze und ihre strategische Ausrichtung:

Dataset-ID	Beschreibung & Inhalt	Primäre Quelle	Strategischer Nutzen
patents-public-data. patents	Globaler bibliografischer	IFI CLAIMS Patent Services / Google	Dient als „Single Source of Truth“ für

	Datensatz, Volltexte (US), Zitationen, CPC/IPC Klassifikationen.		weltweite Recherchen. Deckt über 100 Länder ab und ist ideal für globale Trendanalysen.[5]
patents-public-data.patentsview	Hochdetaillierte US-Daten (ab 1976). Enthält normalisierte Daten zu Erfindern, Anmeldern (Assignees), Anwälten und Geografie.	USPTO / PatentsView Initiative	Unverzichtbar für Netzwerkanalysen (Wer arbeitet mit wem?), Erfinder-Tracking und regionale Cluster-Analysen in den USA.[6]
patents-public-data.google_patents_research	KI-angereicherte Daten: Maschinelle Übersetzungen (Titel/Abstracts), Vektor-Embeddings, Ähnlichkeitscluster.	Google Research	Ermöglicht semantische Suche ("Concept Search") statt reiner Stichwortsuche. Überwindet Sprachbarrieren durch englische Übersetzungen globaler Patente.[7]
patents-public-data.uspto_oce_*	Ökonomische Prozessdaten: Office Actions (Prüfungsbescheide), Litigation (Gerichtsverfahren), Patent Claims, Assignments.	USPTO Office of Chief Economist	Dient der Analyse des Patentwerts, der Verfahrensdauer ("Prosecution Analytics") und der rechtlichen Risikobewertung.[8]
patents-public-data.cpc	Klassifikationsschemata: Definitionen und Hierarchien der Cooperative Patent Classification.	EPO / USPTO	Ermöglicht das Verständnis der technologischen Taxonomie und hierarchische Aggregationen (z.B. Roll-up von Subgruppen auf Hauptklassen).[10]
patents-public-data.ebi_chembl	Chemische Bioaktivitätsdaten	EMBL-EBI	Kritisch für die pharmazeutische

	verknüpft mit Patentdokumenten.		Forschung, um Patente nicht nach Text, sondern nach chemischen Molekülstrukturen zu durchsuchen.[2]

Diese Fragmentierung erfordert vom Analysten oft das „Joinen“ (Verknüpfen) von Tabellen über verschiedene Datasets hinweg, um ein vollständiges Bild zu erhalten. Ein typisches Szenario wäre die Verknüpfung der globalen Zitationsdaten aus patents.publications mit den detaillierten, disambiguierten Erfinderdaten aus patentsview, um die Mobilität von US-Erfindern und deren Einfluss auf globale Zitationsnetzwerke zu analysieren. Die Interoperabilität dieser Tabellen wird durch standardisierte Schlüssel wie die Publikationsnummer (publication_number) gewährleistet, wobei oft Formatierungsunterschiede zu beachten sind.[5]

---

3. Analyse des globalen Kern-Datensatzes: patents-public-data.patents

Der Datensatz patents-public-data.patents.publications stellt das Rückgrat der meisten Patentanalysen in BigQuery dar. Bereitgestellt durch IFI CLAIMS Patent Services, umfasst er bibliografische Daten zu über 98 Millionen Patentpublikationen aus über 100 Ländern.[5] Die Tiefe und Struktur dieses Datensatzes erfordert eine detaillierte Betrachtung.

3.1 Das Konzept der Verschachtelung (Nested & Repeated Fields)

Eine der größten Hürden für SQL-Einsteiger, aber gleichzeitig eines der mächtigsten Features von BigQuery, ist die Nutzung von verschachtelten und wiederholten Feldern (RECORD und REPEATED). Patentdaten sind inhärent hierarchisch und nicht flach: Ein Patent hat nicht nur *einen* Erfinder, sondern oft mehrere. Es hat nicht nur *einen* Klassifikationscode, sondern oft Dutzende. Es zitiert nicht nur *ein* Dokument, sondern Hunderte.
In einer traditionellen relationalen Datenbank würde dies Dutzende von Verknüpfungstabellen (Join-Tables) erfordern (z.B. eine Tabelle patent_inventors, eine Tabelle patent_cpcs). BigQuery hingegen speichert diese Listen direkt in der Zeile des Patents.

- **Implikation:** Dies reduziert den Speicherbedarf und erhöht die Abfragegeschwindigkeit, da keine teuren Joins nötig sind, um alle Attribute eines Patents zu lesen.
- **Herausforderung:** Analysten müssen lernen, diese Strukturen mittels UNNEST zu "entpacken", um Aggregationen durchzuführen.[11]

3.2 Die Anatomie der Metadaten

Die Tabelle publications enthält eine Vielzahl kritischer Felder, deren korrekte Interpretation über den Erfolg einer Analyse entscheidet:

- **Identifikatoren (publication_number vs. application_number):** Die publication_number (z.B. 'US-7650331-B1') ist der primäre Schlüssel und DOCDB-kompatibel. Sie kennzeichnet das physische Dokument. Die application_number hingegen kennzeichnet den rechtlichen Akt der Anmeldung. Für Prozessanalysen (z.B. "Wie lange dauerte die Prüfung?") ist die Anmeldenummer entscheidend, für Technik-Analysen ("Was steht drin?") die Publikationsnummer.[5]
- **Patentfamilien (family_id):** Ein oft übersehenes, aber kritisches Feld für ökonomische Analysen ist die family_id. Patente werden oft in vielen Ländern gleichzeitig angemeldet (z.B. USA, Europa, Japan), um denselben Erfindungsgedanken global zu schützen. Zählt man einfach alle Publikationen, erhält man massive Doppelzählungen. Die family_id gruppiert alle Publikationen, die zur selben „Simple Family“ gehören (d.h. dieselbe Priorität teilen). Wenn man die Innovationskraft eines Unternehmens misst, sollte man in der Regel Patentfamilien zählen, nicht einzelne Publikationen.[5]
- **Klassifikationen (cpc):** Das cpc-Feld ist ein Array von Structs, das nicht nur den Code (z.B. "H04L"), sondern auch Metadaten enthält, wie z.B. ob der Code "inventive" (erfinderisch) oder nur "additional" (zusätzliche Information) ist. Diese Unterscheidung erlaubt feinere Analysen des technologischen Kerns einer Erfindung.[11]
- **Zitationen (citation):** Dieses Array enthält sowohl Vorwärts- als auch Rückwärtszitationen. Es ist die Basis für Netzwerkanalysen. Ein Patent, das häufig von jüngeren Patenten zitiert wird (hohe Forward Citations), gilt ökonomisch als wertvoller und technologisch als einflussreicher. Die Unterscheidung in der Zitationsquelle (z.B. vom Prüfer vs. vom Anmelder hinzugefügt) kann Aufschluss über die Qualität der Prüfung und die Marktkenntnis des Anmelders geben.

3.3 Temporale Dimensionen und Datumsformate

Eine technische Besonderheit des Datensatzes, die zu häufigen Fehlern führt, ist die Speicherung von Datumsfeldern. Felder wie filing_date (Anmeldedatum), publication_date (Veröffentlichungsdatum) und grant_date (Erteilungsdatum) sind oft als **Integer** im Format YYYYMMDD gespeichert, nicht als nativer SQL-Datentyp DATE.[12]

- **Analytische Konsequenz:** Für Zeitreihenanalysen müssen diese Felder transformiert werden. Eine einfache Jahresanalyse kann durch arithmetische Operationen (FLOOR(filing_date / 10000)) erfolgen, was performanter ist als String-Parsing. Für präzise Zeitberechnungen (z.B. Laufzeitberechnung) ist ein Casting notwendig: PARSE_DATE('%Y%m%d', CAST(publication_date AS STRING)).[13]
- **Strategische Bedeutung:** Der Unterschied zwischen filing_date und publication_date ist die "Verzögerung" der Sichtbarkeit. Da Patentanmeldungen in der Regel erst 18 Monate nach Anmeldung veröffentlicht werden, haben alle Analysen auf Basis öffentlicher Daten einen blinden Fleck von 1,5 Jahren. Dies muss bei Trendanalysen ("Ist Technologie X im Rückgang?") zwingend berücksichtigt werden, um Fehlinterpretationen des jüngsten Datenrückgangs zu vermeiden.

---

4. Der US-Fokus: USPTO PatentsView und Entitäten-Disambiguierung

Während der Google-Datensatz durch seine globale Breite besticht, bietet **PatentsView** (patents-public-data.patentsview) eine unübertroffene Tiefe und Datenqualität für den US-amerikanischen Raum (Daten ab 1976). PatentsView ist eine Initiative des USPTO, die sich speziell der Lösung eines der schwierigsten Probleme der Patentdatenanalyse widmet: der **Disambiguierung von Entitäten**.[6]

4.1 Die Herausforderung der Namensvarianz

In Rohdaten variieren Namen von Personen und Unternehmen massiv. Ein Unternehmen wie "International Business Machines" kann als "IBM", "I.B.M. Corp.", "Intl. Bus. Mach." oder mit Tippfehlern auftreten. Ohne Bereinigung würden Analysen des Patentportfolios von IBM fragmentiert und unvollständig bleiben. PatentsView verwendet komplexe Algorithmen (teilweise basierend auf Diskriminanzanalyse und Clustering), um diese Varianten eindeutigen, persistenten Identifikatoren zuzuordnen.

4.2 Die Architektur der PatentsView-Tabellen

PatentsView folgt einem stark relationalen Schema (ähnlich einem Sternschema), im Gegensatz zur verschachtelten Struktur der Google-Haupttabellen. Dies erfordert ein Umdenken bei der Erstellung von Abfragen.

- **Tabelle inventor:** Diese Tabelle ist das Herzstück der Personenforschung. Sie enthält eindeutige ids für Erfinder. Dies ermöglicht es, die Karriere eines Erfinders über Jahrzehnte und verschiedene Arbeitgeber hinweg zu verfolgen ("Inventor Mobility"). Man kann analysieren, wie Wissen zwischen Firmen fließt, wenn Schlüsselerfinder wechseln.
- **Tabelle assignee:** Hier werden die Inhaber der Patente harmonisiert. Ein entscheidendes Feld ist type, das klassifiziert, ob es sich um ein US-Unternehmen, ein ausländisches Unternehmen, eine Regierungsbehörde oder eine Privatperson handelt.[14] Dies ermöglicht strukturelle Analysen des Innovationssystems (z.B. Anteil staatlicher vs. privater Innovation).
- **Tabelle location:** Diese Tabelle verknüpft Erfinder und Anmelder mit präzisen geografischen Koordinaten (latitude, longitude), Städten und Staaten.
    - *Strategischer Nutzen:* Dies ermöglicht hochauflösende regionale Cluster-Analysen. Man kann "Heatmaps" erstellen, die zeigen, wo genau in den USA Biotechnologie-Cluster entstehen, und wie diese mit lokalen Universitäten korrelieren.[15]
- **Linking-Tabellen (Bridge Tables):** Da die Beziehung zwischen Patenten und Erfindern eine n:m-Beziehung ist (ein Patent hat viele Erfinder, ein Erfinder hat viele Patente), nutzt PatentsView Brückentabellen wie patent_inventor oder patent_assignee. Um eine Abfrage zu erstellen, die "alle Patente von Erfindern aus Boston" findet, muss man

location -> location_assignee (oder inventor) -> patent_inventor -> patent verknüpfen.

4.3 Integration und Interoperabilität

Ein wichtiger technischer Hinweis für die Arbeit mit PatentsView in BigQuery ist, dass die IDs oft eine andere Struktur aufweisen als im globalen Google-Datensatz. PatentsView nutzt oft reine Nummern (z.B. 7650331) ohne Ländercode, während Google den ISO-Code voranstellt (US-7650331-B2). Um Daten aus beiden Welten zu kombinieren – etwa um die detaillierten Erfinderdaten aus PatentsView mit den globalen Familieninformationen von Google zu nutzen – ist oft eine Transformation der IDs oder die Nutzung der Tabelle patentsview.match notwendig, die als Übersetzungstabelle fungiert.[5]

---

5. USPTO OCE: Die Ökonomie der Prüfung und Rechtsdurchsetzung

Für Analysten, die sich nicht nur für die technische Seite einer Erfindung interessieren, sondern für deren *ökonomischen Wert*, *Durchsetzbarkeit* und den *Verlauf des Prüfungsprozesses*, sind die Datensätze des **Office of the Chief Economist (OCE)** des USPTO eine Goldgrube.[8] Diese Daten beleuchten den administrativen und rechtlichen Lebenszyklus eines Patents.

5.1 Office Action Research Dataset: Die Black Box der Prüfung

Ein Patent wird nicht einfach angemeldet und erteilt; es durchläuft einen oft jahrelangen, iterativen Prozess der Prüfung ("Prosecution"), der durch einen Austausch von Bescheiden ("Office Actions") zwischen Prüfer und Anwalt gekennzeichnet ist. Das uspto_oce_office_actions Dataset öffnet diese "Black Box".

- **Prozessanalyse (Prosecution Analytics):** Analysten können detailliert untersuchen, wie Prüfer auf Anmeldungen reagieren.
    - *Time-to-Grant:* Wie lange dauert es durchschnittlich in einer bestimmten Technologieklasse (z.B. Halbleiter vs. Pharma), bis ein Patent erteilt wird? Diese Information ist vital für die Budgetplanung von R&D-Abteilungen.
    - *Examiner Analytics:* Die Daten erlauben die Profilierung von Prüfern. Gibt es Prüfer, die statistisch signifikant strenger sind als andere ("Hard Examiners")? Wie hoch ist die Wahrscheinlichkeit einer Zurückweisung ("Rejection") bei einer bestimmten Art Unit?
- **Ablehnungsgründe:** Die Daten kodieren die Gründe für Zurückweisungen, basierend auf dem US-Patentrecht (z.B. 35 U.S.C. § 102 für Neuheit, § 103 für Offensichtlichkeit).
    - *Strategische Implikation:* Eine Analyse, die zeigt, dass in einem bestimmten Technologiefeld 80% der Anmeldungen aufgrund von § 103 (Offensichtlichkeit) scheitern, signalisiert, dass das Feld "überlaufen" ist (crowded art) und inkrementelle Innovationen schwer zu schützen sind.[2]

5.2 Patent Claims Research Data: Der Schutzbereich

Der rechtliche Wert eines Patents wird fast ausschließlich durch seine Ansprüche ("Claims") definiert, nicht durch die Beschreibung. Das Dataset uspto_oce_claims enthält die Volltexte und Metadaten dieser Ansprüche.

- **Scope-Analyse:** Änderungen in der Wortlänge oder der Anzahl der Ansprüche während des Prüfungsverfahrens können auf eine Einschränkung des Schutzbereichs hindeuten. Ein Patent, das mit 20 breiten Ansprüchen beginnt und mit 5 sehr spezifischen, langen Ansprüchen erteilt wird, hat oft massiv an Wert verloren. Die Analyse der "Independent Claims" (unabhängige Ansprüche) vs. "Dependent Claims" hilft, die Breite eines Patents algorithmisch abzuschätzen.[8]

5.3 Litigation Data: Das Risiko-Radar

Das Dataset uspto_oce_litigation enthält strukturierte Daten zu Patentstreitigkeiten an US-Bezirksgerichten.

- **Litigation Risk Profiling:** Durch die Verknüpfung dieser Daten mit Patent-Metadaten lassen sich Risikomodelle erstellen. Welche Technologien sind besonders streitanfällig? Welche Unternehmen agieren als Kläger, welche als Beklagte?
- **NPE-Erkennung:** Die Daten helfen bei der Identifikation von "Non-Practicing Entities" (NPEs oder Patent Trolls), indem man Muster in der Klagehäufigkeit und dem Portfolio-Besitz analysiert. Dies ist essenziell für die "Freedom to Operate" (FTO) Analyse von Unternehmen, die in neue Märkte eintreten.[16]

---

6. Der semantische Wandel: KI, Embeddings und Vektor-Suche

Ein revolutionärer Aspekt der in BigQuery verfügbaren Patentdaten ist die Integration von maschinellem Lernen und **Vektor-Embeddings** in der Tabelle patents-public-data.google_patents_research.publications.[5] Dies markiert den Übergang von der syntaktischen Suche (Wörter) zur semantischen Suche (Bedeutung).

6.1 Die Grenzen der Booleschen Suche

Klassische Patentrecherchen basieren auf komplexen Booleschen Ketten (z.B. "((Drone OR UAV) AND (Rotary OR Propeller))"). Dieses Verfahren hat zwei fundamentale Schwächen:

1. **Vokabular-Mismatch:** Patentanwälte nutzen oft bewusst abstrakte oder obskure Sprache ("Aerial Vehicle with rotatable propulsors" statt "Helicopter"), um Begriffe zu verschleiern oder den Schutzbereich zu maximieren (Lexikographen-Privileg). Keyword-Suchen übersehen daher oft relevante Dokumente (False Negatives).
2. **Kontext-Blindheit:** Ein Wort wie "Jaguar" kann ein Auto, ein Tier oder eine Software-Version bedeuten. Stichwortsuchen liefern daher oft irrelevante Ergebnisse (False Positives).

6.2 Die Lösung: Embeddings im Vektorraum

Google hat maschinelle Lernmodelle auf den riesigen Korpus von Patenttexten trainiert, um den *Inhalt* eines Patents in einen Vektor (eine Reihe von 64 Fließkommazahlen, gespeichert im Feld embedding_v1) zu transformieren.[17] In diesem hochdimensionalen Vektorraum liegen Patente mit ähnlichem technischem Inhalt geometrisch nahe beieinander, selbst wenn sie völlig unterschiedliche Wörter verwenden.

- **Trainingsbasis:** Die Embeddings basieren auf Modellen, die ursprünglich darauf trainiert wurden, CPC-Klassifikationen aus dem Text vorherzusagen (z.B. WSABIE-Modelle). Sie haben also "gelernt", welche technischen Konzepte zu welchen Technologieklassen gehören.[5]
- **Vektor-Suche in BigQuery:** Mit der Einführung von VECTOR_SEARCH und Vektor-Indizes in BigQuery kann dieser Ansatz nun nativ skaliert werden. Anstatt Vektoren nach Python zu exportieren, kann die Suche direkt in der Datenbank erfolgen.
    - *Anwendungsszenario:* Ein Forscher gibt einen Absatz technischer Beschreibung ein (oder wählt ein "Seed-Patent"). Das System generiert daraus einen Vektor und sucht mittels Cosinus-Ähnlichkeit (Cosine Distance) die nächsten Nachbarn im 100-Millionen-Dokumente-Raum. Dies fördert oft "Prior Art" zutage, die durch Keywords unauffindbar war, weil sie in einer anderen Sprache verfasst wurde oder andere Terminologie nutzte.[18]

6.3 Multilinguale Intelligenz

Ein weiterer Aspekt der google_patents_research Tabelle ist die Überwindung von Sprachbarrieren. Die Tabelle enthält maschinelle Übersetzungen (ins Englische) für Titel und Zusammenfassungen von nicht-englischen Patenten. Dies ermöglicht eine globale Volltextsuche, die besonders für die Analyse des asiatischen Marktes (China, Japan, Korea) essenziell ist. Da China inzwischen jährlich die meisten Patentanmeldungen weltweit generiert, wäre eine Analyse ohne diese Übersetzungen blind für einen Großteil der globalen Innovationstätigkeit.[7]

---

7. Chemische Intelligenz: Die Verbindung zur Life Science

Innovation findet nicht nur im Text statt, sondern auch in der Molekülstruktur. Für die pharmazeutische und chemische Industrie ist die Integration von **PubChem** und **ChEMBL** Daten in BigQuery ein entscheidender Vorteil.

- **Struktur-Suche:** Patente beschreiben chemische Verbindungen oft generisch (Markush-Strukturen) oder durch IUPAC-Namen, die schwer zu suchen sind. Google und Partner haben Millionen von chemischen Strukturen aus Patenten extrahiert und mit Datenbanken wie PubChem verknüpft (patents-public-data.ebi_chembl oder ncbi-research-pubchem).

- **Strategischer Join:** Durch die Verknüpfung der Patent-Metadaten mit den bioaktiven Daten aus ChEMBL können Analysten Fragen beantworten wie: "Welche Unternehmen halten Patente auf Verbindungen, die ähnlich zu meinem Wirkstoffkandidaten sind?" oder "Welche Patente decken Moleküle ab, die eine hohe Affinität zu Protein X haben?". Dies ermöglicht eine "FTO-Analyse" (Freedom to Operate) auf molekularer Ebene, lange bevor klinische Studien beginnen.[2]

---

8. Technische Implementierung: SQL-Strategien und Best Practices

Die Arbeit mit diesen Datenmengen erfordert fortgeschrittene SQL-Techniken. Die naive Herangehensweise ("SELECT * FROM table") führt bei Terabyte-großen Tabellen zu ineffizienten und teuren Abfragen. Im Folgenden werden zentrale Entwurfsmuster (Design Patterns) vorgestellt, die für effiziente Patentanalysen in BigQuery unerlässlich sind.

8.1 Partitionierung und Pruning: Die Kostenbremse

Da BigQuery nach verarbeiteter Datenmenge abrechnet, ist das Ziel jeder Query, so wenig Daten wie möglich zu scannen.

- **Partitionierung:** Die großen Tabellen sind oft nach Datum partitioniert (z.B. publication_date). Eine WHERE-Klausel, die auf das Datum filtert (z.B. publication_date > 20200101), sorgt dafür, dass BigQuery nur die relevanten Partitionen liest und den Rest ignoriert (Partition Pruning). Dies kann die Kosten einer Abfrage um 99% senken.[22]
- **Spalten-Projektion:** Da BigQuery ein spaltenorientierter Speicher (Columnar Store) ist, ist das Lesen einer einzigen Spalte sehr billig, das Lesen aller Spalten (SELECT *) sehr teuer. Analysten sollten *niemals* SELECT * verwenden, wenn sie nur an Metadaten interessiert sind, da dies riesige Textfelder wie description oder claims mitlädt, die Terabytes an Daten umfassen können.[23]

8.2 Das UNNEST-Pattern: Umgang mit Arrays

Wie in Abschnitt 3.1 beschrieben, sind viele Daten als Arrays gespeichert. Um diese Daten zu aggregieren (z.B. "Zähle die Top-Anmelder pro CPC-Klasse"), muss man das Array "flachklopfen" (flatten). Der Standard-SQL-Befehl hierfür ist CROSS JOIN UNNEST.
Beispiel-Szenario: Identifikation der führenden Unternehmen im Bereich "Quantum Computing" (definiert durch CPC-Codes, die mit 'G06N10' beginnen).

SQL

```
SELECT
  -- Extrahiere den harmonisierten Namen des Anmelders
  assignee.name AS Company_Name,
```

```
  -- Zähle die Anzahl der eindeutigen Patentfamilien (nicht Publikationen!)
  COUNT(DISTINCT pubs.family_id) AS Family_Count
FROM
  `patents-public-data.patents.publications` AS pubs,
  -- Entpacke das CPC-Array, um auf einzelne Codes zuzugreifen
  UNNEST(pubs.cpc) AS cpc_code,
  -- Entpacke das Assignee-Array, um auf Anmelder zuzugreifen
  UNNEST(pubs.assignee_harmonized) AS assignee
WHERE
  -- Filtere auf den gewünschten Technologiebereich
  cpc_code.code LIKE 'G06N10%'
  -- Betrachte nur Anmeldungen der letzten 10 Jahre
  AND pubs.filing_date >= 20140101
  -- Nur US-Patente für dieses Beispiel
  AND pubs.country_code = 'US'
GROUP BY
  Company_Name
ORDER BY
  Family_Count DESC
LIMIT 20;
```

Dieses Pattern – FROM table, UNNEST(field) – ist der Standardmechanismus für fast alle tiefgehenden Analysen in BigQuery.[11] Es erzeugt temporär eine Tabelle, in der jede Kombination aus Patent und CPC-Code eine eigene Zeile ist, was Filterungen und Gruppierungen ermöglicht.

8.3 Semantische Ähnlichkeitssuche (SQL Implementierung)

Die Nutzung der Embeddings für eine Ähnlichkeitssuche ("Finde Patente, die diesem ähnlich sind") kann direkt in SQL erfolgen, indem man das Skalarprodukt (Dot Product) oder die Cosinus-Ähnlichkeit berechnet.

Beispiel-Logik:

1. Extrahiere den Vektor (embedding_v1) eines Zielpatents (Seed).
2. Verbinde diesen Vektor mit allen anderen Vektoren in der Datenbank.
3. Berechne die Ähnlichkeit. Da die Vektoren oft normalisiert sind, entspricht die Cosinus-Ähnlichkeit dem Skalarprodukt.

SQL

```
WITH TargetPatent AS (
  SELECT embedding_v1 AS target_vec
  FROM `patents-public-data.google_patents_research.publications`
  WHERE publication_number = 'US-9000000-B2'
```

```
)
SELECT
  candidate.publication_number,
  -- Manuelle Berechnung des Dot Products als Maß für Ähnlichkeit
  (
    SELECT SUM(t * c)
    FROM UNNEST(target.target_vec) t WITH OFFSET i
    JOIN UNNEST(candidate.embedding_v1) c WITH OFFSET j
    ON i = j
  ) AS similarity_score
FROM
  TargetPatent target,
  `patents-public-data.google_patents_research.publications` candidate
WHERE
  -- Vorfilterung zur Performance-Steigerung (z.B. nur gleiches CPC-Feld)
  EXISTS (SELECT 1 FROM UNNEST(candidate.top_terms) term WHERE term = 'neural network')
ORDER BY
  similarity_score DESC
LIMIT 50;
```

Hinweis: Für sehr große Datenmengen ist die Funktion VECTOR_SEARCH (Teil von BigQuery Vector Search) performanter, da sie spezialisierte Indizes (wie IVFFlat) nutzt, anstatt eine Brute-Force-Berechnung durchzuführen.[20]

---

9. Governance, Kostenmanagement und Data Sharing

Die Nutzung von Public Datasets ist zwar im Zugriff "kostenlos", die Verarbeitung (Compute) jedoch nicht. Daher ist Governance entscheidend.

9.1 Kostenkontrolle

BigQuery bietet verschiedene Preismodelle. Das "On-Demand"-Modell (ca. $5 pro TB Scan) ist flexibel, kann aber bei unvorsichtigen Abfragen teuer werden.

- **Best Practice:** Setzen von Quotas auf Projekt- oder Nutzerebene (z.B. max. 10 TB pro Tag).
- **Slot-Reservierung:** Für Unternehmen mit hohem, vorhersehbarem Aufkommen lohnt sich oft der Wechsel zu "Editions" (Pauschale für Rechenkapazität/Slots), was die Kosten deckelt und planbar macht.[24]
- **Storage-Kosten:** Daten, die länger als 90 Tage nicht modifiziert wurden, fallen automatisch in den günstigeren "Long-Term Storage" Tarif (ca. 50% Rabatt). Da Public Datasets von Google bezahlt werden, betrifft dies vor allem die *eigenen* Ergebnistabellen, die man speichert.[23]

9.2 Data Sharing: Analytics Hub vs. Authorized Views

Wie teilt man Erkenntnisse sicher innerhalb eines Konzerns oder mit Partnern?

- **Authorized Views:** Erlauben es, einem Nutzer Zugriff auf das *Ergebnis* einer Abfrage zu geben, ohne ihm Zugriff auf die zugrundeliegenden Rohdaten zu gewähren. Dies ist ideal, um z.B. dem Management ein Dashboard bereitzustellen, ohne direkten Datenbankzugriff zu erlauben.[25]
- **Analytics Hub:** Eine modernere Lösung für den Datenaustausch. Unternehmen können "Data Exchanges" erstellen. Ein Datensatz (z.B. eine kuratierte Liste von Wettbewerber-Patenten) kann dort "publiziert" werden. Abonnenten erhalten einen "Linked Dataset" in ihrem eigenen Projekt. Der Vorteil: Die Daten werden nicht kopiert (keine Redundanz), bleiben aber immer aktuell. Dies ermöglicht eine zentrale Governance bei dezentraler Nutzung.[26]

---

10. Strategische Anwendungsfälle (Use Cases) und Ausblick

Die Kombination der technischen Infrastruktur und der Datenvielfalt ermöglicht völlig neue strategische Anwendungen.

10.1 White Space Analysis (Innovationslücken)

Durch Clustering der Vektor-Embeddings lassen sich "Landkarten" der Technologie erstellen. Dichte Cluster repräsentieren etablierte Felder ("Red Ocean"). Die leeren Räume dazwischen ("White Spaces") können auf ungenutzte Innovationspotenziale hinweisen – oder auf technologische Sackgassen. Die Kombination von Embeddings mit zeitlichen Daten erlaubt es, die *Entstehung* neuer Cluster in Echtzeit zu beobachten (Emerging Technology Detection).[17]

10.2 M&A Due Diligence und Wettbewerber-Radar

Vor einer Firmenübernahme kann das Patentportfolio des Ziels in Minuten analysiert werden.

- **Qualität:** Wie oft werden die Patente zitiert? (Forward Citations)
- **Abdeckung:** Deckt das Portfolio wirklich die Produkte ab? (Vergleich von Produktbeschreibungen via Embeddings mit Patentansprüchen).
- **Risiko:** Gibt es anhängige Rechtsstreitigkeiten? (Litigation Data).
- **Abhängigkeit:** Zitiert das Zielunternehmen stark die Patente eines Konkurrenten? Dies könnte auf Lizenzbedarf hinweisen.

10.3 Fazit: Die Zukunft ist hybrid

Die Verfügbarkeit der Google Patents Public Datasets in BigQuery markiert einen Paradigmenwechsel. Die Hürden für den Zugang zu globalen Innovationsdaten wurden drastisch gesenkt. Was früher teure Spezialsoftware und Wochen an Arbeit erforderte, kann heute mit Standard-SQL und Cloud-Ressourcen in Minuten gelöst werden.

Die Zukunft der Patentanalyse liegt in der **hybriden Intelligenz**: Die Kombination aus strukturierter Analyse (SQL auf Metadaten für Trends und Statistiken) und unstrukturierter KI-Analyse (Vector Search und Large Language Models auf Volltexten für inhaltliches Verständnis). Mit der Integration von Modellen wie Gemini direkt in die Vertex AI und BigQuery Plattform bewegt sich die Patentanalyse weg vom reinen "Suchen" hin zum intelligenten "Generieren" von Insights – etwa dem automatischen Zusammenfassen von Patent-Claims, der Generierung von Invaliditätsargumenten oder der automatisierten Überwachung von Wettbewerbern. Für Unternehmen bedeutet dies: Wer diese Datenquellen beherrscht, besitzt einen signifikanten Informationsvorsprung im globalen Wettbewerb um die Technologien von morgen.

Referenzen

1. Patent PDF Samples with Extracted Structured Data – Marketplace - Google Cloud Console, Zugriff am Januar 18, 2026, https://console.cloud.google.com/marketplace/product/global-patents/labeled-patents
2. Google Patents Public Datasets: connecting public, paid, and private patent data, Zugriff am Januar 18, 2026, https://cloud.google.com/blog/topics/public-datasets/google-patents-public-datasets-connecting-public-paid-and-private-patent-data
3. BigQuery public datasets - Google Cloud Documentation, Zugriff am Januar 18, 2026, https://docs.cloud.google.com/bigquery/public-data
4. Programmatic Patent Searches Using Google's BigQuery & Public Patent Data, Zugriff am Januar 18, 2026, https://www.aipla.org/list/innovate-articles/programmatic-patent-searches-using-google-s-bigquery-public-patent-data
5. patents-public-data/tables/dataset_Google Patents Public Datasets.md at master · google/patents-public-data - GitHub, Zugriff am Januar 18, 2026, https://github.com/google/patents-public-data/blob/master/tables/dataset_Google%20Patents%20Public%20Datasets.md
6. PatentsView Data - Kaggle, Zugriff am Januar 18, 2026, https://www.kaggle.com/datasets/bigquery/patentsview
7. Google Patents Research Data – Marketplace, Zugriff am Januar 18, 2026, https://console.cloud.google.com/marketplace/product/google_patents_public_datasets/google-patents-research-data?hl=en-GB
8. USPTO OCE Patent Claims Research Data – Marketplace - Google Cloud Console, Zugriff am Januar 18, 2026, https://console.cloud.google.com/marketplace/product/google_patents_public_datasets/uspto-oce-claims
9. USPTO OCE Office Actions Data – Marketplace - Google Cloud Console, Zugriff am Januar 18, 2026, https://console.cloud.google.com/marketplace/product/google_patents_public_datasets/uspto-oce-office-actions
10. patents-public-data/tables/dataset_Other.md at master · google, Zugriff am

Januar 18, 2026, https://github.com/google/patents-public-data/blob/master/tables/dataset_Other.md
11. Google Public Patent Data SQL (BigQuery) - Stack Overflow, Zugriff am Januar 18, 2026, https://stackoverflow.com/questions/68325713/google-public-patent-data-sql-bigquery
12. Programmatic Patent Searches Using Google's BigQuery & Public Patent Data - Medium, Zugriff am Januar 18, 2026, https://medium.com/@AlphaDataIQ/programmatic-patent-searches-using-googles-bigquery-public-patent-data-293adad3d30c
13. Big Data & Public Databases for Patent Research and Analysis - Intellectual Property Owners Association, Zugriff am Januar 18, 2026, https://ipo.org/wp-content/uploads/2018/09/Patent-Searching-Public-Databases.pdf
14. BigQuery patent dataset documentation - Kaggle, Zugriff am Januar 18, 2026, https://www.kaggle.com/code/wetherbeei/bigquery-patent-dataset-documentation
15. patents-public-data/tables/index.md at master - GitHub, Zugriff am Januar 18, 2026, https://github.com/google/patents-public-data/blob/master/tables/index.md
16. USPTO OCE Patent Litigation Docket Reports Data - Kaggle, Zugriff am Januar 18, 2026, https://www.kaggle.com/datasets/bigquery/uspto-oce-litigation
17. Expanding your patent set with ML and BigQuery | Google Cloud Blog, Zugriff am Januar 18, 2026, https://cloud.google.com/blog/products/data-analytics/expanding-your-patent-set-with-ml-and-bigquery
18. Build a Patent Search App with AlloyDB, Vector Search & Vertex AI! | Google Codelabs, Zugriff am Januar 18, 2026, https://codelabs.developers.google.com/patent-search-alloydb-gemini
19. Search embeddings with vector search | BigQuery - Google Cloud Documentation, Zugriff am Januar 18, 2026, https://docs.cloud.google.com/bigquery/docs/vector-search
20. BigQuery vector search now in preview. | Google Cloud Blog, Zugriff am Januar 18, 2026, https://cloud.google.com/blog/products/data-analytics/introducing-new-vector-search-capabilities-in-bigquery
21. Docs - PubChem - NIH, Zugriff am Januar 18, 2026, https://pubchem.ncbi.nlm.nih.gov/docs/
22. Querying the Google Patent Data on Big Query processes way too much data, Zugriff am Januar 18, 2026, https://stackoverflow.com/questions/79329120/querying-the-google-patent-data-on-big-query-processes-way-too-much-data
23. Cost optimization best practices for BigQuery | Google Cloud Blog, Zugriff am Januar 18, 2026,

https://cloud.google.com/blog/products/data-analytics/cost-optimization-best-practices-for-bigquery
24. Optimizing computational analysis costs with BigQuery editions | Google Cloud Blog, Zugriff am Januar 18, 2026, https://cloud.google.com/blog/products/data-analytics/optimizing-computational-analysis-costs-with-bigquery-editions/
25. Authorized views | BigQuery - Google Cloud Documentation, Zugriff am Januar 18, 2026, https://docs.cloud.google.com/bigquery/docs/authorized-views
26. BigQuery data sharing | Analytics Hub - Google Cloud, Zugriff am Januar 18, 2026, https://cloud.google.com/analytics-hub
27. Introduction to BigQuery sharing - Google Cloud Documentation, Zugriff am Januar 18, 2026, https://docs.cloud.google.com/bigquery/docs/analytics-hub-introduction