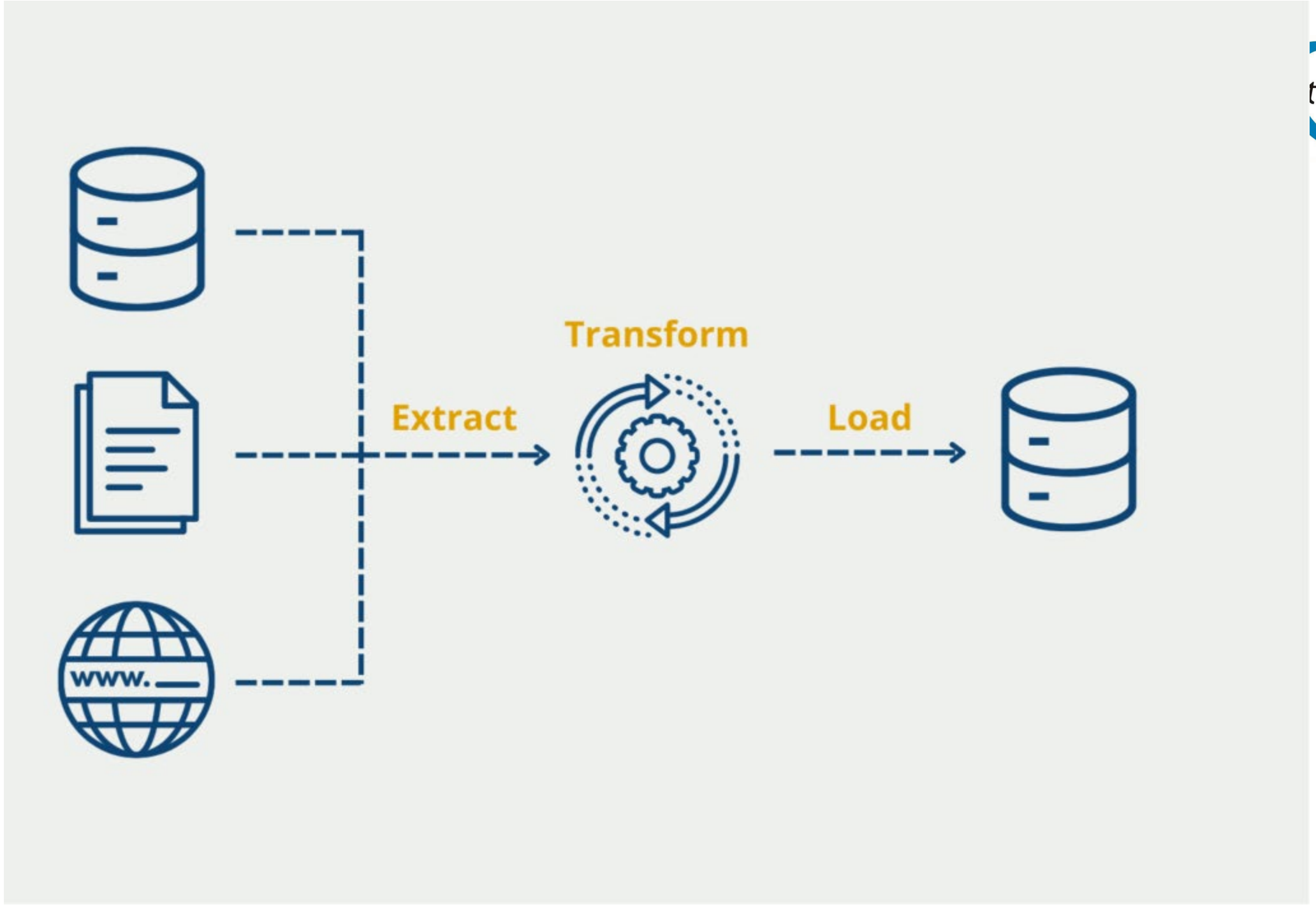
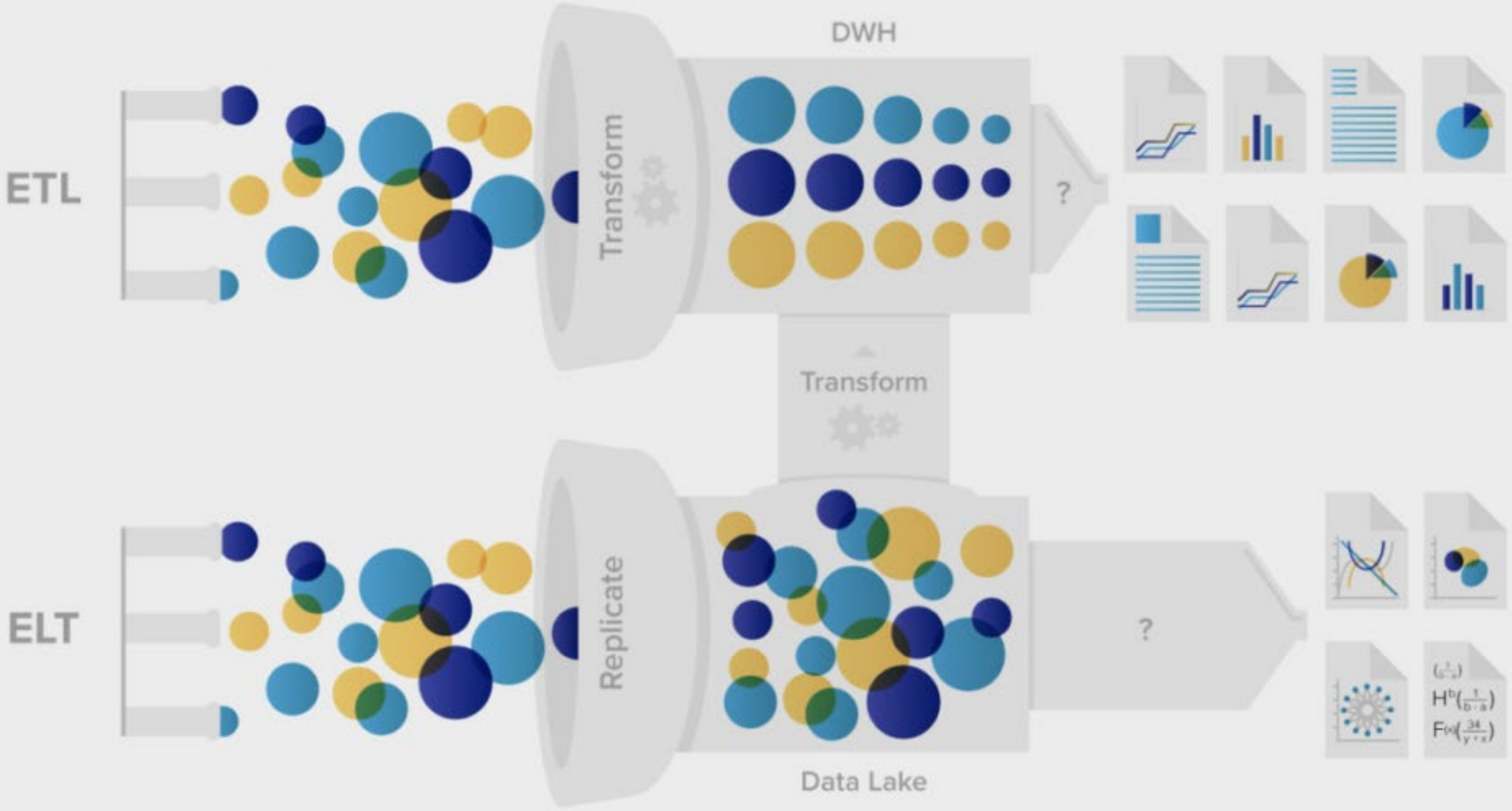
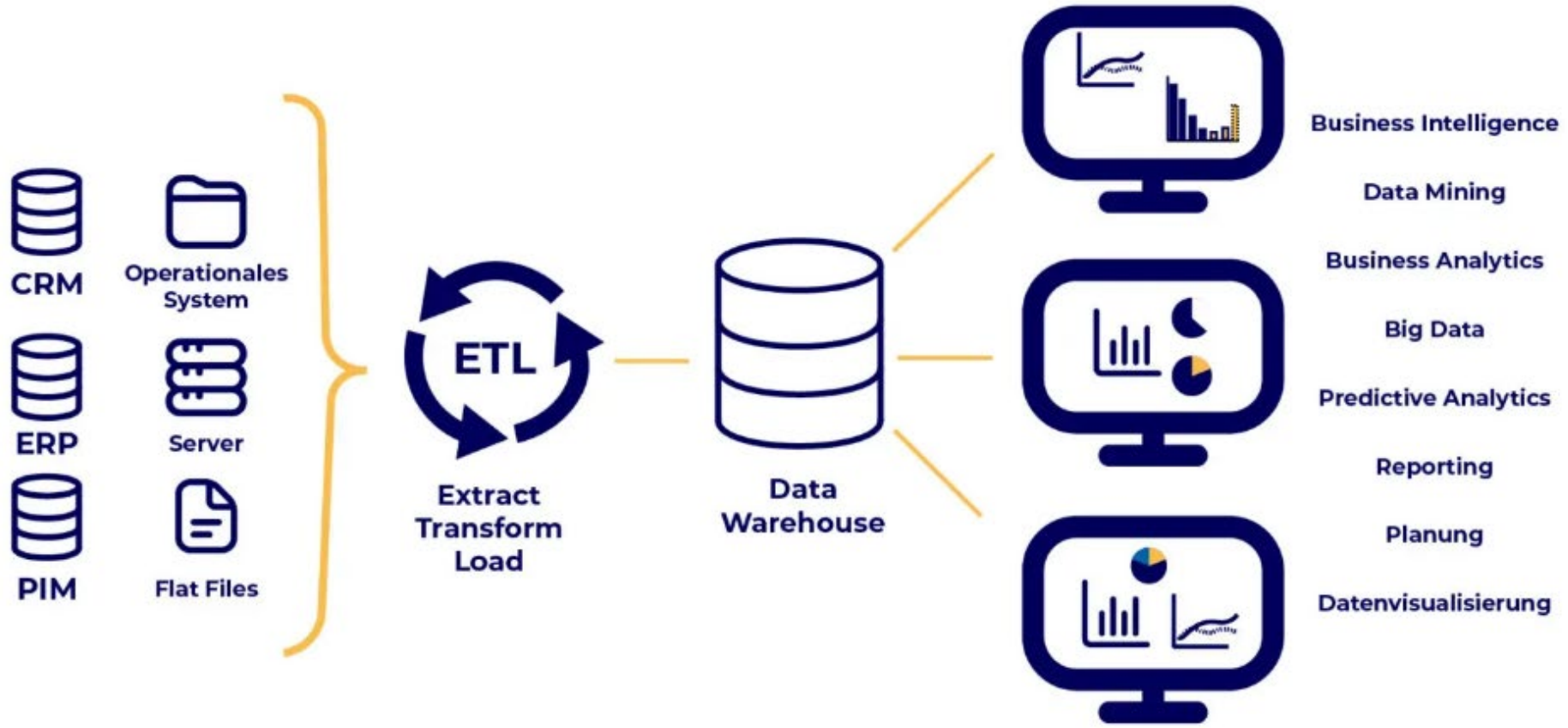
Data Engineer

ETL:


Transform
Extract
Load
www.

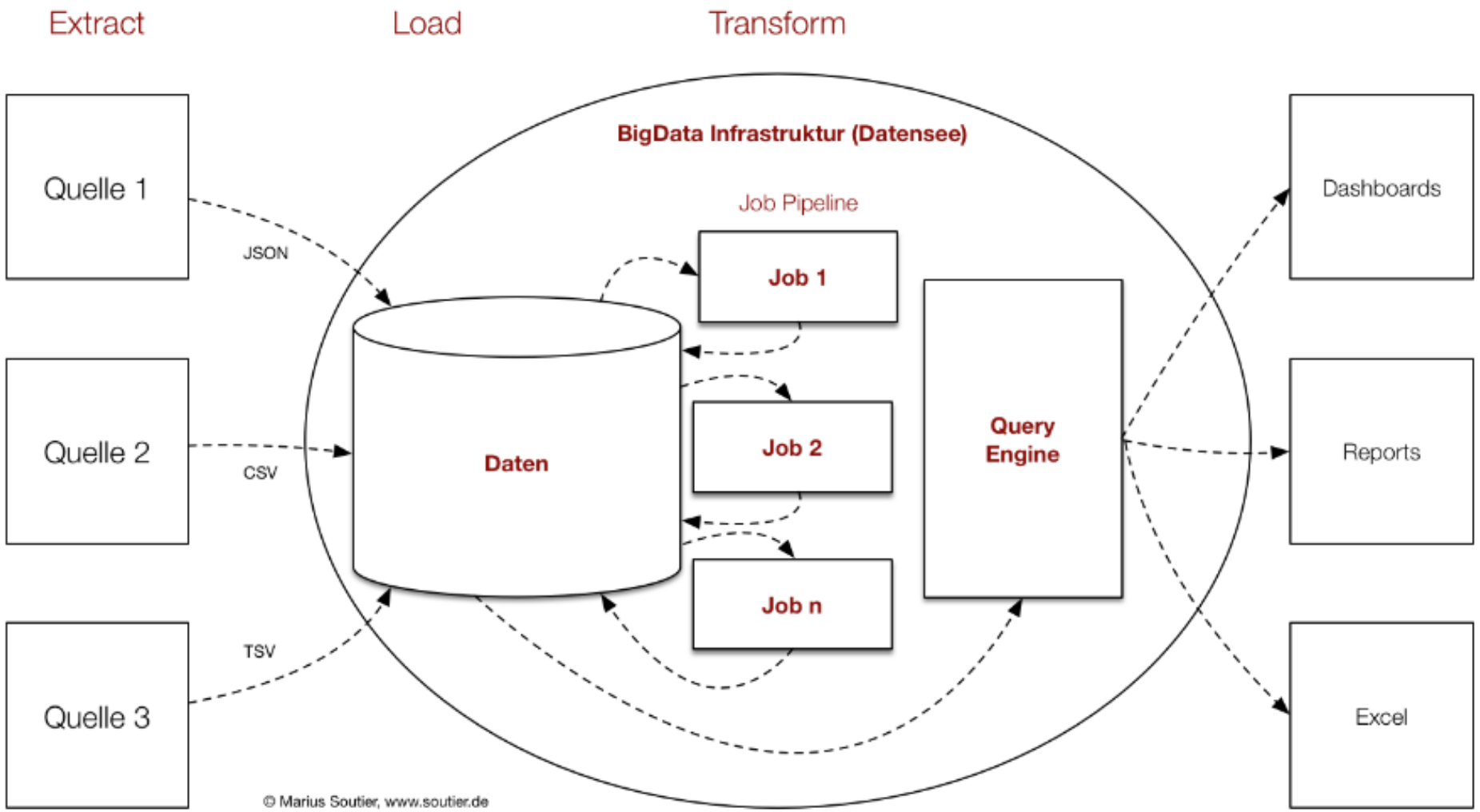

ETL
Transform
DWH
?
Transform
ELT
Replicate
?
Data Lake

ETL (extract, transform, load) 1

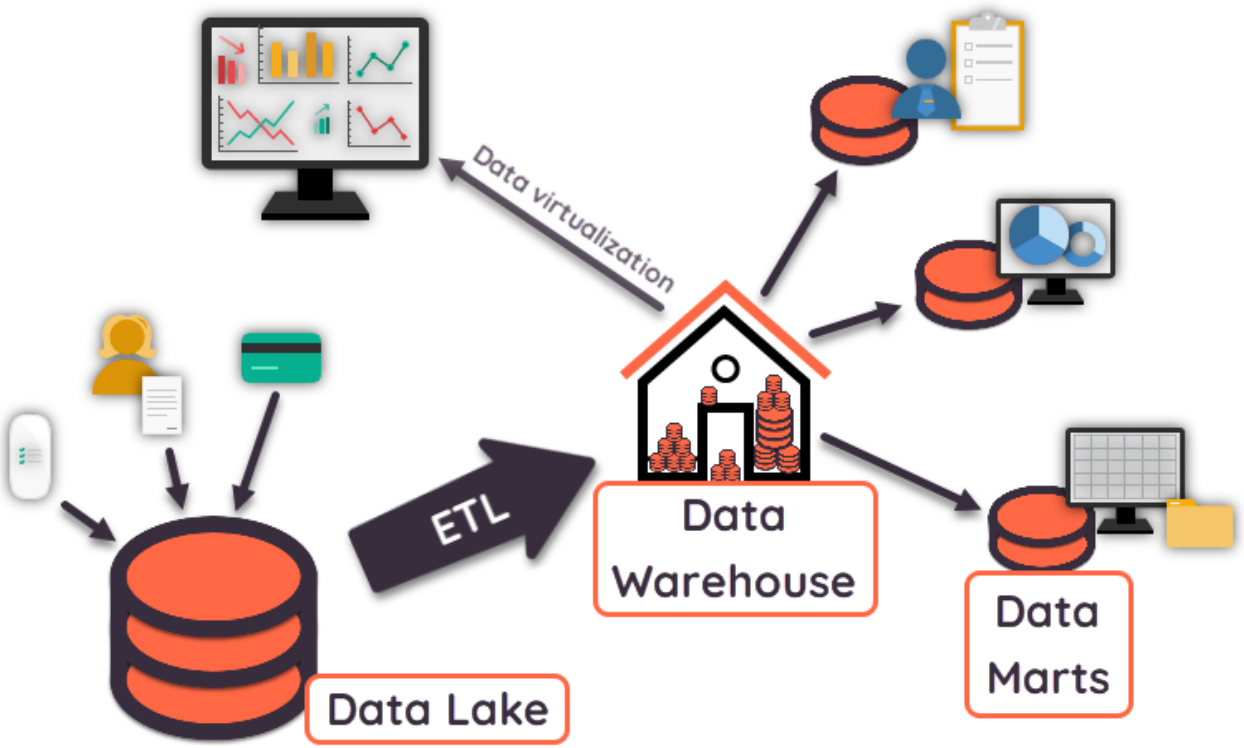


https://datasolut.com/was-ist-ein-etl-prozess/

ELT (extract, load, transform) 2



ETL Verwendung in der Praxis



Datenbereinigung 1

Datenbereinigung, auch Data Cleaning, Data Cleansing oder Data Scrubbing genannt, ist der Prozess der Korrektur falscher, unvollständiger, doppelter oder anderweitig fehlerhafter Daten in einem Datensatz. Es geht darum, Datenfehler zu identifizieren und dann Daten zu ändern, zu aktualisieren oder zu entfernen, um sie zu korrigieren. Datenbereinigung verbessert die Datenqualität und trägt dazu bei, genauere, konsistentere und zuverlässigere Informationen für die Entscheidungsfindung in einem Unternehmen bereitzustellen.

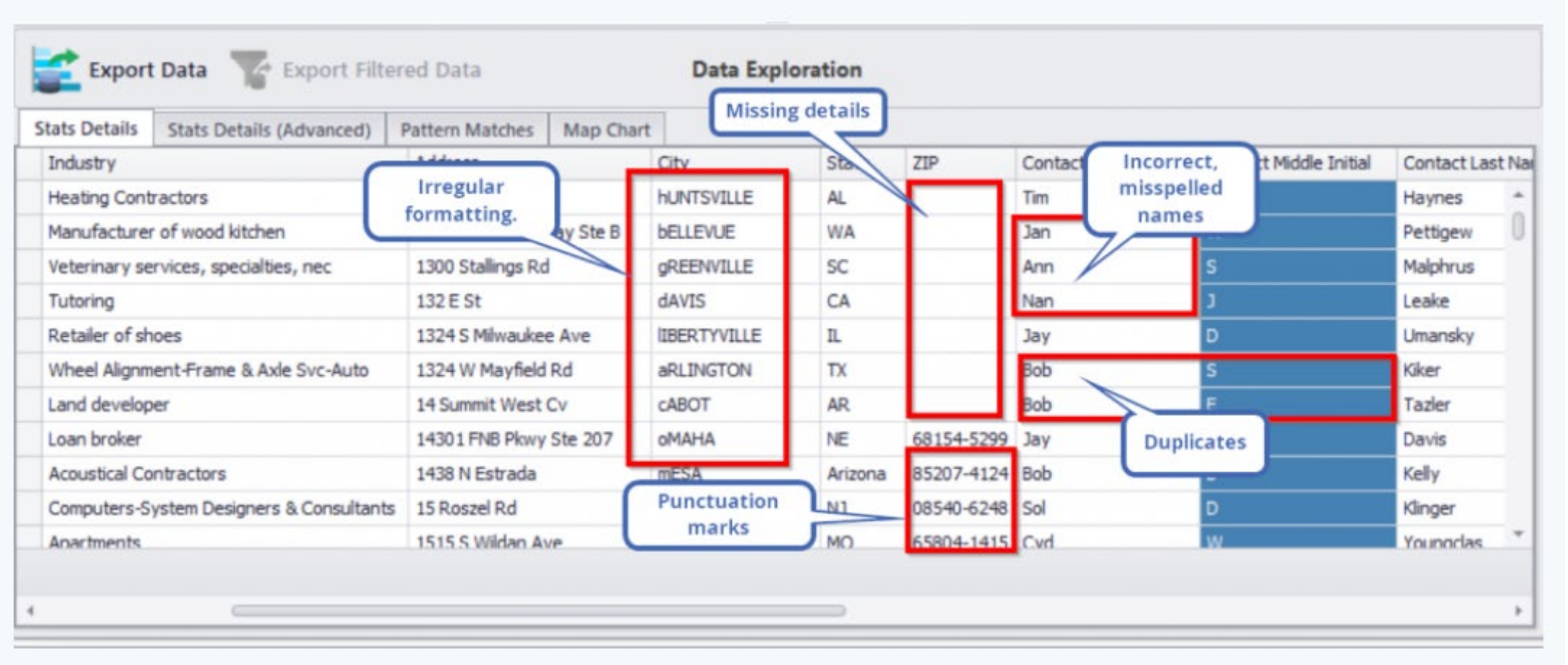
https://www.bigdata-insider.de/was-ist-datenbereinigung-a-843546/

https://dataladder.com/de/grundlagen-der-datenbereinigung-wie-man-mit-schlechten-daten-einfach-umgeht/

https://www.astera.com/de/type/blog/data-cleansing-tools/

Data Cleansing versus Data Cleaning versus Data Scrubbing

Datenbereinigung 2



Datenbereinigung 3

Was sind die Schritte im Datenbereinigungsprozess?

1. Inspektion und Profiling
2. Bereinigung
3. Verifizierung
4. Reporting

Datenbereinigung 4

Merkmale von bereinigten Daten.

Verschiedene Datenmerkmale und -attribute werden verwendet, um die Sauberkeit und Gesamtqualität von Datensätzen zu messen, darunter:

- Genauigkeit
- Vollständigkeit
- Konsistenz
- Integrität
- Aktualität
- Einheitlichkeit
- Gültigkeit

Datenbereinigung 5

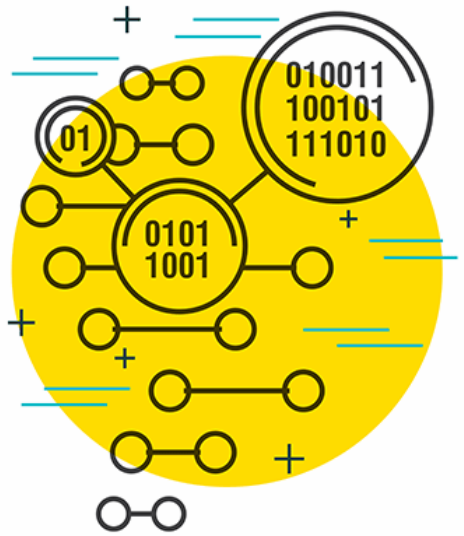
Vorteile der Daten-bereinigung

- Beseitigung von Fehlern, Verzerrungen und Inkonsistenzen
- Spart Zeit und Geld
- Rationalisiert Geschäftsprozesse
- Verbessert die Datenintegrität und -qualität

ILLUSTRATION: YELLOWLINE/ADOBE STOCK

Datenbereinigung 6

Allgemeine Metriken zur Datenqualität



- Statistiken über die Anzahl der erkannten und behobenen Datenfehler pro Monat oder Quartal
- Genauigkeit und Fehlerquoten in den Datensätzen, mit Warnmeldungen, wenn die Fehler ein akzeptables Niveau überschreiten
- Quantitative Messungen von Datenvollständigkeit, -konsistenz, -integrität und -aktualität
- Berechnungen der Auswirkungen von Datenqualitätsproblemen auf das Geschäft und potenzielle Abhilfemaßnahmen
- Bewertungen der Qualitätsniveaus in Datendefinitionen, Metadaten und Datenkatalogen
- Umfragedaten und andere Rückmeldungen zur Datenqualität von Endnutzern gesammelt

ILLUSTRATION: GSTUDIO GROUP/ADOBE STOCK

Links

https://www.tecchannel.de/a/bi-datenmanagement-teil-1-datenaufbereitung-durch-den-etl-prozess,1746250

---- laut vorigenen Folien

https://datasolut.com/was-ist-ein-etl-prozess/
https://www.bigdata-insider.de/was-ist-datenbereinigung-a-843546/
https://dataladder.com/de/grundlagen-der-datenbereinigung-wie-man-mit-schlechten-daten-einfach-umgeht/
https://www.astera.com/de/type/blog/data-cleansing-tools/
Top 5 Data Cleansing Tools Every Data Professional Should Know (syncari.com)

Scrubbing vs. Harmonisierung von Daten

Datenbereinigung, auch Data Scrubbing genannt, bezieht sich auf den Prozess der Identifizierung und Korrektur von doppelten, inkonsistenten, ungenauen oder unvollständigen Datensätzen in den Stammdaten. Bei der **Datenharmonisierung** geht es um die Standardisierung und Organisation von Daten aus unterschiedlichen Quellen oder Formaten in einer einheitlichen und konsistenten Struktur. Dadurch soll sichergestellt werden, dass die Daten kompatibel sind, gemeinsamen Standards folgen und effektiv integriert und analysiert werden können.

Was beim **Abgleichen von Kodierungen, Synonymen und Homonymen** passiert, lässt sich an folgenden Beispielen veranschaulichen:

- Einzelne Datenbestände können unterschiedlich kodiert sein. So können Attribute wie Geschlecht in Datenquelle 1 als „M“/ “W“ kodiert sein, in Datenquelle 2 als 0-1-Variable.

- Hierbei werden die gefilterten und bereinigten Daten zusammengeführt. Vor allem drei Problemklassen müssen hier angegangen werden:
- Erstens, das Abgleichen von Kodierungen, Synonymen und Homonymen,
- zweitens, das Auflösen von Schlüsseldisharmonien und
- drittens, die betriebswirtschaftliche Harmonisierung.

- Unterschiedliche Attributnamen können die gleiche Bedeutung haben (Synonymie). Beispielsweise kann in Datenquelle 1 für den Namen von Betriebsmitarbeitern das Attribut „Personal“ vorgesehen sein, in Datenquelle 2 aber „Mitarbeiter“.
- Umgekehrt können gleiche Attributnamen unterschiedliche Bedeutungen haben (Homonymie). In Datenquelle 1 kann „Partner“ beispielsweise den Namen von Kunden bezeichnen, in Datenquelle 2 den Namen von Lieferanten

In allen drei Fällen müssen die Daten harmonisiert werden. Im ersten Fall muss der Attributwert einheitlich z.B. auf 0-1-Werte festgelegt werden, im zweiten Fall ist ein identischer Attributname zu wählen und im dritten Fall ein unterschiedlicher Attributname. Auch hier werden für den Abgleich in der Regel Mapping Tables implementiert, die die gefilterten Dateien über Namensabgleichungen und Kodierungsabstimmungen zu themenorientierten Datensammlungen zusammenführen.

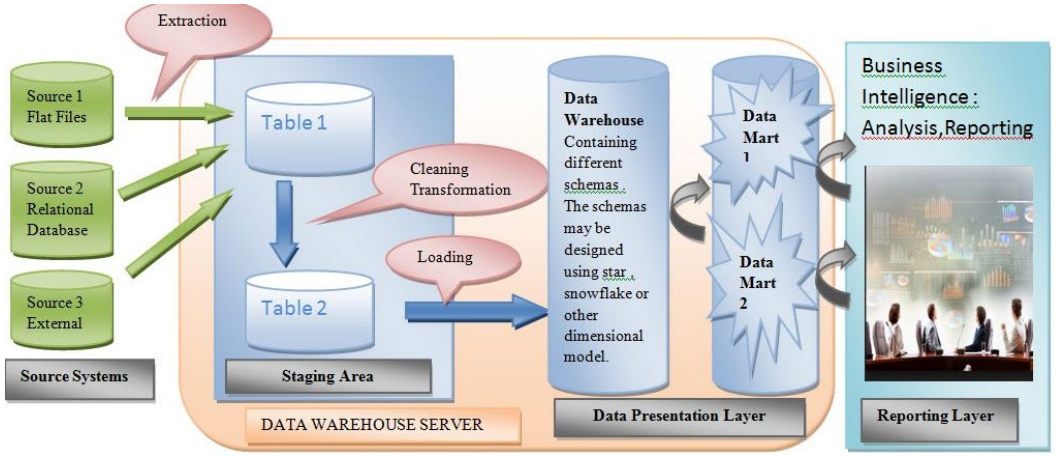
Implementierung

Quell-DB – stored proc (gescheduled) – DWH-DB Table 1 (staging) – stored proc /gescheduled) – DWH-DB Table 2 – DWH

Quell-DB – Phyton + stored proc (gescheduled) – DWH-DB Table 1 (staging) – Python + stored proc /gescheduled) – DWH-DB Table 2 – DWH

SSIS – Packages

ETL-Tools



Stored proc vs. python

[The Benefits Of Using Python And T-SQL Over SSIS For ETL | by Bob Wakefield | Data Driven Perspectives]