ANALYTIQ

• CONCEPTS AT A GLANCE •

# Data Lake Architektur

Karim Hanke
amotIQ solutions

## ANALYSE

Zugänglich über
- Web Browser
- Client Applikationen

Die proaktiven Methoden der Analyse sind

- Predictive Modellierung
- Descriptive Modellierung
- Data Mining
- Text Mining
- Statistische/ Quantitative Analyse
- Simulation & Optimierung

Vom Endanwender erstellte Skripte sollten skalierbar und parallelisierbar sein. Die Verarbeitung von großen Datenmengen muss für diverse Workload-Kategorien wie

- Abfragen
- ETL
- Analytik
- Maschinelles Lernen
- Maschinenübersetzung
- Bildverarbeitung
- Sentimentanalyse

durch den Einsatz bestehender Bibliotheken gewährleistet sein.

## AUSWERTUNG

Zugänglich über
- Web Browser
- Mobiles Web
- Mobile Apps

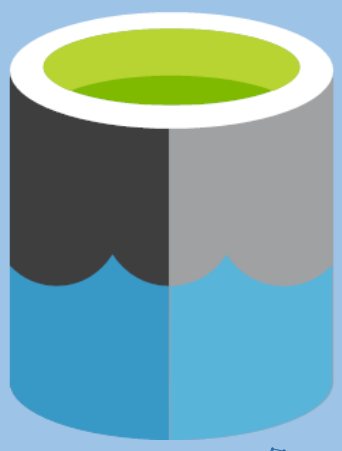
Die reaktiven Methoden des Reportings sind

- KPIs und Metriken
- Automatisierte Überwachung und Alarmierung
- Dashboards
- Scorecards
- OLAP
- Ad-hoc-Abfragen

Daten werden in folgender Art und Weise visualisiert:

- Abstract und Kurz - Für Endanwender mit wenig Zeit und einen besonderem Fokus
- Jahresberichte - Hoch formalisierte/ standardisierte Berichten mit allen Aspekten
- Fact Sheet - Detailinformationen zu bestimmten Sachverhalten
- Empirische Publikation - Forschungs- oder Evaluationsergebnisse

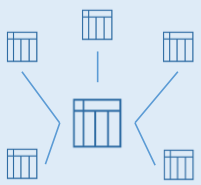
## ENTERPRISE DATA LAKE



**Daten Veredelungs-prozess** während des lesens

**Transformieren relevanter Daten** in Datensätze

**Data Warehouse** mit leseoptimierten Strukturen

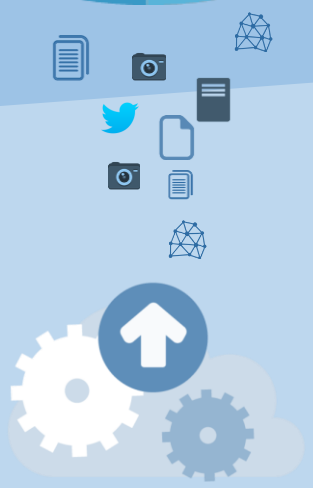


Im EDL (Enterprise Data Lake) werden die Fragen beantwortet

- Was wird geschehen?
- Was wird geschehen wenn wir etwas ändern?

Daten werden beim lesen verwandelt und veredelt, in kuratierte Datensätze transformiert um schlussendlich in verschiedenen Schemas abgespeichert und als Data Warehouses zur Verfügung gestellt zu werden.
Die Nutzer sind hauptsächlich Data Scientists, Business Analysten und Fachanwender.

## DATENINTEGRATION

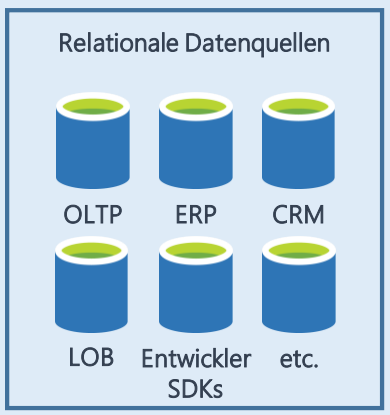


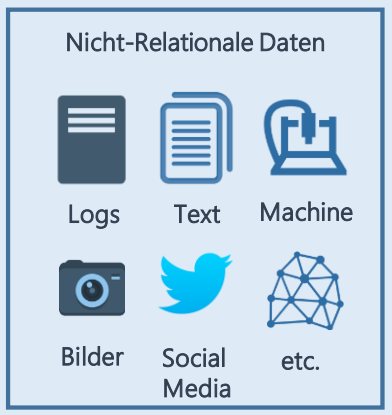
Zur ordnungsgemäßen Integration müssen folgende Management-funktionen dem User zur Verfügung gestellt werden:

- Starten von lokalen und remote gespeicherten Paketen
- Stoppen von lokalen und remote laufenden Paketen
- Überwachung von lokalen und remote laufenden Paketen
- Importieren und Exportieren von Paketen
- Paketspeicher verwalten
- Anpassen von Speicherordnern
- Stoppen von laufenden Pakete, wenn Dienste gestoppt werden
- Anzeigen der Event Logs

Die Datenintegration extrahiert und lädt die Daten hauptsächlich (EL statt ETL). Transformationen gilt es zu vermeiden.
Dadurch werden die Daten in ihrer nativen Form im Enterprise Data Lake gespeichert.
So wird dem Endanwender die Orchestrierung und das Streamen von Daten möglich gemacht.

## DATENQUELLEN




Es sollten alle Typen von Datenbanken in Betracht gezogen werden, um das größte Wertschöpfungspotential zu erhalten.

- Relationale
- Analytische (OLAP)
- Key-value
- Column-family
- Grafen
- Dokument

Die Daten sollten idealerweise in Rohform und nicht verdichtet vorliegen.

Der **Data Lake** ist ein großes, leicht zugängliches, **zentralisiertes Repository von großen Mengen an strukturierten und unstrukturierten Daten**. Die Daten werden nicht klassifiziert, wenn sie im Repository gespeichert werden, da der Wert der Daten am Anfang nicht klar ist.

aNALYTIQ

• CONCEPTS AT A GLANCE •

# Data Lake Architecture

Karim Hanke
amotIQ solutions

## ANALYZE

Accessable via
- web browser
- Client applications

The proactive methods used for analysis are

- Predictive Modeling
- Descriptive Modeling
- Data Mining
- Text Mining
- Statistical/ Quantitative Analysis
- Simulation & Optimization

The analysis tools should be simple and extensible that allows you to write code once and have it automatically parallelized for the scale you need. Process GB of data for diverse workload categories such as

- Querying
- ETL
- Analytics
- Machine learning
- Machine translation
- Sentiment analysis

by leveraging existing libraries.

## REPORTING

Accessable via
- web browser
- mobile web
- mobile apps

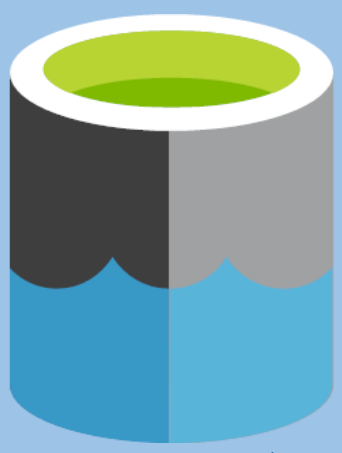
The reactive methods used for reporting are

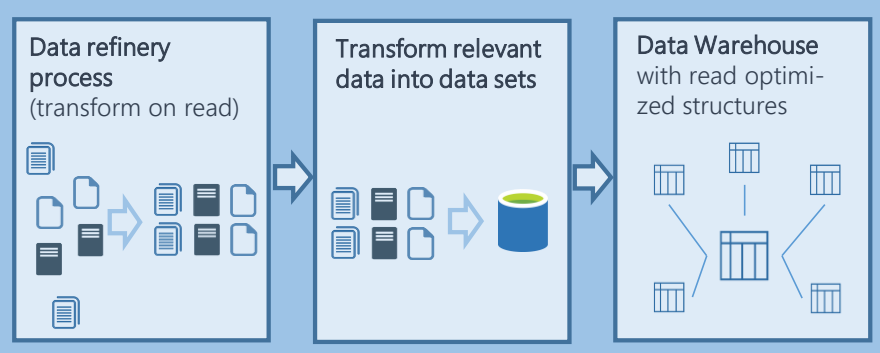
- KPIs and metrics
- Automated monitoring and altering (thresholds)
- Dashboards
- Scorecards
- OLAP (Cubes, Slice and Dice, Drilling)
- Ad hoc query

Information will be visualized by the following methods:

- Abstract and Briefing - For audiences who are short on time or focus
- Annual Reports Highly formal report on all aspects of a program and the evaluation
- Fact Sheet - To pick out relevant facts about the data at a glance
- Empirical Publication - Research or evaluation findings

## ENTERPRISE DATA LAKE



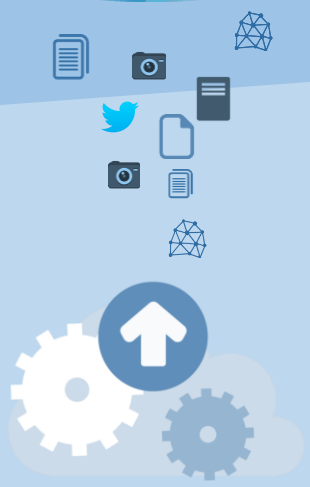



The purpose of the EDL (enterprise data lake) is to answer the questions

- What will happen?
- What will happen if we change this one thing?

Data will be transformed and refined on read, transformed into curated data sets to be finally stored in different schemas and made available as data warehouses.
The users are mainly data scientists, business analysts and professional users.

## DATA INTEGRATION

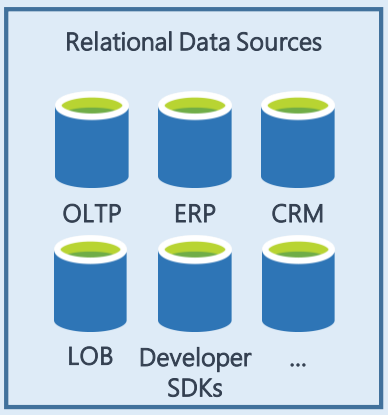


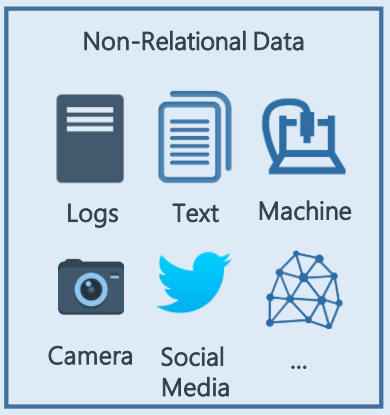
To run the integration properly, the following **management capabilities** have to be provided:

- Starting remote and locally stored packages
- Stopping remote and locally running packages
- Monitoring remote and locally running packages
- Importing and exporting packages
- Managing package storage
- Customizing storage folders
- Stopping running packages when the service is stopped
- Viewing the Event log

The data integration does mainly extract and load the data (EL instead of ETL). There should be no to minimal transformation during the integration.
Hereby the data will be stored in a near-native format to the data lake. Orchestration and streaming data accommodation becomes possible for the end user.

## DATA SOURCES




All types of databases should be concidered, such as

- Relational
- analytical (OLAP)
- Key-value
- Column-family
- Graph
- Document

to have the biggest data potential.

Leverages the power of on-premise technologies and the cloud for storage and capture. The data should be idealy offered in a native format.

The **Data Lake** is a **large**, easily **accessible, centralized repository of large amounts of structured and unstructured data**.
The data is not classified when stored in the repository because the value of the data is not clear at the beginning.