ANALYTIQ
· Konzepte auf einen Blick ·

Data Lake Architektur

amotIQ solutions

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.

ANALYSE

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.

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**

AUSWERTUNG

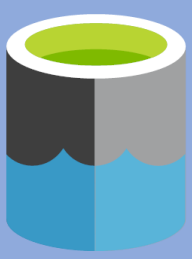
Daten werden in folgender Art und Weise visualisiert:
- Abstract und Kurz - Für Endanwender mit wenig Zeit und einen besonderem Fokus
- Jahresberichte - Hoch formalisierte/ standardisierte Berichten mit allen Aspekten
- Fact Sheet - Detailinformationen zu bestimmten Sachverhalten
- Empirische Publikation - Forschungs- oder Evaluationsergebnisse

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**

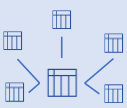
ENTERPRISE DATA LAKE



Daten Veredelungs-prozess während des Lesens

Transformie-ren relevanter Daten in Datensätze

Data Warehouse mit lese-optimierten 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.

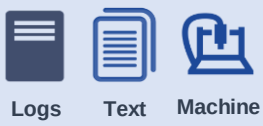
DATENQUELLEN

Relationale Datenquellen

OLTP ERP CRM

LOB Entwickler SDKs etc.

Nicht-Relationale Daten



Logs Text Machine

Bilder Social Media etc.

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.