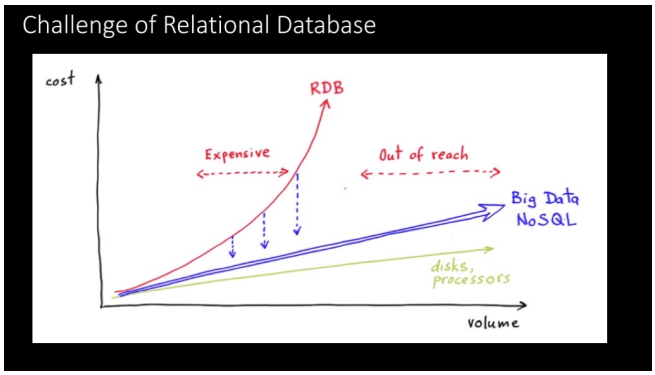
CDP Data Lake on AWS



Inhalt:

Video zu Stand und Ausblick CDP 07/2017:

Was sind die Herausforderungen für eine Relational Datenbank. Sehen wir uns ein einfaches Diagramm an. Auf der x-Achse bilden wir das Datenvolumen ab und auf der y-Ache die Kosten für die Datenbank bzw. für das Equipment. Wir sehen, dass wir beim Speicher und bei den Prozessoren einen linearem Kostenverlauf haben. D.h. für den doppelten Speicher bezahlen wir den doppelten Preis. Analog für die Prozessoren. Wenn wir nun die Kosten für eine Datenbank eintragen, sehen wir, dass sich diese nicht linear entwickeln. Start mit kleinen, Open Source Systemen, endet mit teuren großen Systemen (Exadata).

Zwei Probleme:

1. Processing ist sehr langsam. Je mehr Daten desto langsamer
2. Für mehr Daten zahlen wir einen immer höheren Preis, und bei richtig großen Datenmangen können diese von Datenbanken nicht mehr verarbeitet werden.

Das war die Motivation für die großen Web Unternehmen Google and Yahoo

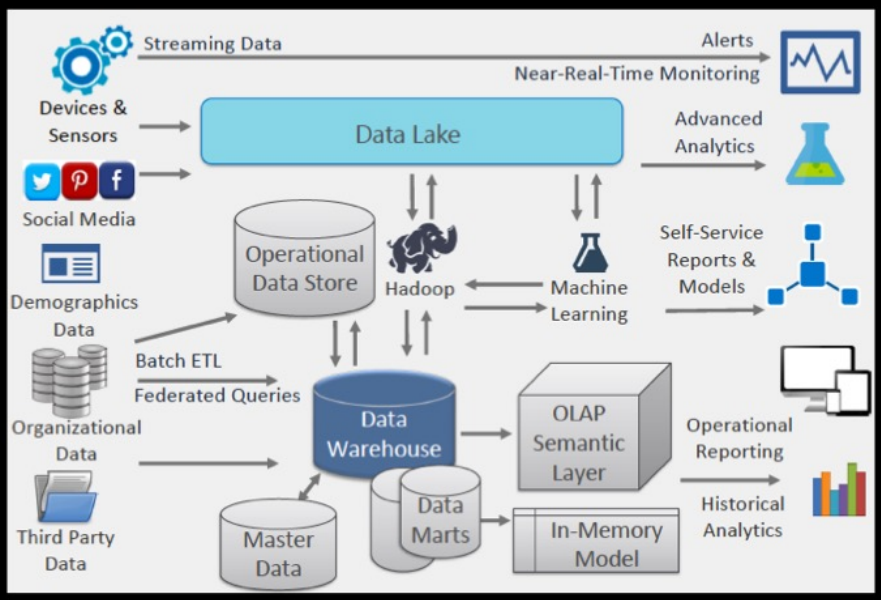
Eine neue Lösung finden -> zwei Ziele.

1. Performance verbessern und die Kosten senken
2. Daten verarbeiten können, die RDB nicht mehr verarbeiten können.

Dies führt uns zum Konzept des Data Lakes. Ein Data Lake ist ein Repository für Data. Hier werden alle Daten abgelegt die für die Weiterverarbeitung in Big Data notwendig sind.

CDP Architektur

Einordnung des Data Lake



Modernes Data Warehousing