Big Data 2.2

Das faktenbasierte Modell zur Repräsentation von Daten

- Daten müssen im Stammdatensatz repräsentiert werden
- Empfehlung:
  - Faktenbasiertes Datenmodell: Daten werden in grundlegende Bestandteile aufgegliedert, die als Fakten bezeichnet werden
- Eigenschaften von Fakten: atomar und mit einem Zeitstempel versehen
  - Atomar: Fakten können nicht weiter in kleinere, noch aussagekräftige Komponenten unterteilt werden und es gibt keine redundanten Informationen
  - Zeitstempel: Zeichnen die Fakten als unveränderlich und dauerhaft richtig aus
- Weitere Eigenschaft von Fakten: Identifizierbarkeit
  - Fakten sollten einzigartige Kennzeichnungen zugeordnet sein, die diese eindeutig identifizierbar machen
  - Die Identifizierbarkeit gestattet es, denselben Fakt mehrfach in den Stammdatensatz aufzunehmen, ohne dessen Semantik zu ändern
- Zusammenfassung des faktenbasierten Datenmodells:
  - Rohdaten werden als atomare Fakten gespeichert
  - Fakten werden mit einem Zeitstempel versehen und sind dadurch unveränderlich und dauerhaft richtig
  - Fakten sind identifizierbar, sodass Dubletten bei Abfragen erkannt werden können

Vorteile des faktenbasierten Modells

- Stammdatensatz: kontinuierlich wachsende Liste von unveränderlichen und atomaren Fakten
  - Relationale Datenbanken sind nicht dafür ausgelegt
- Vorteile:
  - Daten können für jeden beliebigen Zeitpunkt abgefragt werden
  - Daten sind fehlertolerant gegenüber menschlichem Versagen
  - Unvollständige Daten können gehandhabt werden
  - Daten weisen sowohl die Vorteile der normalisierten als auch der nicht normalisierten Form auf

Vorteile des faktenbasierten Modells

Abfragen für jeden beliebigen Zeitpunkt

- Zustand zu jedem beliebigen im Datensatz enthaltenen Zeitpunkt kann abgefragt werden
  - Folge daraus, dass Fakten unveränderlich und mit eine Zeitstempel versehen sind
  - Da keine Daten entfernt werden, kann der Zustand zu dem abgefragten Zeitpunkt rekonstruiert werden

Fehlertoleranz gegenüber menschlichem Versagen

- Fehlerhafte Daten werden einfach gelöscht

Unvollständige Informationen

- Da jeder Datensatz nur einen Fakt enthält, können leicht unvollständige Informationen über ein Objekt gespeichert werden, ohne das NULL-Werte in den Datensatz aufgenommen werden müssen – fehlende Fakten entsprechen logisch einem NULL-Wert

Vorteile des faktenbasierten Modells

Speicherung und Verarbeitung der Daten finden auf verschiedenen Layern statt

- Informationen werden sowohl im Batch- als auch im Serving-Layer gespeichert: Daten liegen in normalisierter und nicht normalisierte Form vor
- Normalisierung: strukturierte Speicherung der Daten zur Minimierung der Redundanz und Förderung der Konsistenz
- Lambda-Architektur: Vorteile einer vollständigen Normalisierung und die Performancevorteile einer Indizierung der Daten zur schnelleren Beantwortung von Abfragen