- Data Warehouse
  - DWH soll: "Reich an Daten, arm an Informationen und Wissen" wiederlegen
  - Unstrukturierte Daten
    - Herausforderungen
      - DWH führ strukturierte und unstrukturierte Daten eine einem Ort zusammen
      - Kein vorgegebenes Schema
      - Zugriff abhängig von Semantik (Inteprationsspeilraum)
      - hohes Datenvolumen, kein manuelles Extrahieren möglich
      - Prioritisierung je nach Semantik und Kontext
      - Datenqualität kann nicht Autonom gemessen werden
    - Schritte für logische und Physische Integaration
      - Extraktion
        - komplex
        - Information Retreival
        - Analyse für Integration (NLP - Natural Language Processing)
        - Kategorisieren, Auswertung von Dok /Flatfile
      - Datenbereinigung (Data Cleansing)
        - Standard Cleaning
        - NLP
      - Laden in DBMS, DWH
      - ETL Extract Transform Load
        - Connect & Collect, Trnasform & Enrich, Publish, Monitor à la definition Microsoft
  - Definiton
    - Ein Datawarehouse ist eine konsistente, speziell analytische entwickelte Datenbank, welche ein Unternehmen massgeblich bei der Enscheidungin hinsicht auf das Decision Mangement unterstützten kann. Die Hauptfunktion ist eine einheitliche Umgebung zu schaffen aus welcher BI Tools, sprich OLAP do. Multidemensionale Abfragen generiert werden können
    - Ein Data - Warehouse integriert aus vielen unterscheidlichen Quellen in einer für die Entscheidungsfindung optimierten Datenbank
  - Referenzmodell für die Architektur von DHW-Systemen
    - Dies soll unzeigen wie in wir ein DWH in eine Unternehmen integrieren können.
    - Bestandteile des Referenzmodells Definitionen
      - Reportingbereich SSRS
        - Im Referenzmodell nicht erwähnt
        - SQL Server Reporting Services SSRS
      - Auswertungsbereich SSAS
        - SQL Server Analysis Services = SSAS
        - Ableitungsdatenbank
        - Zentrales Data DWH
        - möglichst homogener Datenbestand
        - Data Mart
        - Teilkopien von DW für u.a Abteilungen
        - Cube
        - für mehrdimensionale Abfragen
      - Integrationsbereich SSIS
        - SQL Server Integration Service = SSIS
        - Arbeitsbereich = Staging Area
        - periodische Transformation od. Homogenisierung
        - Bais DB od. Operative DB
        - Datendrehscheibe
        - ODS Operational Data Sotrage
      - + externe Systeme
        - DB's
        - Flatfiles
        - Webserver
        - CRM- Systeme
      - Verwaltungsbereich
        - Repository
    - Bestanteile der Referenzarchitektur im Detail
      - 1. Datenquellen - Externe Daten
        - Meist Falsche Tabellen oder Flache Files
          - Daten aus Legacy Systemen
          - Daten aus Anwendungsprogrammen SAP
          - Zeichencodierung ASCII, ANSI, UTF 8
          - Trenzeichen Abstand,Tabulator, Semikolon
          - Zeilenschaltung (Paragraph) CR/ LF Windows, LF Unix+Mac
        - XML Datenströme soll und muss heute Standard werden
      - 2. Arbeitsbereich Staging Area
        - Integriert Daten, die Heterogenen
        - Das E von ETL
        - Datenbank Tabellen, Daten Stammen aus sehr unterschiedlichen Systemen. Alle Notwendigen Berschreibungen zum System können zur Umwelt darin enthalten sein.
        - Date werden mit Hilfe der Metadaten zusammengeführt und abgelegt
        - Metadaten Respsitory
        - Arbeitsdepot, das flüchtige Zwischendepot
        - Auch für Simulationen, Berechnungen usw
      - 3. Basis DB (Operational Data Sorage ODS)
        - Zwischenstufe von Quellsystem zum DWH
        - Auch als Last und Leistungsausgleich
        - z.b für voranalysen
        - Daten in der BasisDatenbank sind Bereinigt
        - Data Cleaning & Data Wrangling
        - ODS und Stating Area kann und muss nicht immer Zusammen sein
        - Integrationsbereich
      - 4. Ableitungsdatenbank Data Warehouse
        - Data Vault
      - 5. Auswertungsdatenbank (Data Mart)
        - Teilkopie des DWH
          - bessere Leistung
          - Verlagerung der Rechenleistung
          - Weniger Network Traffic
        - Unterschied zu DWH
          - Verwendung
          - Anwendungsorientiert nicht wie dwh das anwendungsneutral ist
          - Anzahl
          - Abteiltung vs. Unternemen
          - Ein Data Mart im Unterscheid zum Data Lake hat viel Potenzial. Man Stelle sich vor das man ein Wein auch in einem Delikatessenladen kauft für einen Speziellen Anlass. Der Spezialist im Laden kennt sich viel besser mit den Produkten aus als in einer Warenkette. Es kann sicher möglch sein beide Produkte an beiden Orten zu bekommen, aber im Delikatessenladen wird man höchstwarscheindlich nicht sie schalchten weinse bekommen die man höchstwarschiendlich kaufen würde in eineem Supermarkt.
  - Referenzmodell Bauer & Günzel - Seite 42 (Buch)
    - SSAS
    - SSIS
  - Datenstrukturen
    - Stern/ Star Schema
      - Es steht nich die Normalisierung sondern die Optimierung auf effiziente Leseoperationen
      - Im Zentrum die Faktentabelle und aussen rum sind die Dimension
      - Sternschem ist denormalisiert, Anomalien und erhöhter Speicher-bedarf ist möglich
    - Snowflake
      - Hier müssen allerdings mehrstufige Dimensionstabellen über Join-Abfragen verknüpft werden.
      - Hat die Dimensionen in der 3.NF
      - Erweiterte Unterteilungen etc.
    - Galaxy
      - Teilt die gleichen Diemensionen mit mehreren Faktentabellen
  - Referenzmodell in der Praxis umgesetzt
    - Hub and Spoke Architektur
      - Nabe Speicher Architektur
      - Zentrales DWH (HUB) mit angeschlissenen DM (Spokes)
    - Legacy / Middelware
      - aus verschiedenen Legacy Systemen direkt ins Data Warehouse, daraus in die DataMarts via Middleware in Anwendungen (Analysen)
    - High Level System Architecture
      - Alles auf ändlicher Ebene nahe am Präsentations Layer