Workshop

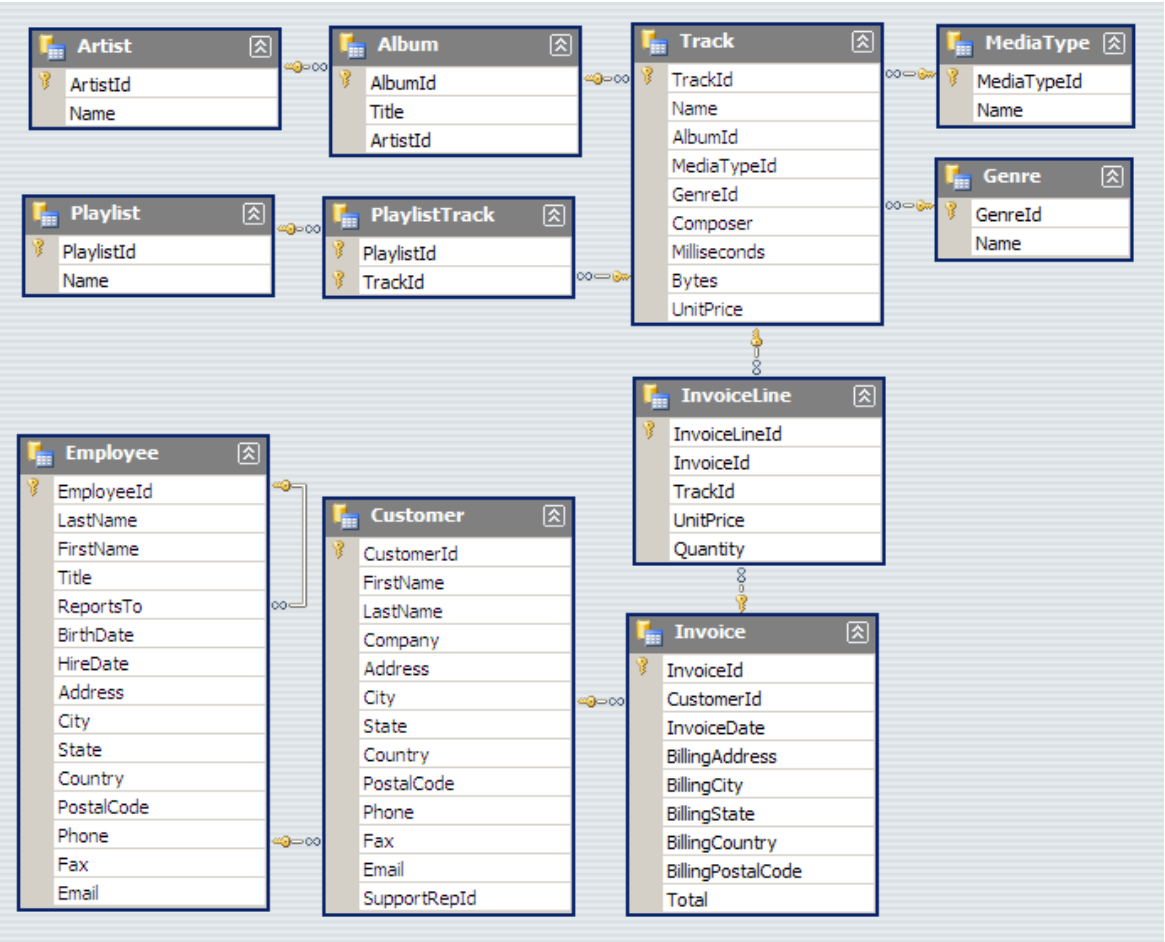
Chinook

- Convert ER to Dimensional Model
- Load Dimensional Model

Chinook

Sample Database

Chinook Database



The Chinook data model represents a digital media store, including tables for artists, albums, media tracks, invoices and customers.

Chinook data model is an Entity Relationship (ER) Model.

Chinook Database

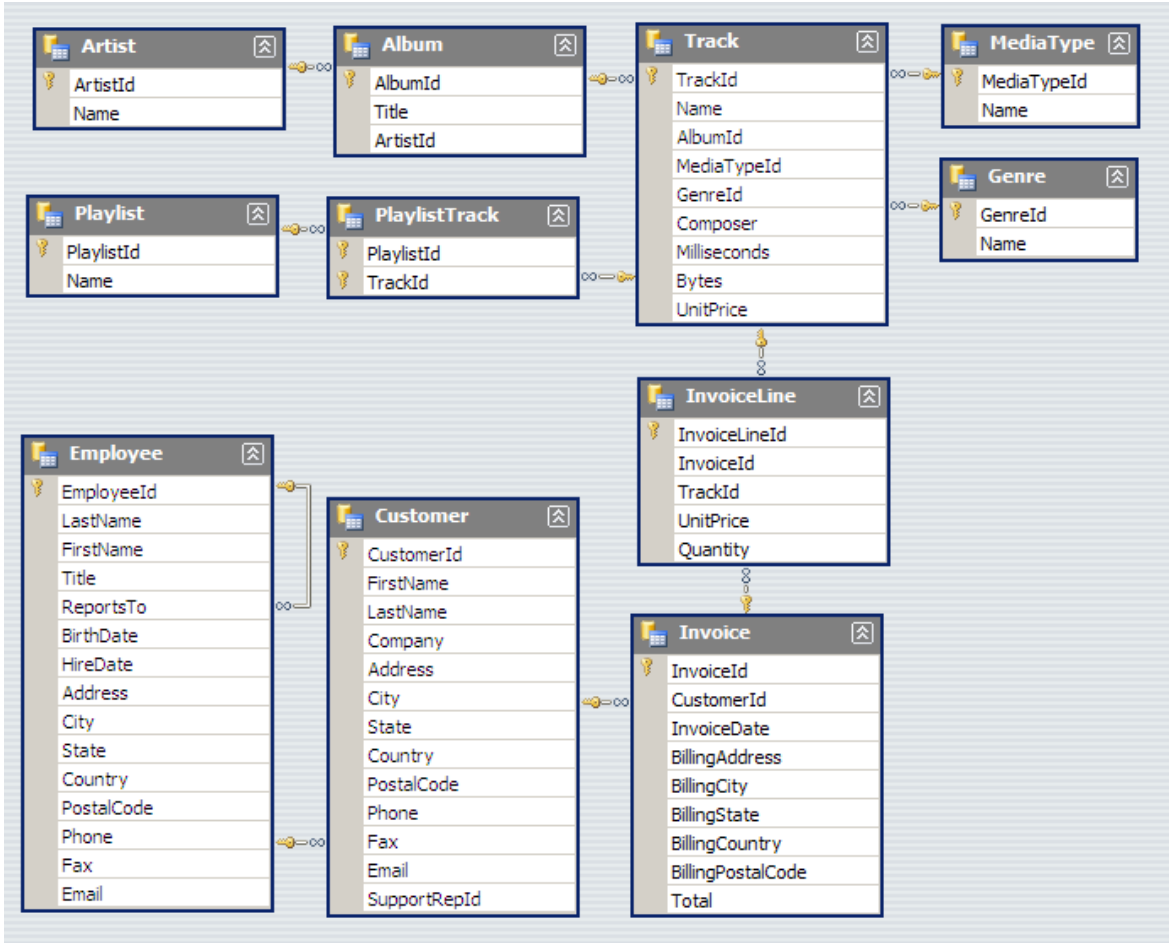
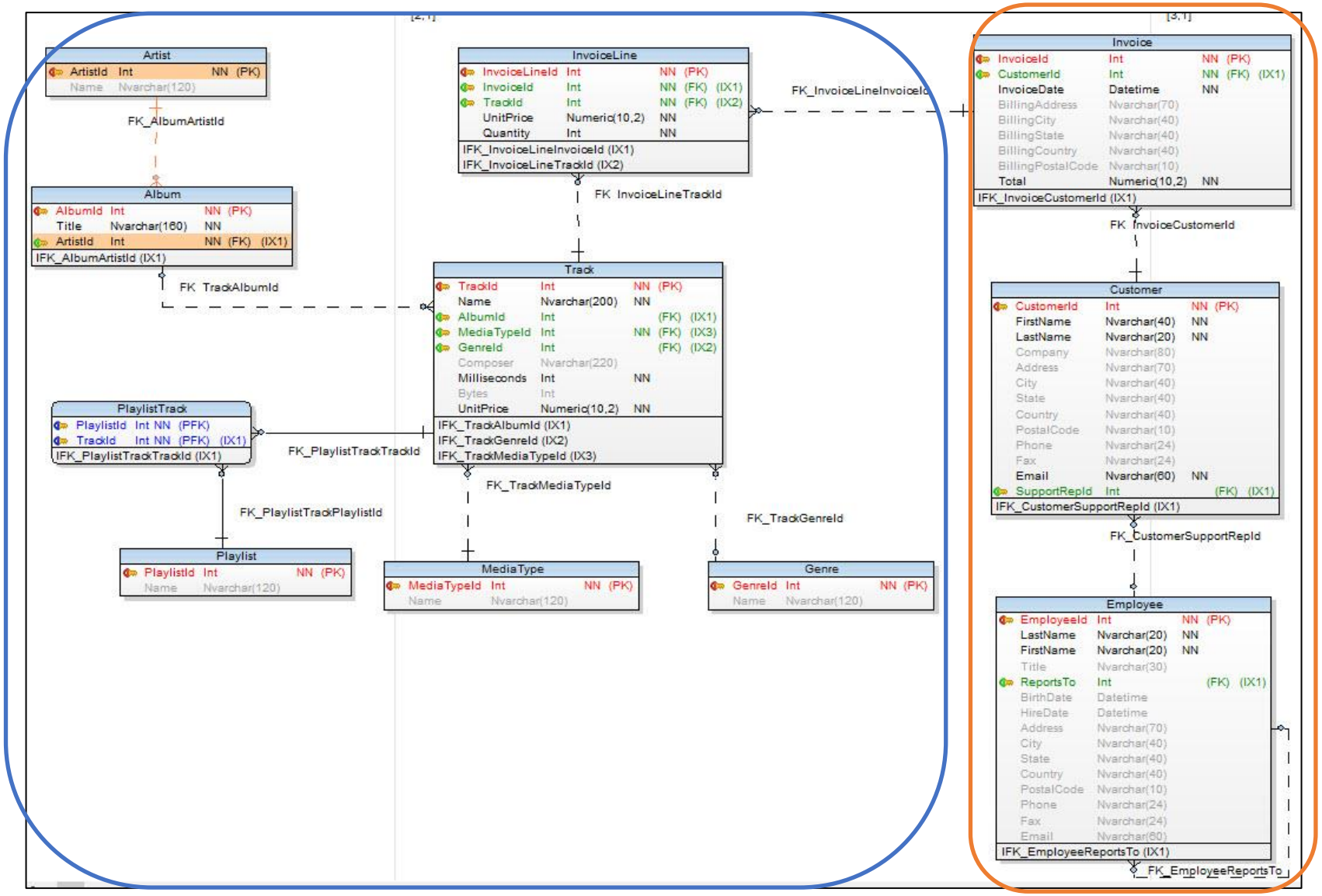


Table Name	Row
Album	347
Artist	275
Customer	59
Employee	8
Genre	25
Invoice	412
InvoiceLine	2,240
MediaType	5
Playlist	18
PlaylistTrack	8,715
Track	3,503

Chinook Database: Data Model (ER/Studio)



- Sale $ in two entities
  - Invoice
  - Invoice Line Item
- Entities tied to Invoice
  - Customer
  - Employee
- Entities tied to Invoice Line
  - Track (Song)
  - Album
  - Artist
  - Genre
  - Media Type

Chinook

Convert ER Model to Dimensional Data Model

Deliverables

- Reverse engineer Chinook creating an ER Model (3NF) using ER/Studio
  - Upload ER/Studio file and screenshot of data model
- Convert ER Model to Dimensional Model using ER/studio
  - See next slide for description of process
  - Upload ER/Studio file and screenshot of data model
- Create DDL scripts for Dimensional Model using ER/studio
  - Upload sql script for
    - MySQL
    - SQL Server
    - Oracle
    - PostgreSQL

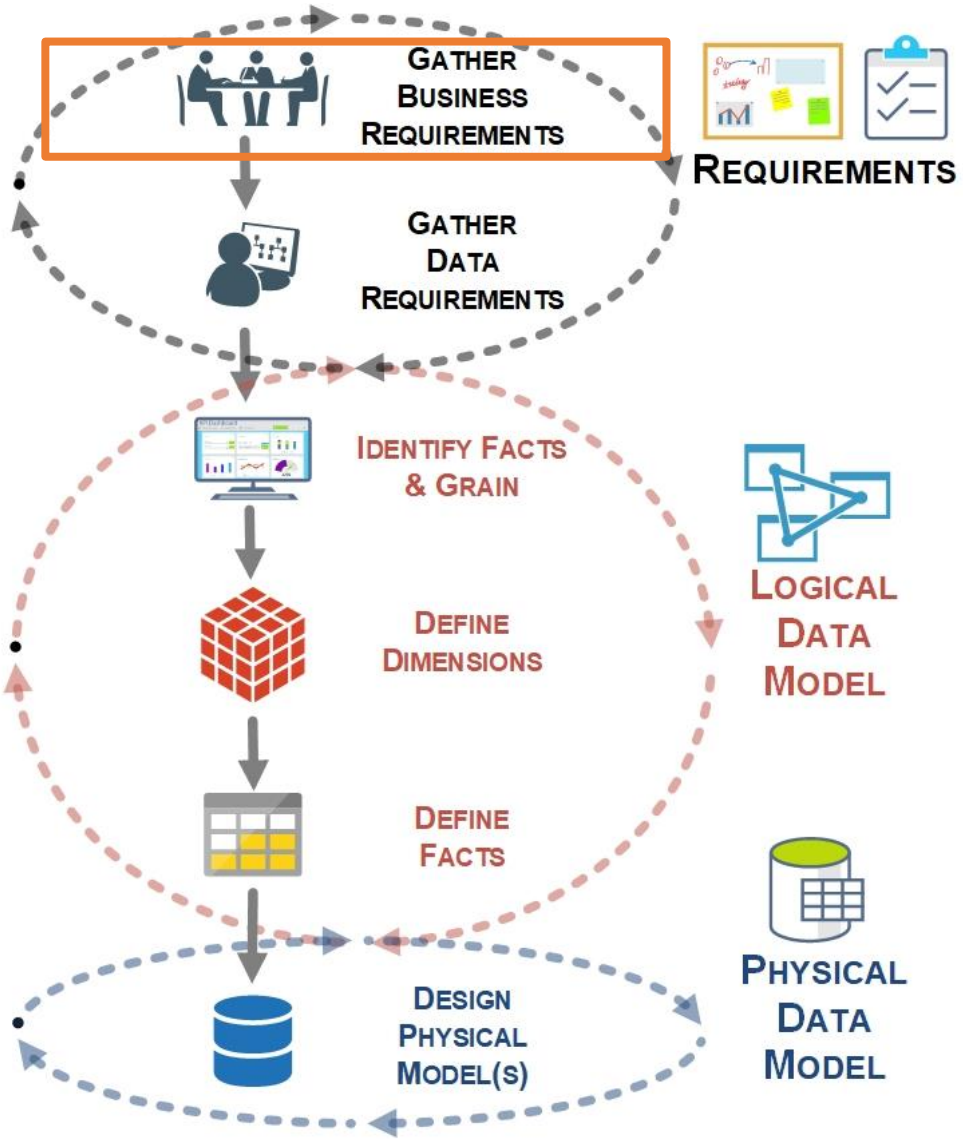
Deliverables

- Convert ER Model to Dimensional Model
  - List fact(s) & dimensions
  - What tables will be combined?
  - Create date/calendar dimension
  - Create tables with surrogate SKs, NKs & FKs
  - Create geography table
  - Determine table attributes
  - Map source table(s) to target table

Data Modeling Lifecycle

Business Requirements

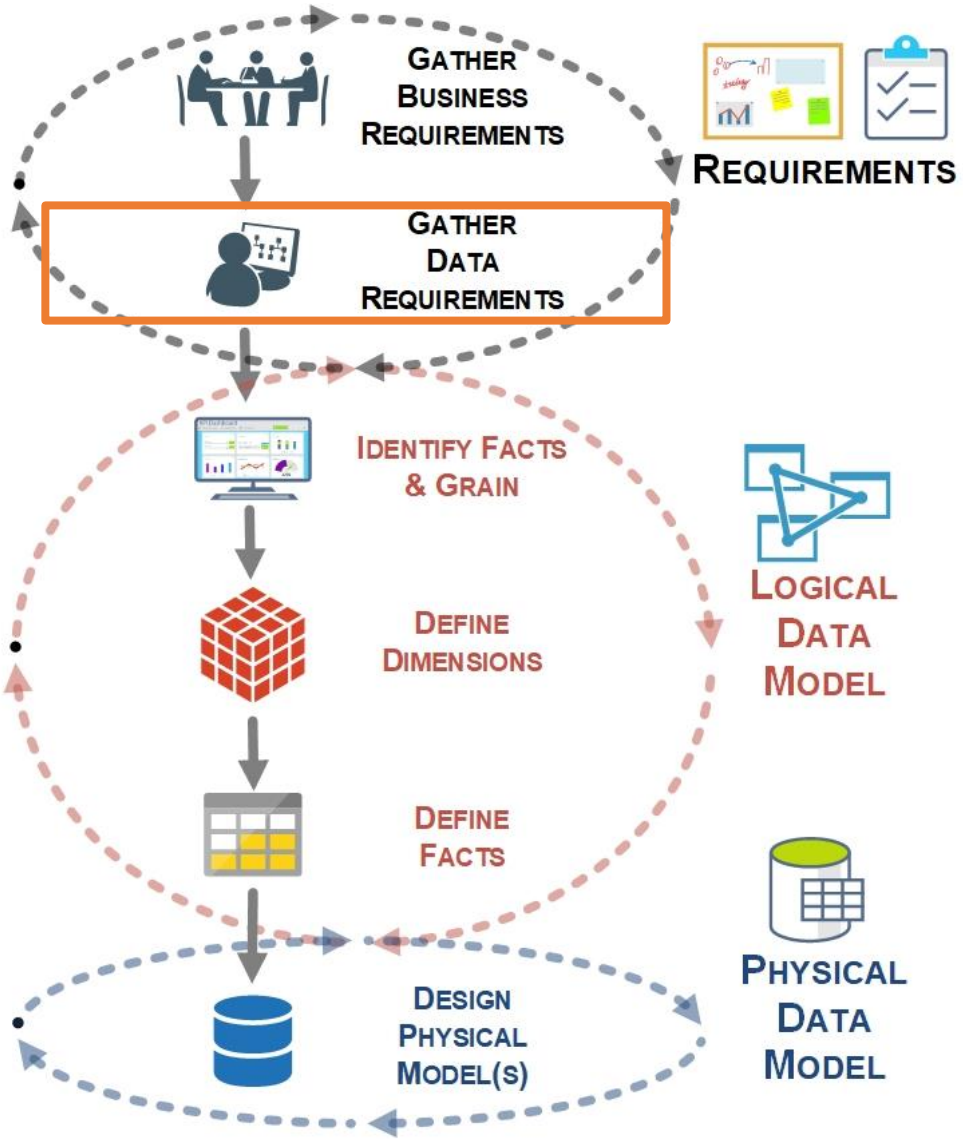
- Gather, analyze & prioritize business requirements
- Identify business processes or business analysis
- Identify high level entities and measures (metrics)



Data Modeling Lifecycle

Data Requirements

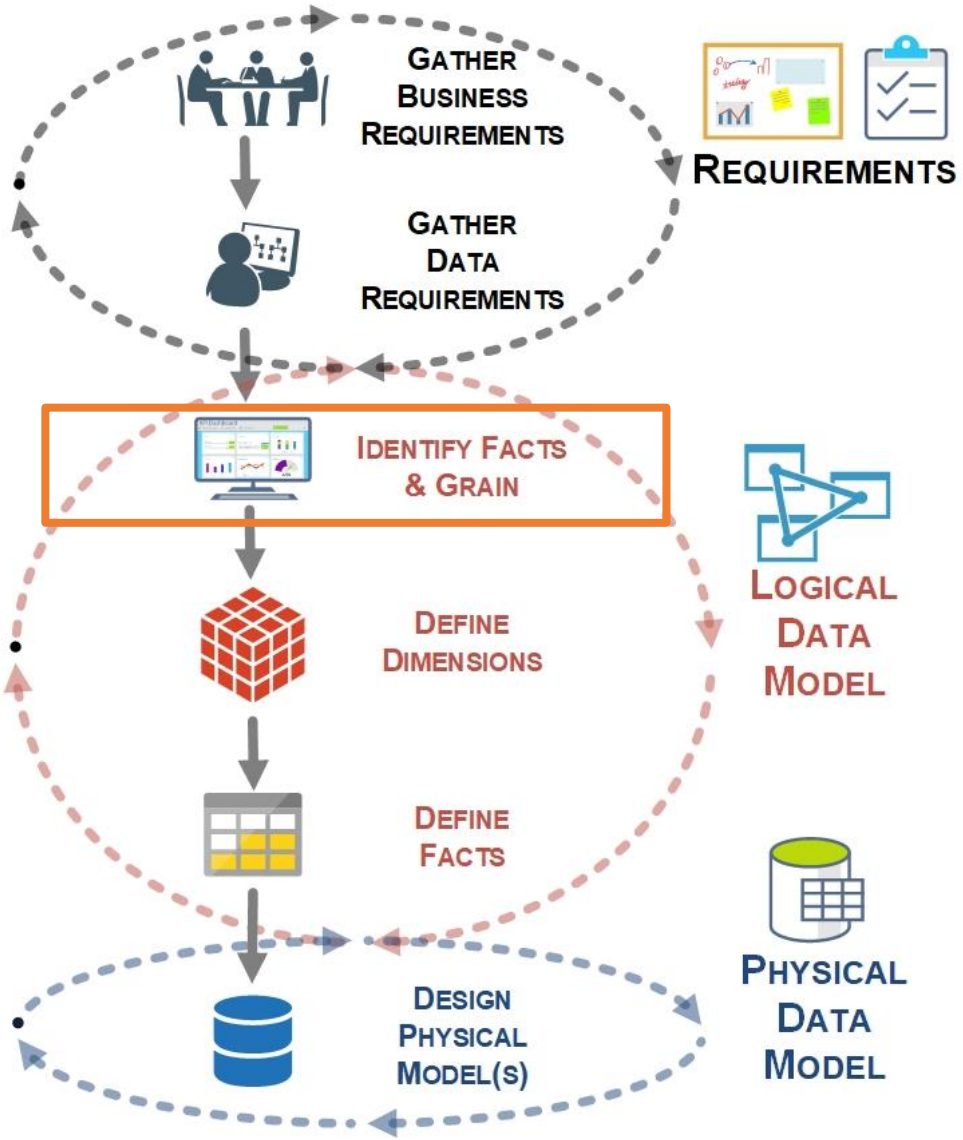
- Identify data sources
- Determine if data requirements is user-based or source-based
- Review existing data models or data structures
- Perform data profiling



Data Modeling Lifecycle

Identify Facts & Determine Grain

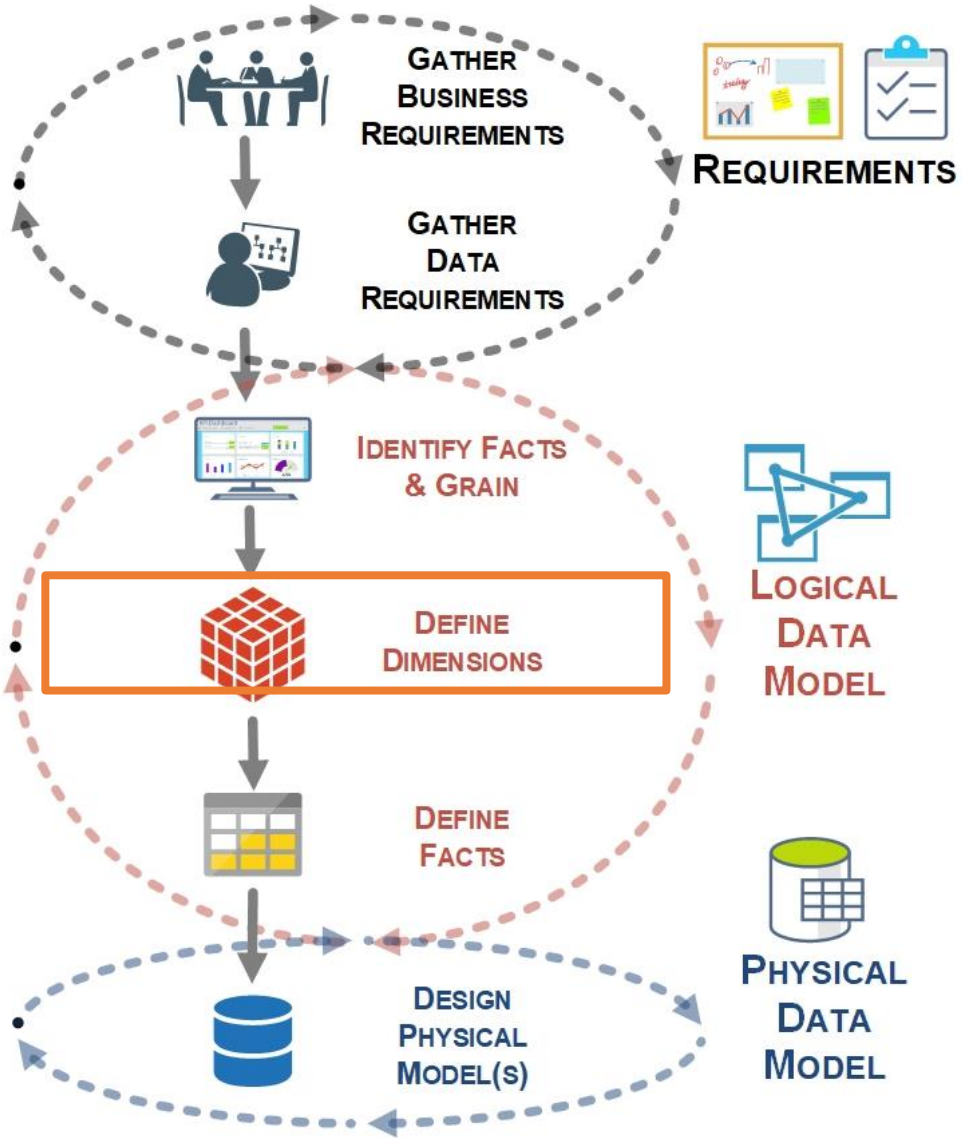
- Identify grain(s) in business processes
- Identify Fact Tables
- Identify Fact Table Types
  - Transaction, Periodic & Accumulating
- Identify Fact Table Granularity
- Identify preliminary dimensions



Data Modeling Lifecycle

Define Dimensions

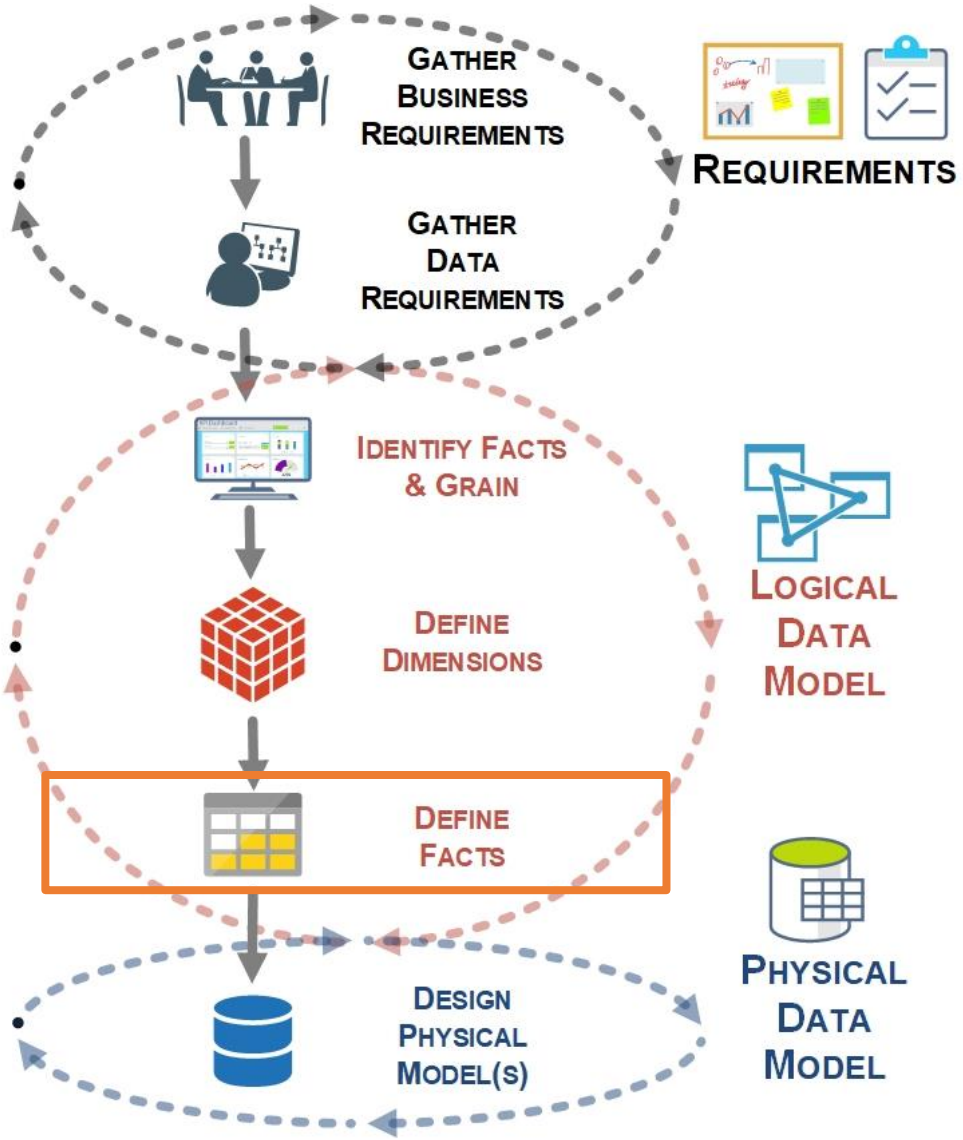
- Determine all dimensions
- Identify degenerate & conformed dimensions
- Identify dimensional attributes & validate granularity
- Identify hierarchies & attributes
- Identify date & time attributes
- Identify slowly changing dimensions (SCD) & types
- Identify multi-valued dimensions & define approach
- Identify role-playing dimensions
- Identify & classify specialized dimensions
  - Junk, Rapidly Changing, Hot Swappable, etc.
- Define surrogate keys (SKs), identify natural keys (NKs) and alternative keys (AKs)
- Define change data capture (CDC) attributes



Data Modeling Lifecycle

Define Facts

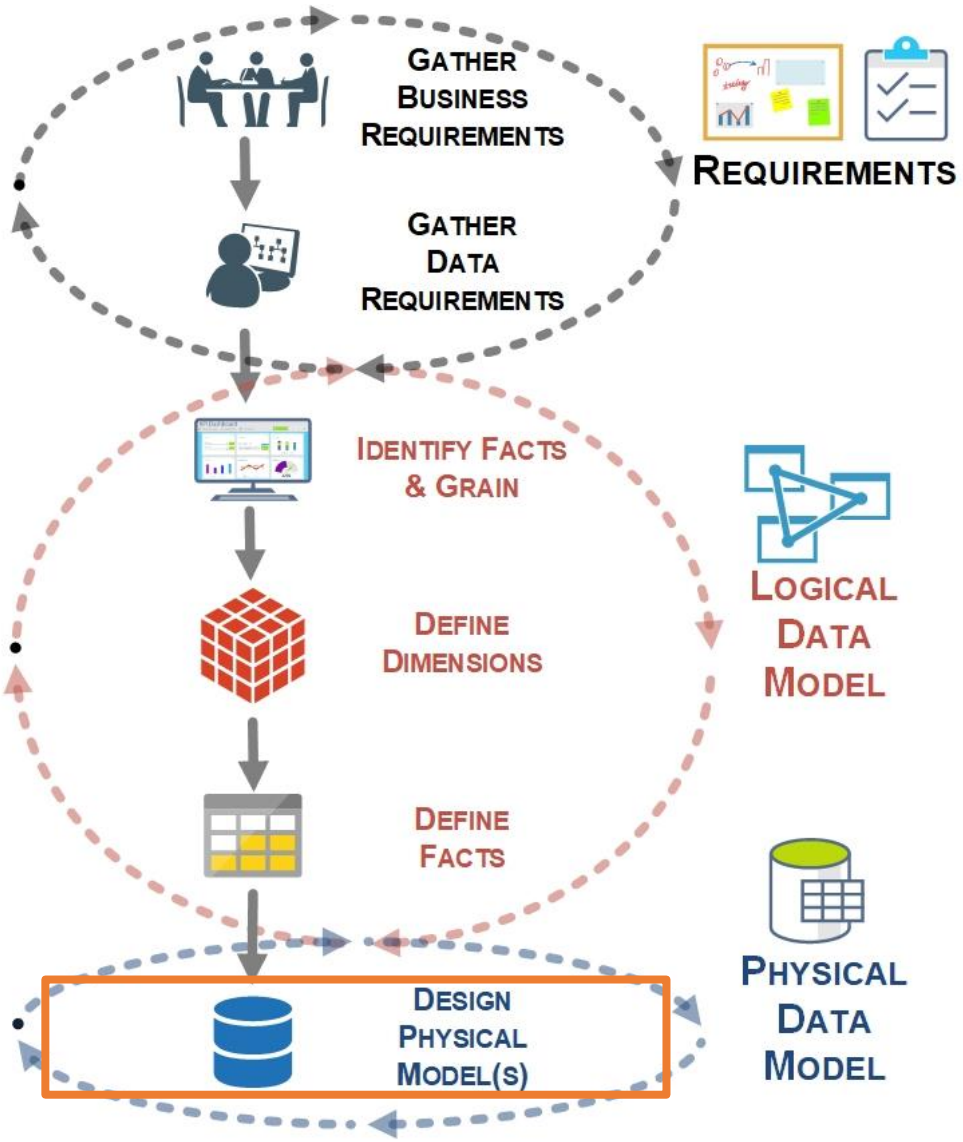
- Determine all facts
- Identify conformed facts
- Identify fact attribute types
  - Additive, semi-additive & non-additive
- Identify derived attributes & define approach
- Identify aggregates with associated hierarchies & define approach
- Identify composite keys & design PK approach
- Identify “snapshot” tables & define approach
- Identify event tables & define approach



Data Modeling Lifecycle

Design Physical Data Model

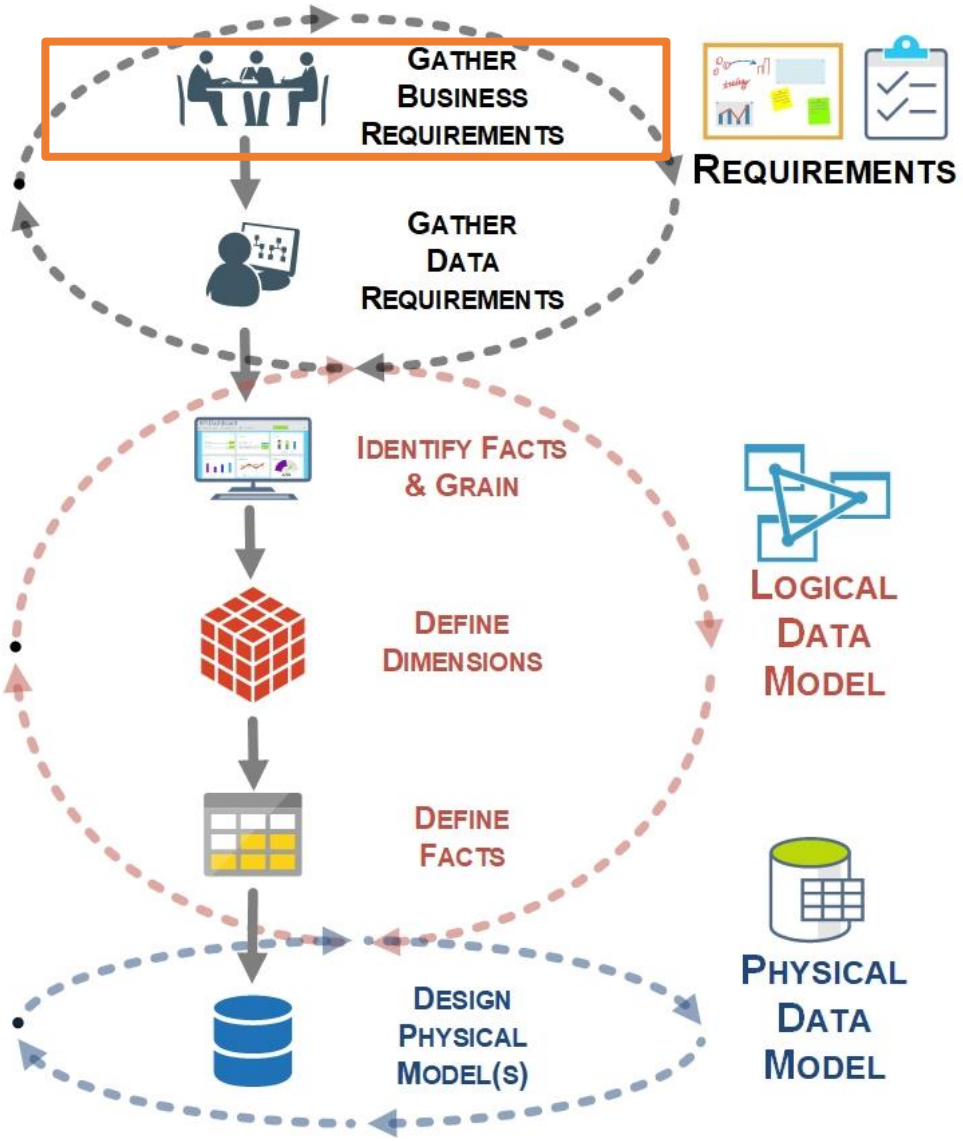
- Estimate dimension & fact tables sizing & growth
- Determine target database(s)
  - DBMS type
  - Specific DBMS
- Define tables according to specific DBMS
- Define keys as appropriate - PKs, SKs, FKs
- Determine use cases for views such as role-playing dimensions
- Define performance tuning approach
  - Different types of indexes, partitioning, etc.



Data Modeling Lifecycle

Business Requirements

- Gather, analyze & prioritize business requirements
- Identify business processes or business analysis
- Identify high level entities and measures (metrics)



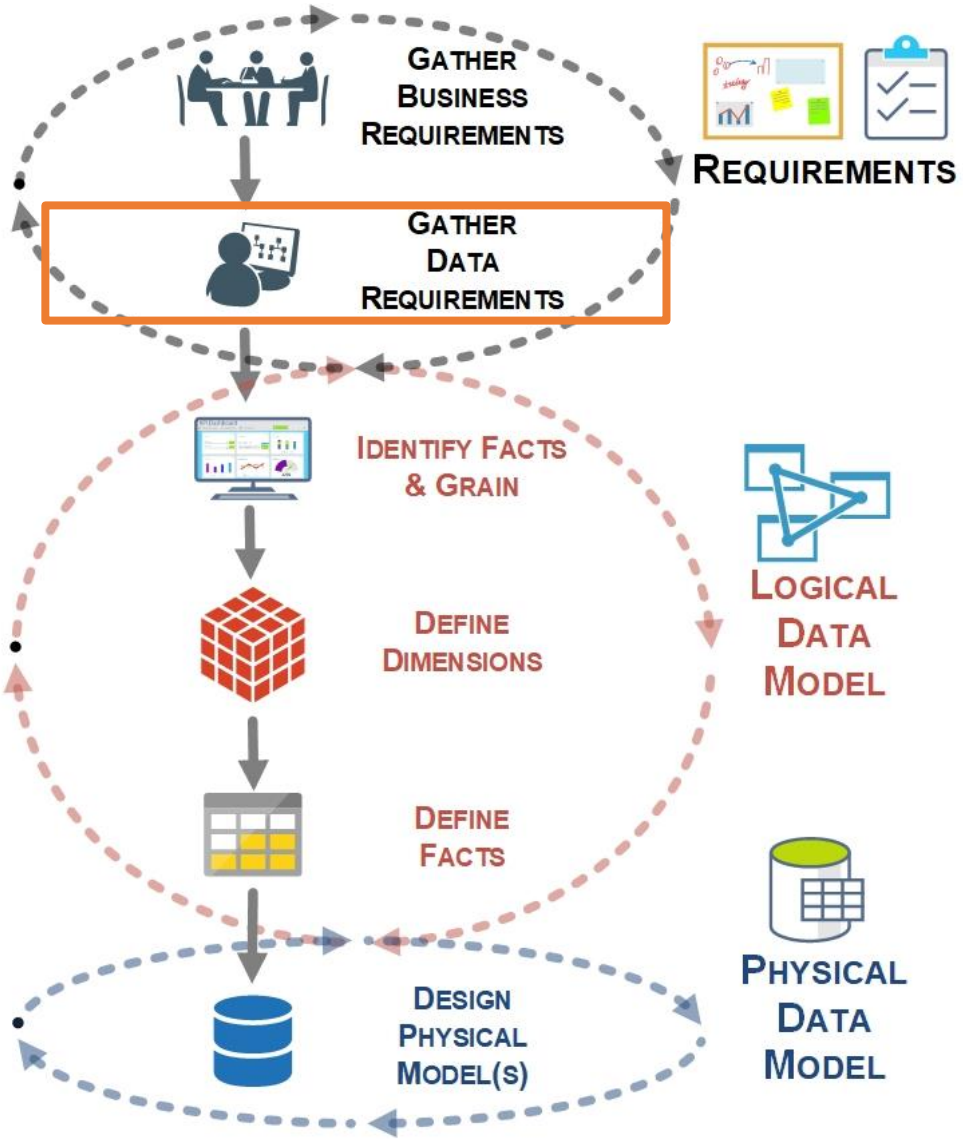
Chinook: Business Requirements

- Create & run query for each database: place SQL queries in Word document & paste query results into an individual worksheet in an Excel spreadsheet
  1. Total sales $
  2. Total sales $ by country – ranked (or at least sorted largest to smallest)
  3. Total sales $ by country, state & city
  4. Total sales $ by customer (a person with last name & first name) – ranked (or at least sorted largest to smallest)
  5. Total sales $ by artist – ranked (or at least sorted largest to smallest)
  6. Total sales $ by albums
  7. Total sales $ by sales person (employee)
  8. Total tracks bought and total revenue $ by media type
  9. Total sales $ by genre
  10. Total sales $ by company
- Create data visualizations for above

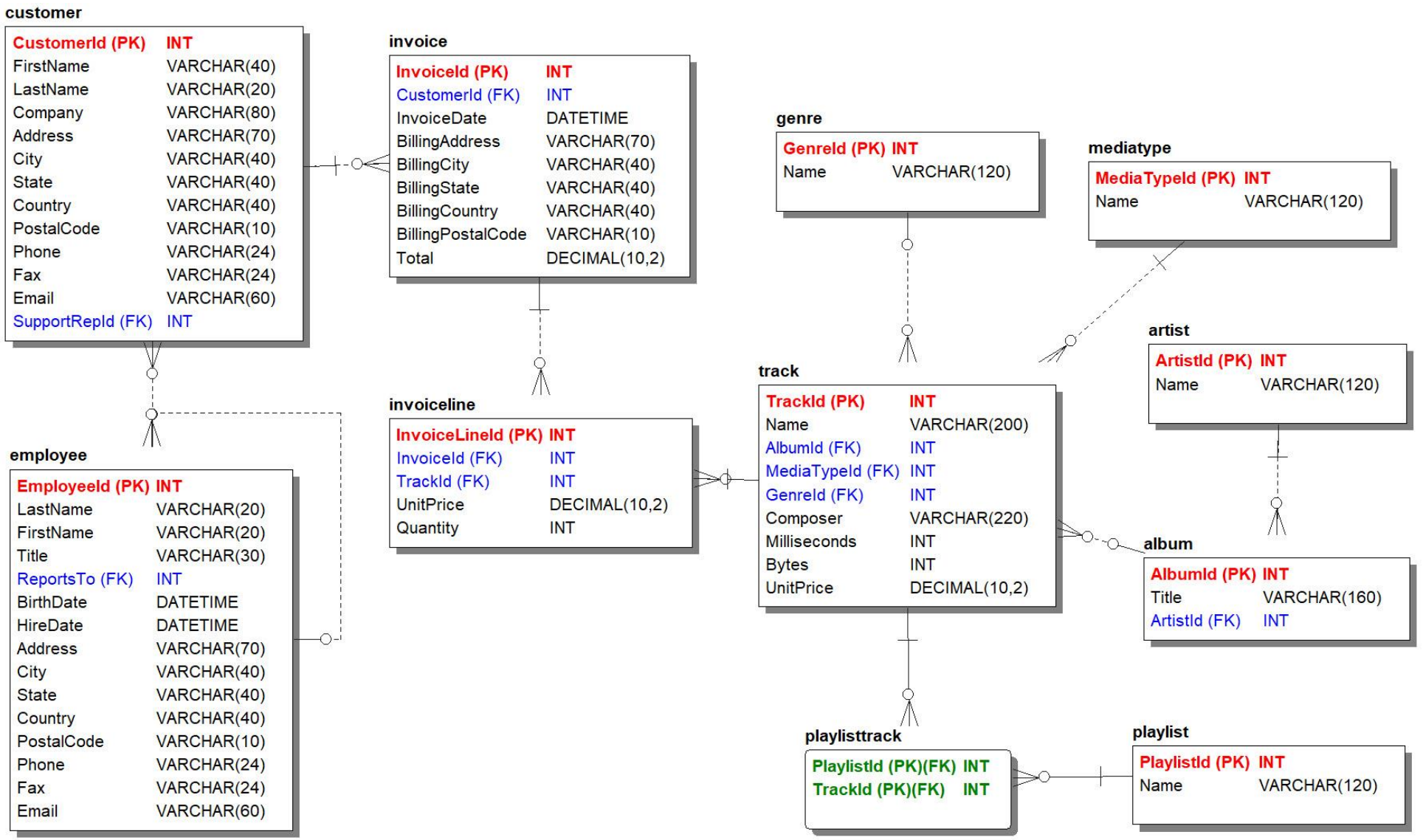
Data Modeling Lifecycle

Data Requirements

- Identify data sources
- Determine if data requirements is user-based or source-based
- Review existing data models or data structures
- Perform data profiling



Chinook Database: ER Model

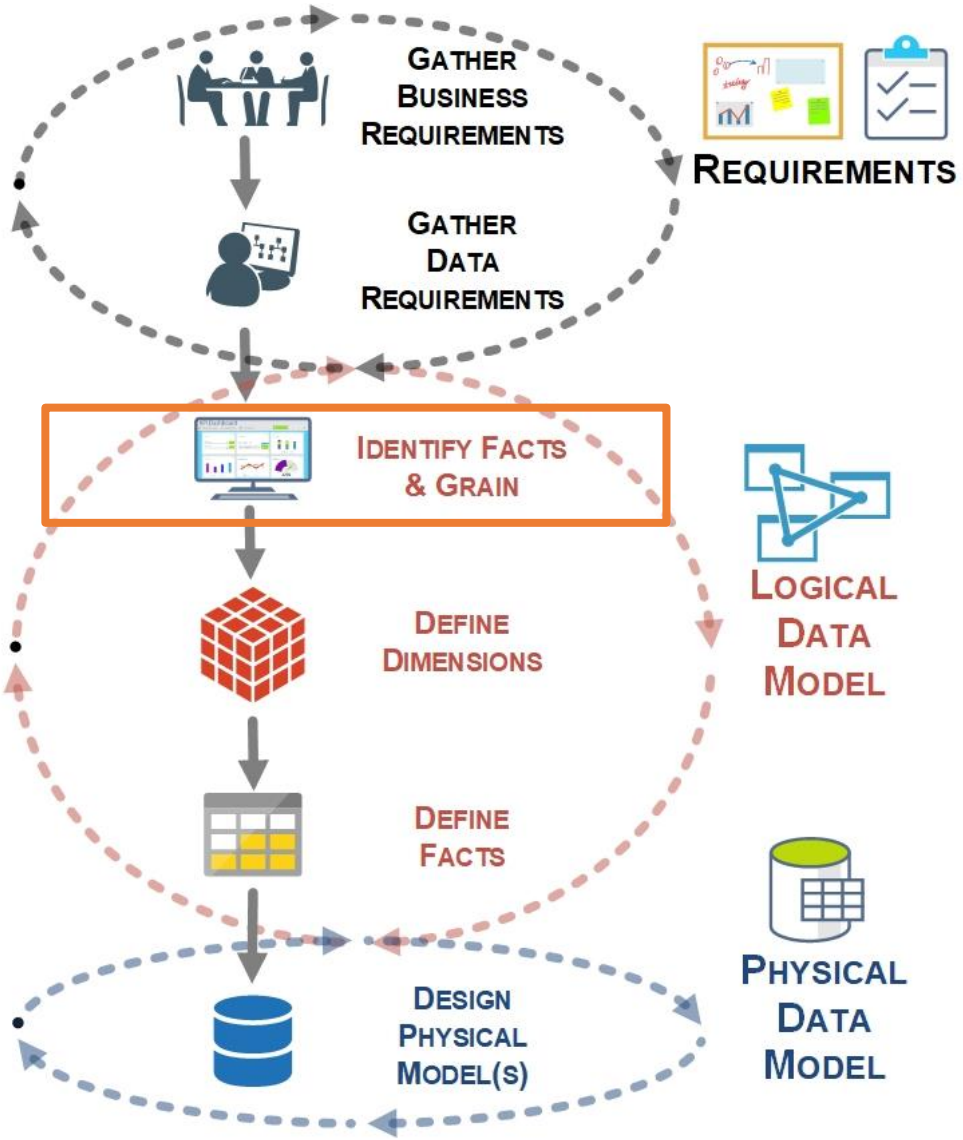


ATHENA IT SOLUTIONS

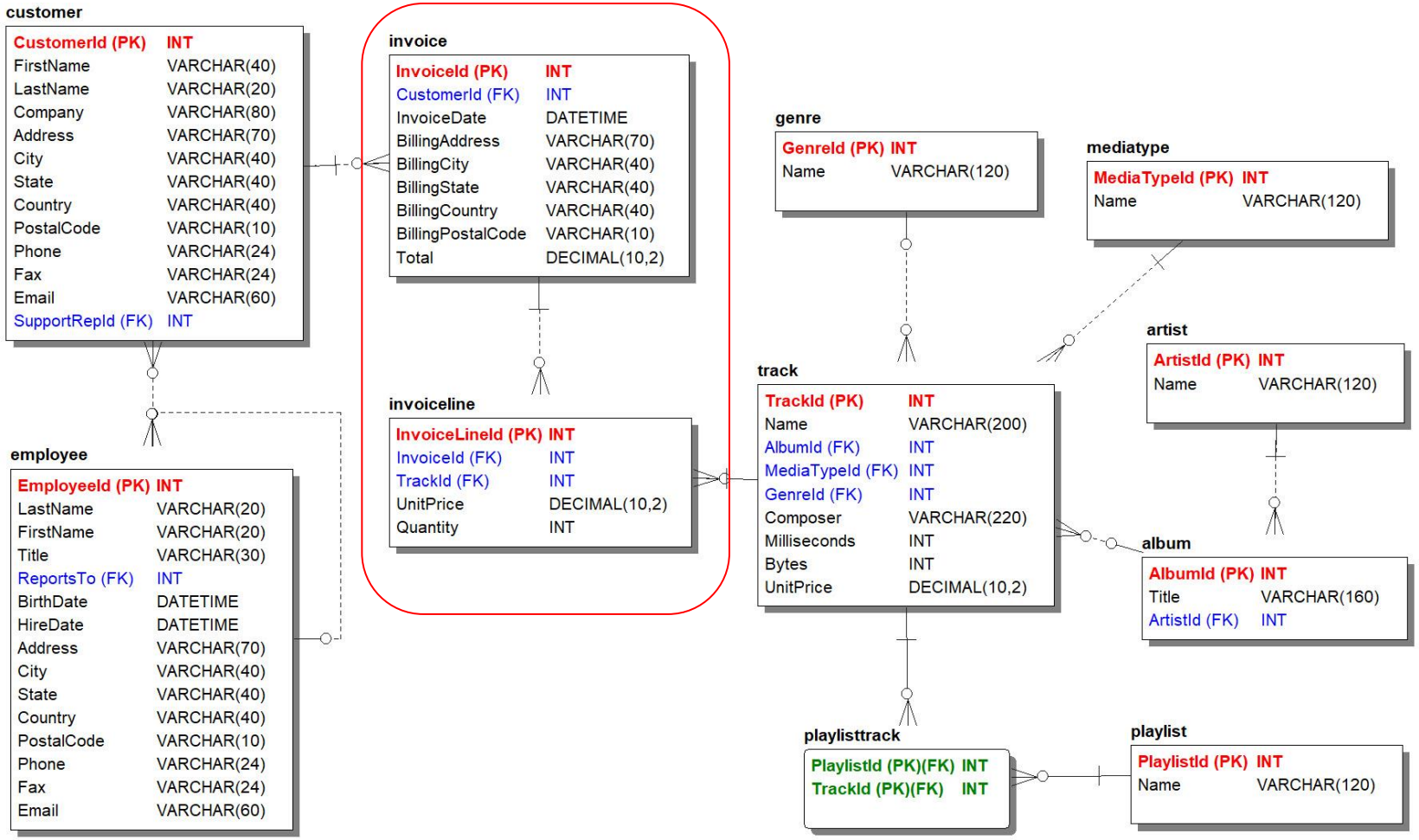
Data Modeling Lifecycle

Identify Facts & Determine Grain

- Identify grain(s) in business processes
- Identify Fact Tables
- Identify Fact Table Types
  - Transaction, Periodic & Accumulating
- Identify Fact Table Granularity
- Identify preliminary dimensions



Identify Facts & Determine Grain

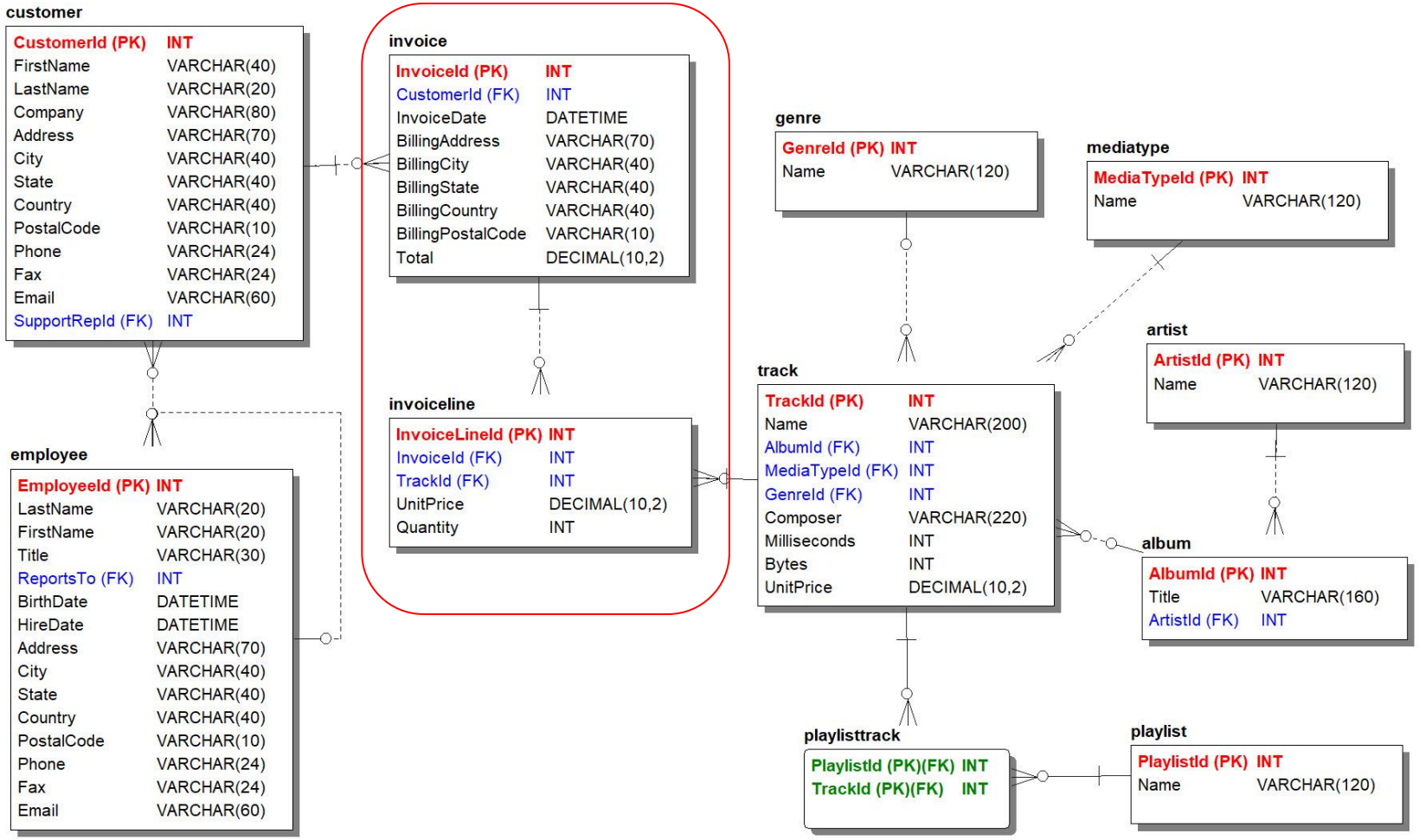


- Identify Facts
  - Invoice
  - InvoiceLine

Chinook Dimensional Data Model

Determine Dimensions

Identify Facts & Determine Grain

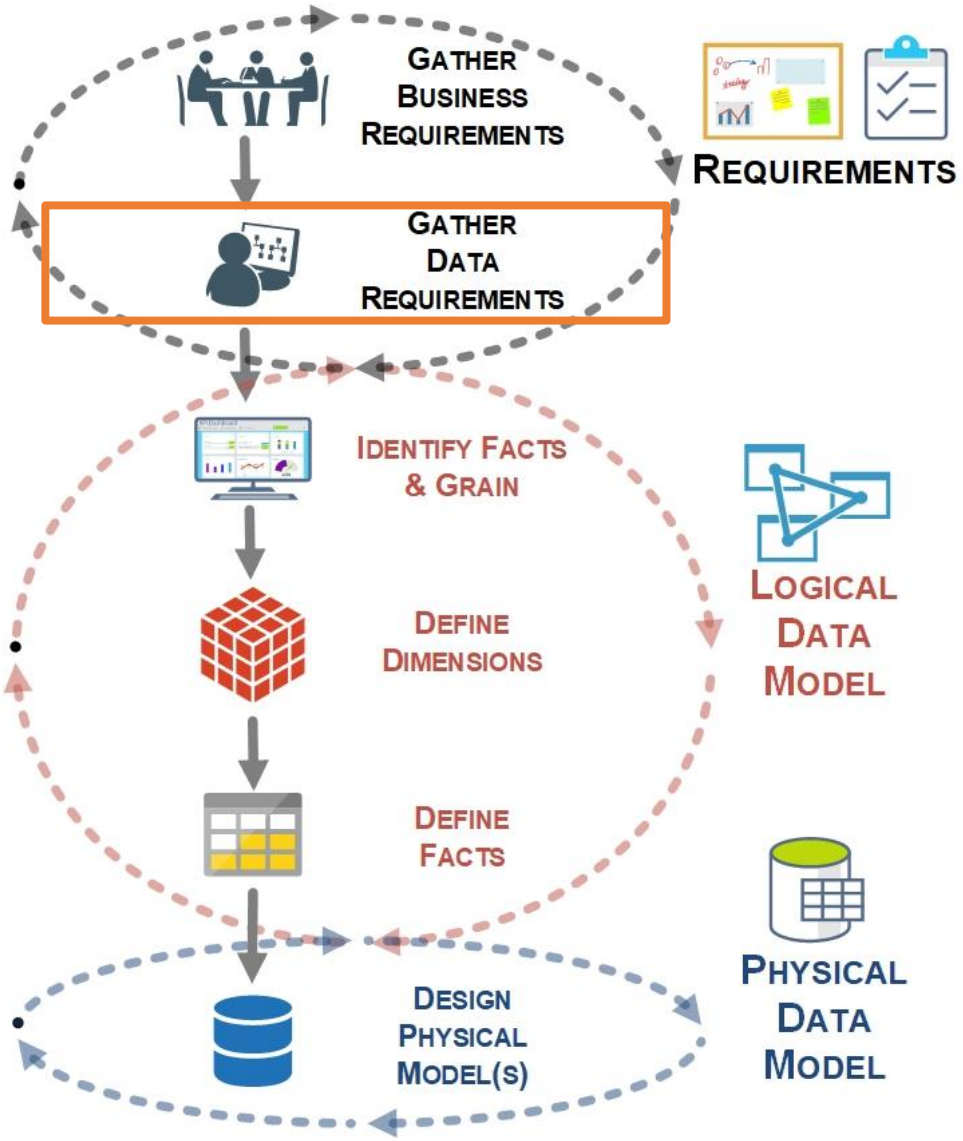


- Identify Facts
  - Invoice
  - InvoiceLine

Data Modeling Lifecycle

Data Requirements

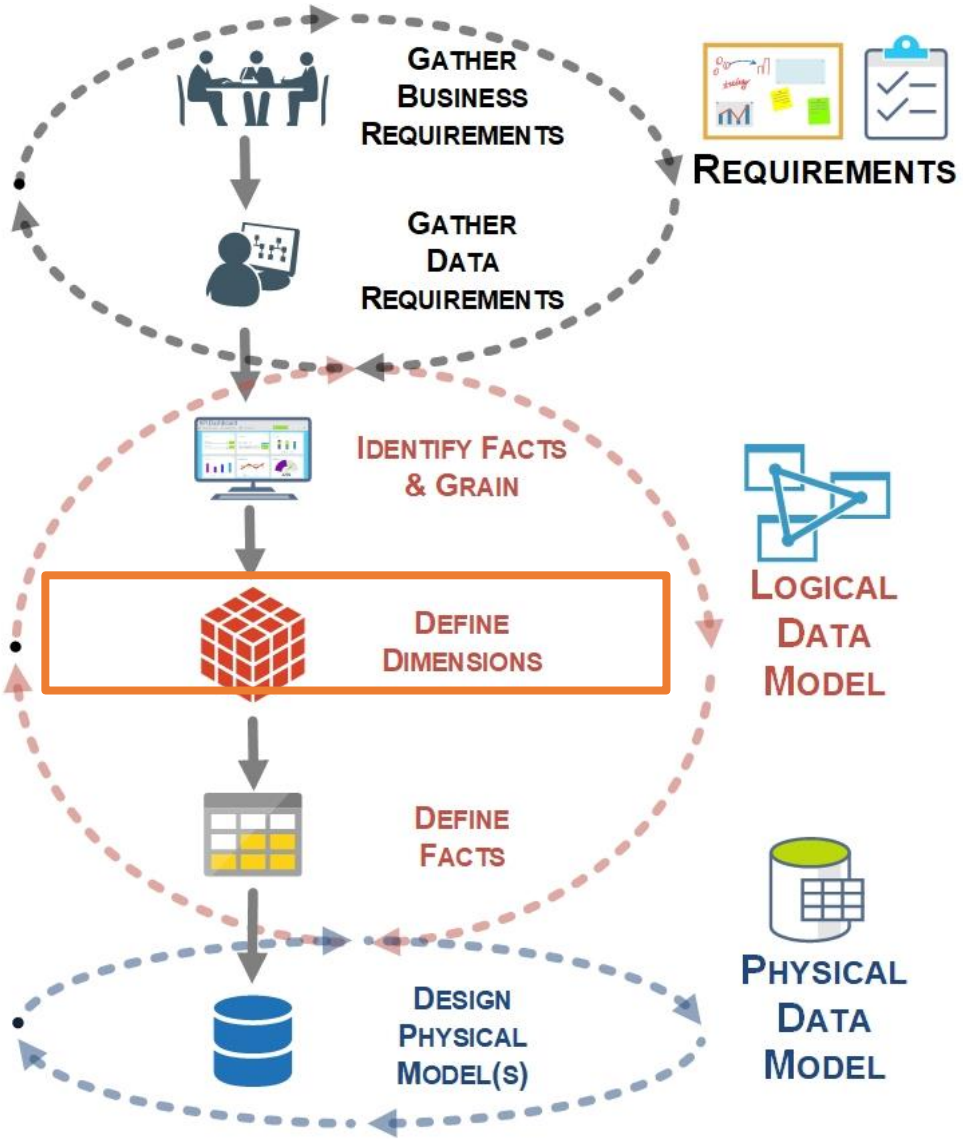
- Identify data sources
- Determine if data requirements is user-based or source-based
- Review existing data models or data structures
- Perform data profiling



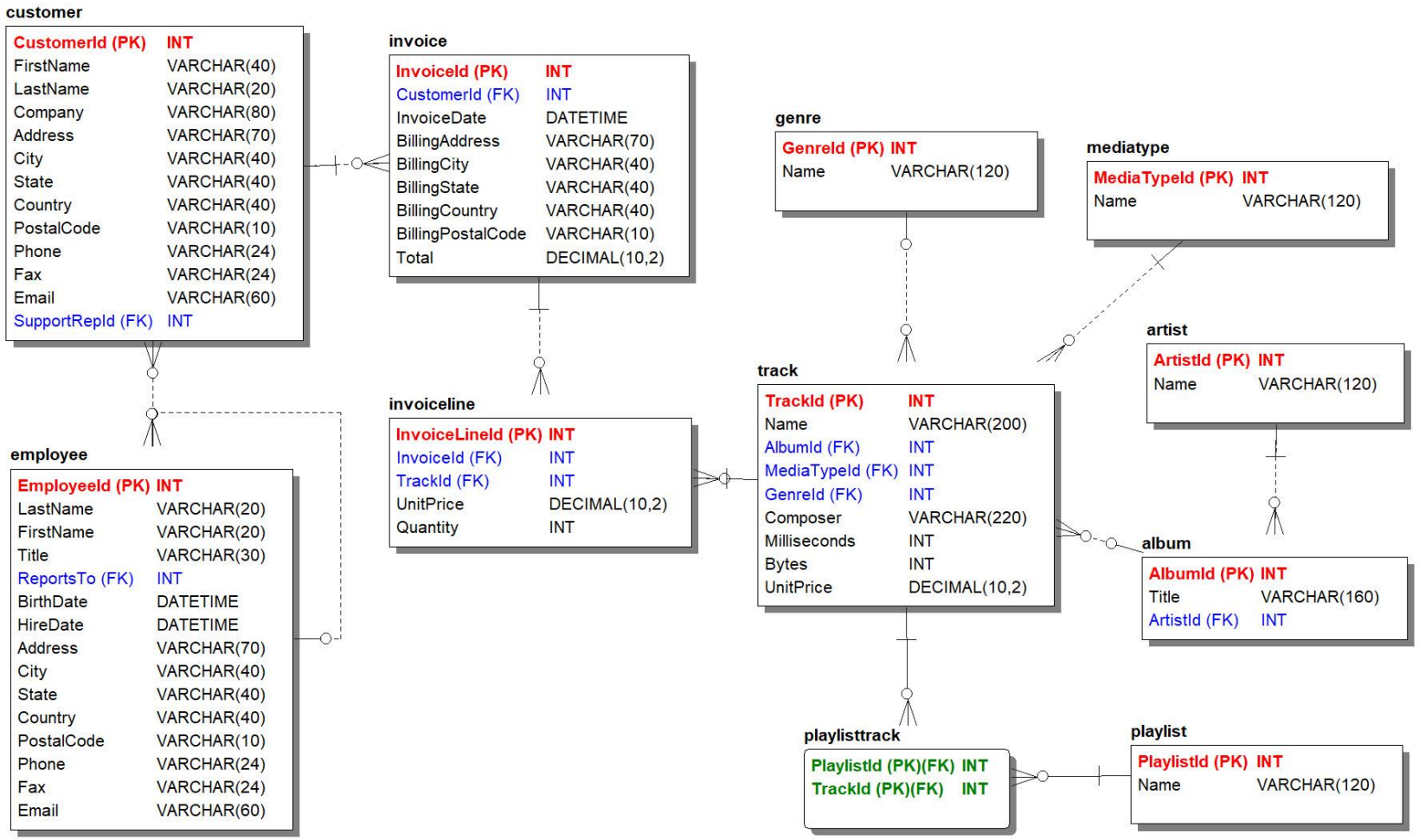
Data Modeling Lifecycle

Define Dimensions

- Determine all dimensions
- Identify degenerate & conformed dimensions
- Identify dimensional attributes & validate granularity
- Identify hierarchies & attributes
- Identify date & time attributes
- Identify slowly changing dimensions (SCD) & types
- Identify multi-valued dimensions & define approach
- Identify role-playing dimensions
- Identify & classify specialized dimensions
  - Junk, Rapidly Changing, Hot Swappable, etc.
- Define surrogate keys (SKs), identify natural keys (NKs) and alternative keys (AKs)
- Define change data capture (CDC) attributes

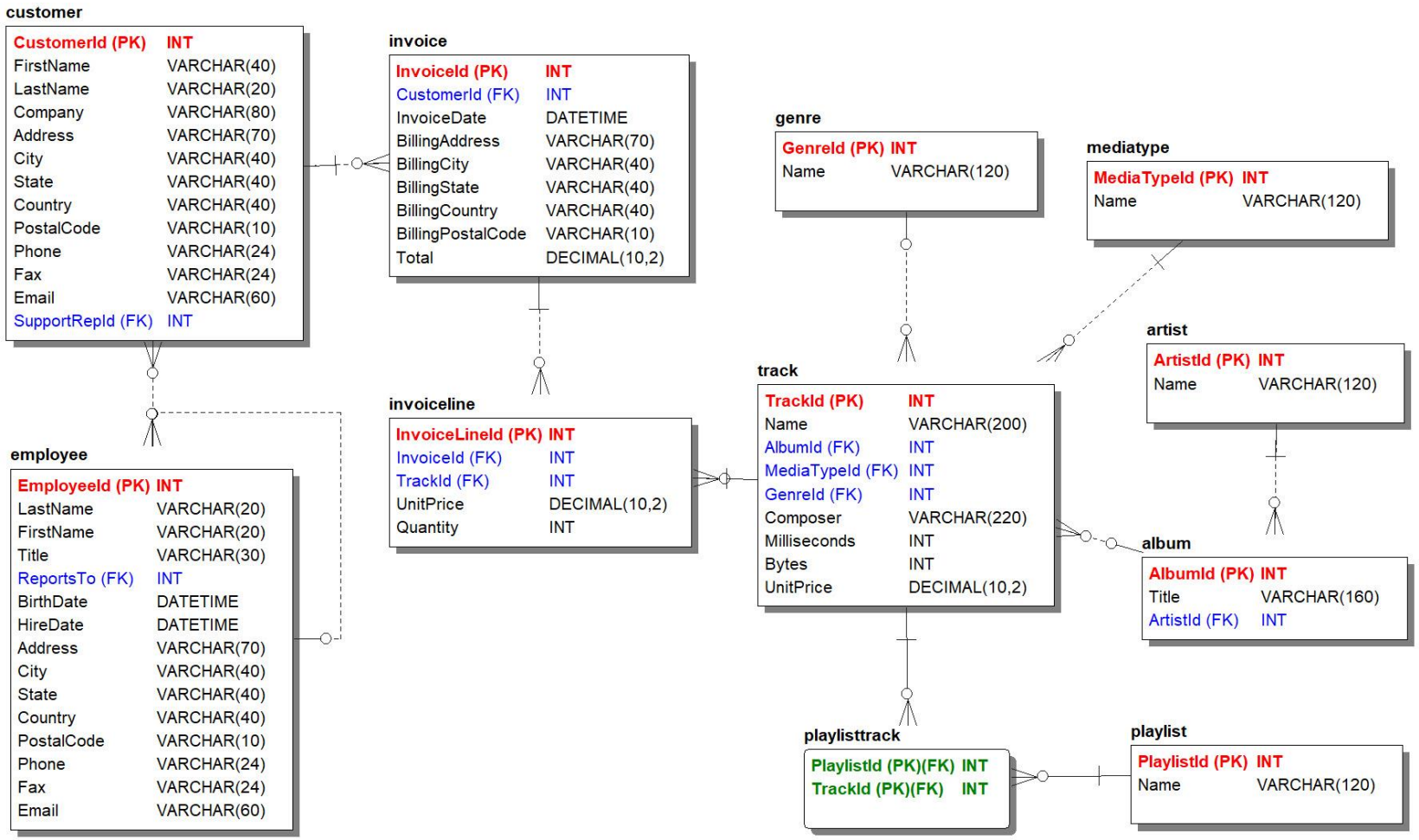


Determine all dimensions



- Initial draft of dimensions:
  - Album
  - Artist
  - Customer
  - Employee
  - Genre
  - MediaType
  - Playlist
  - Track
- Bridge table:
  - PlaylistTrack

Determine all dimensions



Identify Bridge table:

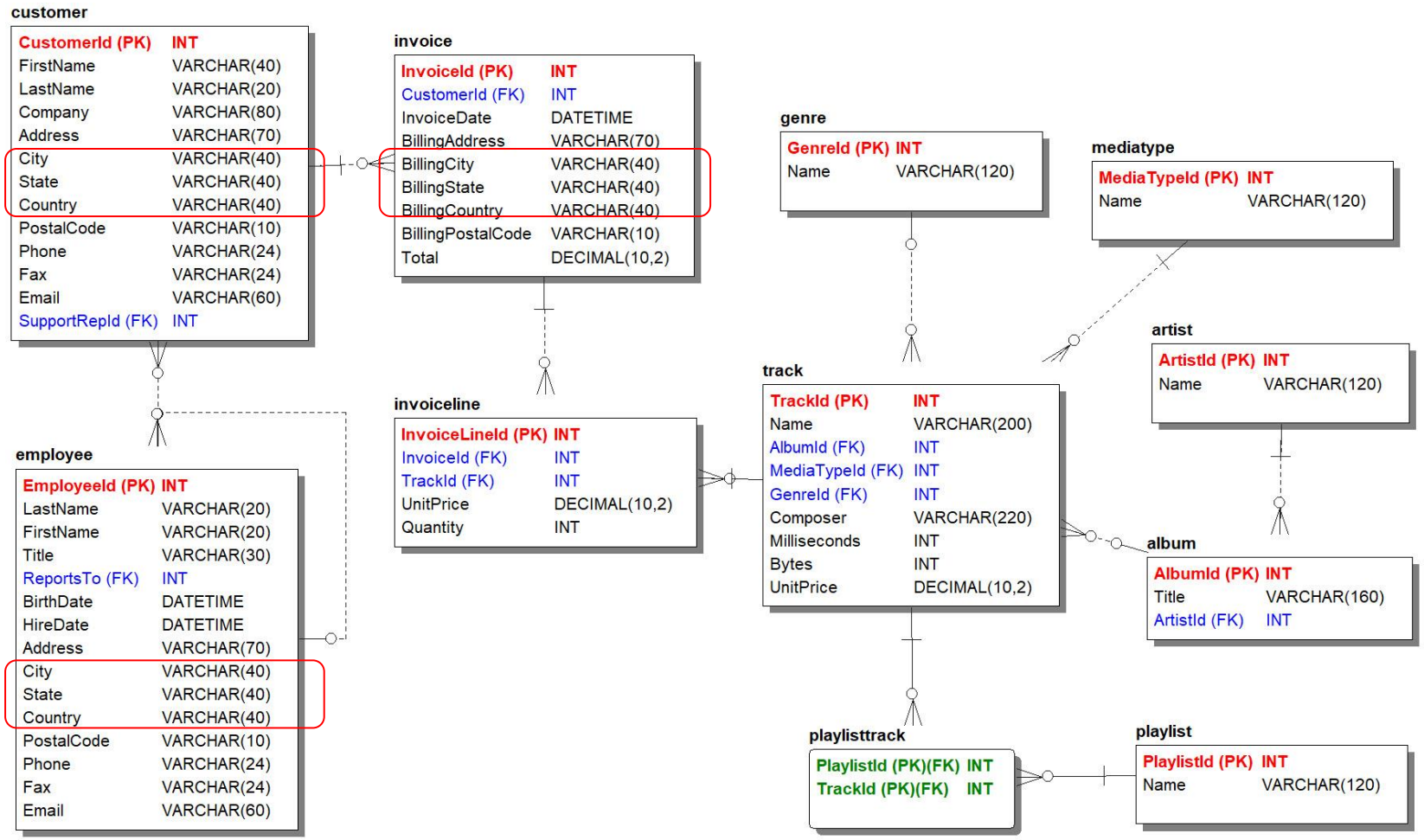
- PlaylistTrack

Notes:

There would be other bridge tables IF there were other many-to-many relationships

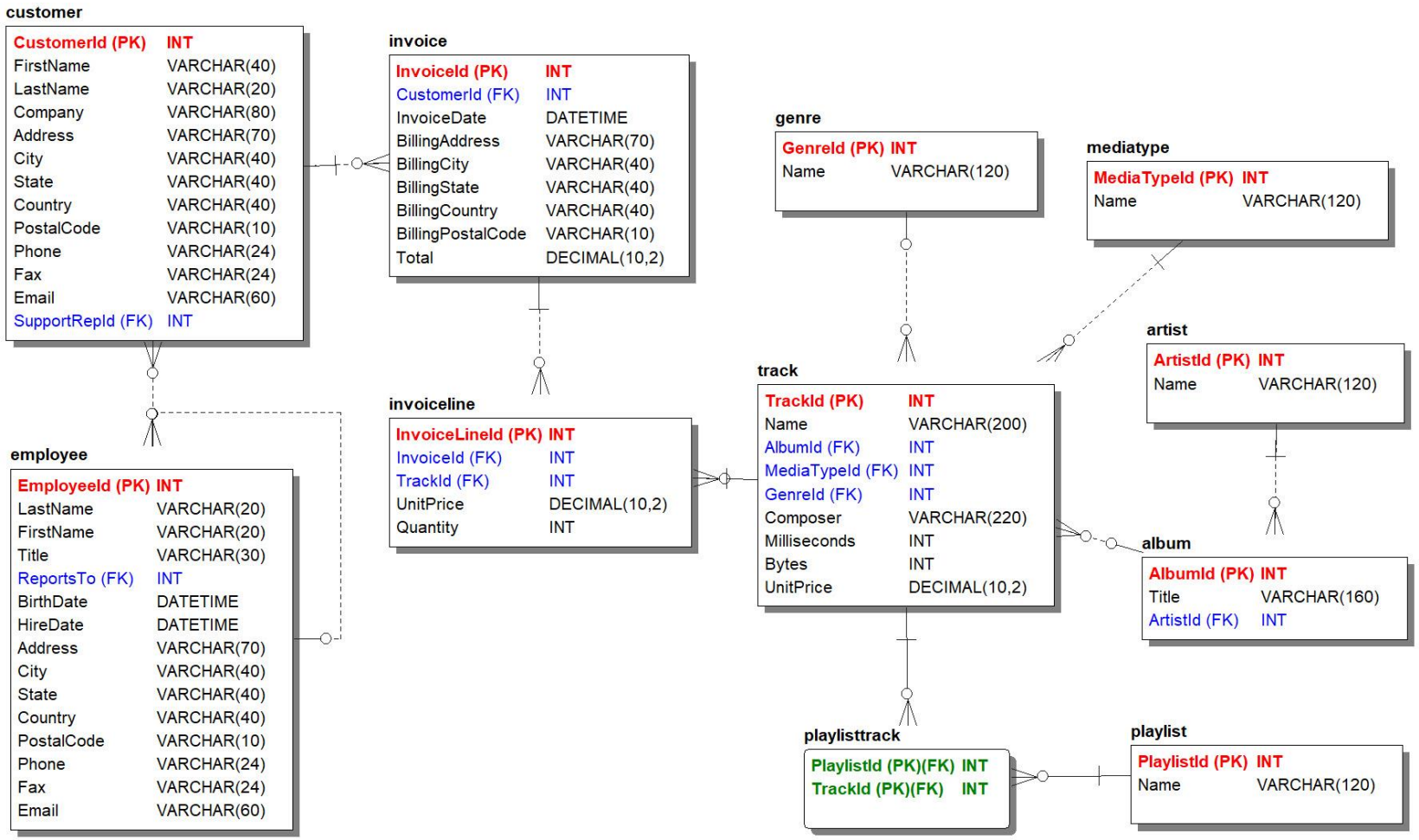
- Track can only have one Genre
- Track can only be on one Album
- Album can only have one Artist
- Track can only have one MediaType

Determine all dimensions



- Identify outrigger(s):
  - DimGeography
    - City
    - State
    - Country
  - An alternative would be DimAddress
    - Address
    - City
    - State
    - Country
    - PostalCode
- Create role playing dimensions (as Views) from Outriggers

Determine all dimensions



- Create a DimDate dimension and store dates as Surrogate Key (SK), i.e. YYYYMMDD

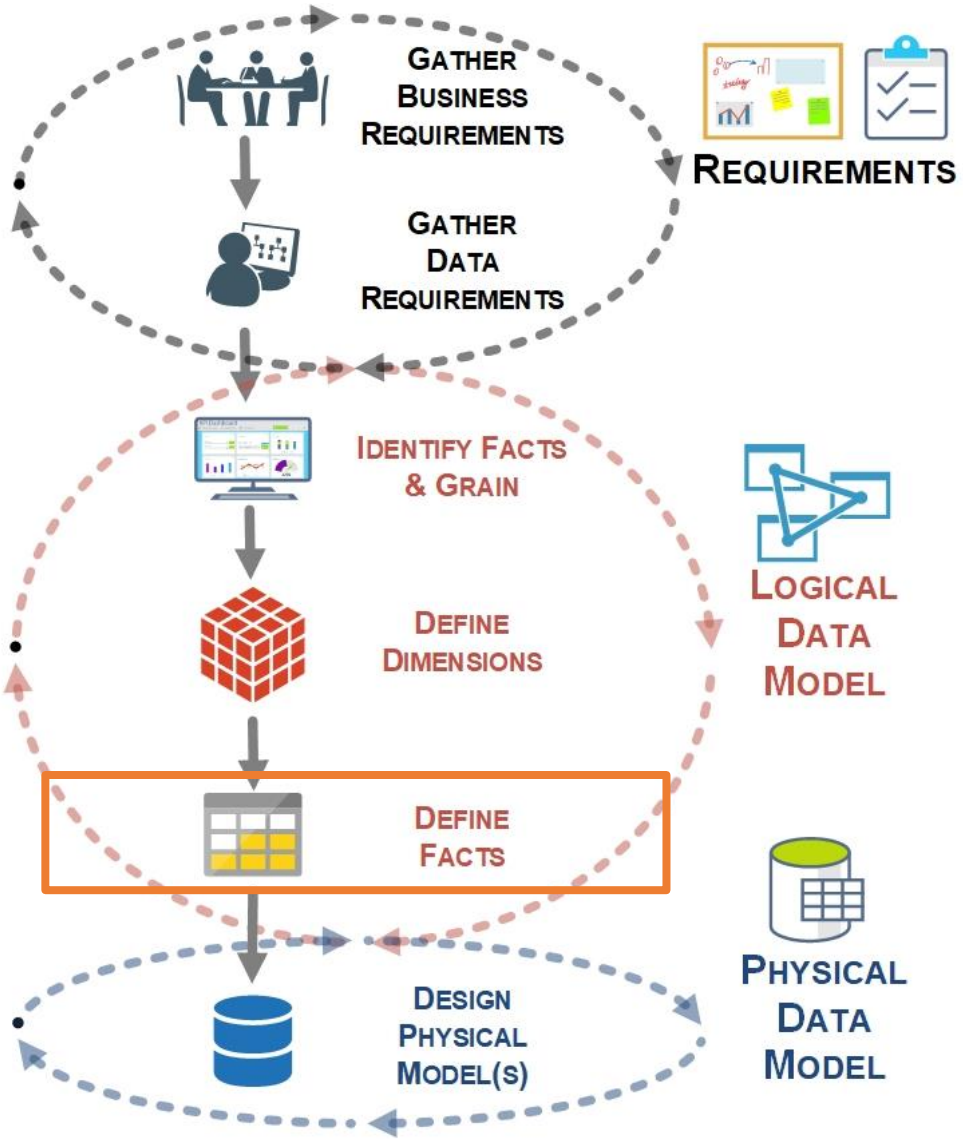
Chinook Dimensional Data Model

Determine Fact(s)

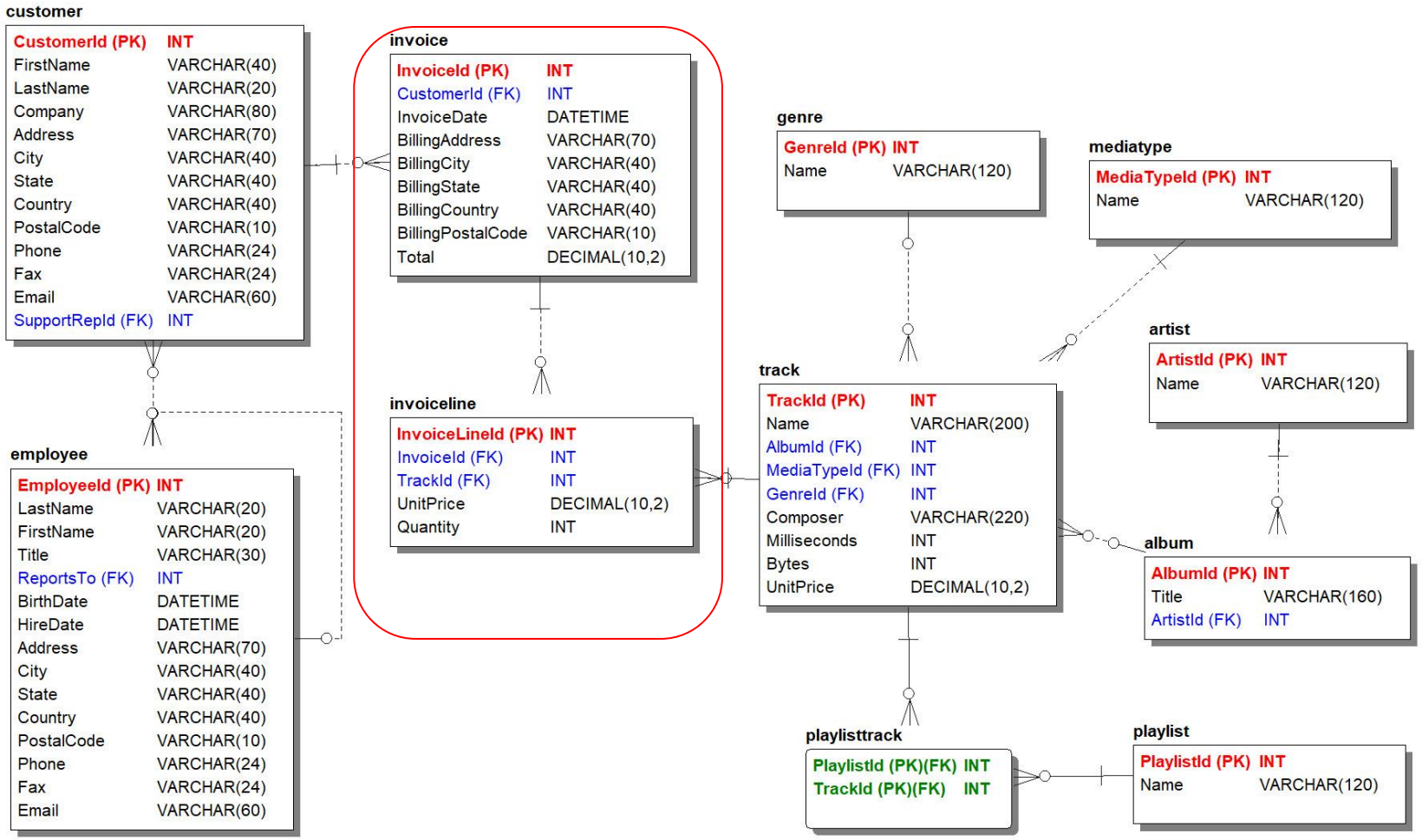
Data Modeling Lifecycle

Define Facts

- Determine all facts
- Identify conformed facts
- Identify fact attribute types
  - Additive, semi-additive & non-additive
- Identify derived attributes & define approach
- Identify aggregates with associated hierarchies & define approach
- Identify composite keys & design PK approach
- Identify “snapshot” tables & define approach
- Identify event tables & define approach

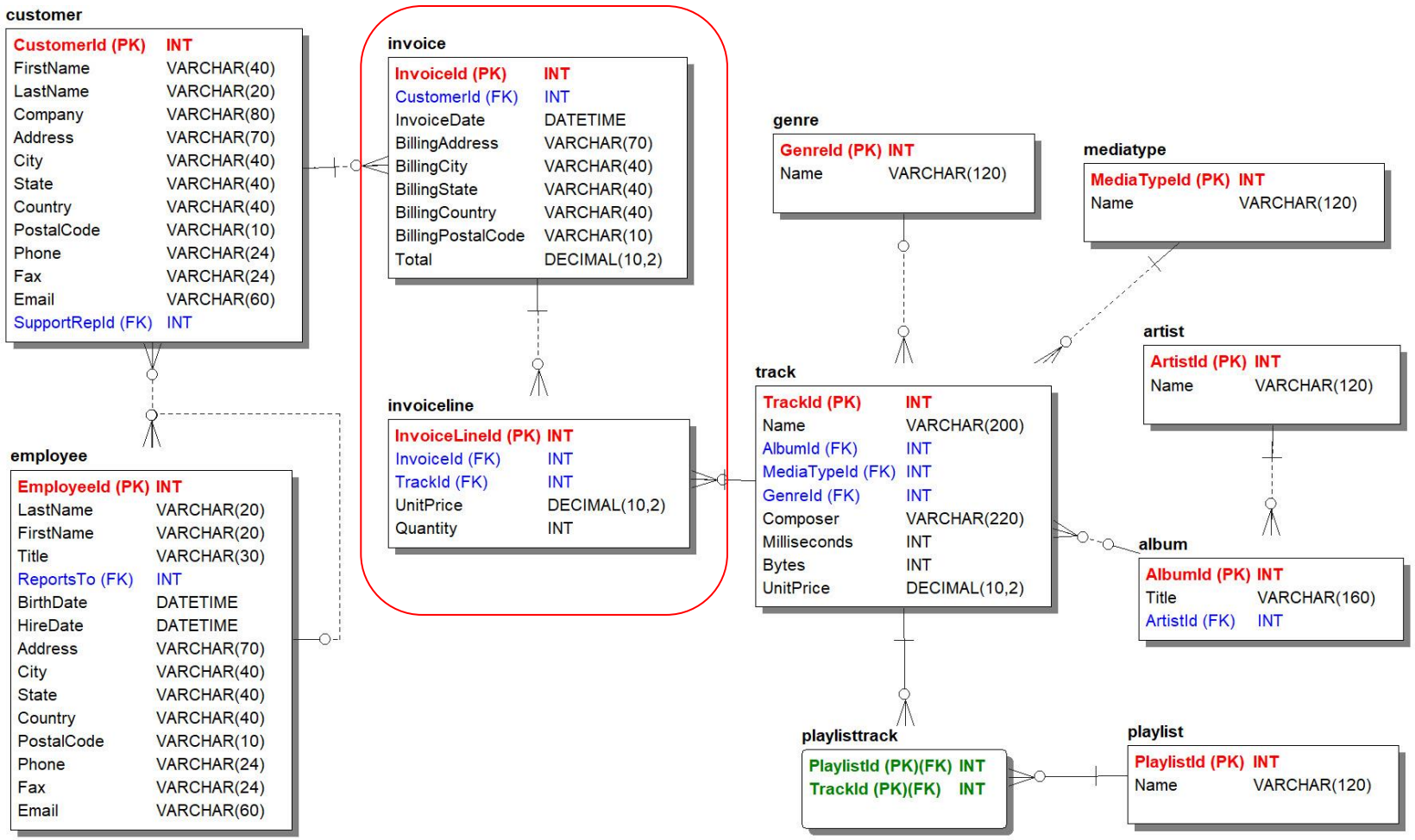


Define Facts



- Classic Header & Line Item entity examples:
  - Sales
  - Orders
  - Invoices
  - Purchases
- Handling Header & Line Item entities
  - Combine 2 entities & Denormalize
  - Granularity Consistency
  - Fact Attribute Consistency
  - Avoid “double counting”

Define Facts



- Sales (Fact)
  - Combine Invoice & InvoiceLine entities
  - UnitPrice removed
  - SalesTotal = UnitPrice * SalesQuantity
  - InvoiceID & InvoiceLineID are degenerate

Chinook

Dimensional Data Model

Chinook Dimensional Model