CIS 467 – Week 1

Chapters 1, 6 & 3 in Lemahieu text, Chapter 1 & 10 in Murach text

- **Introduction to Data Management**
  - Some definitions to get started and for future reference
    - Note: terminology varies across texts, platforms and applications – concept is the same
  - Database system basic elements

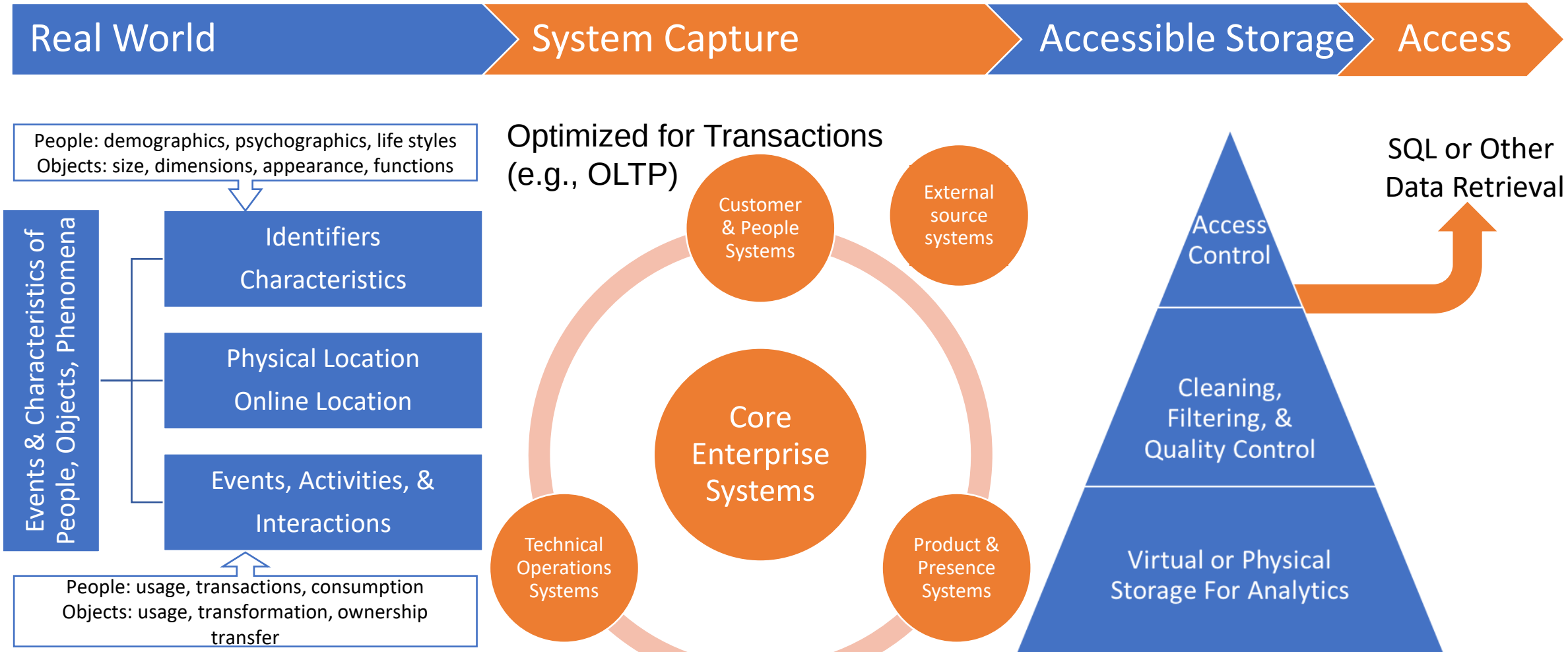
- **Database Design**
  - Use database specifications/real-world models to identify tables, columns and keys for a database
  - (E)ER diagrams, translating ER diagrams to relational DB schema

- **Data Normalization**
  - Rules of data normalization
  - Steps for normalizing database schema
  - Concepts of “tidy” data, transforming raw data to processed data

Database Terminology

- Data
  - known, recorded facts
- Database
  - a collection of data
- Database System
  - “People, Processes/Technology & Data”
- Database Management System (DBMS)
  - software component(s) that support database system functionality

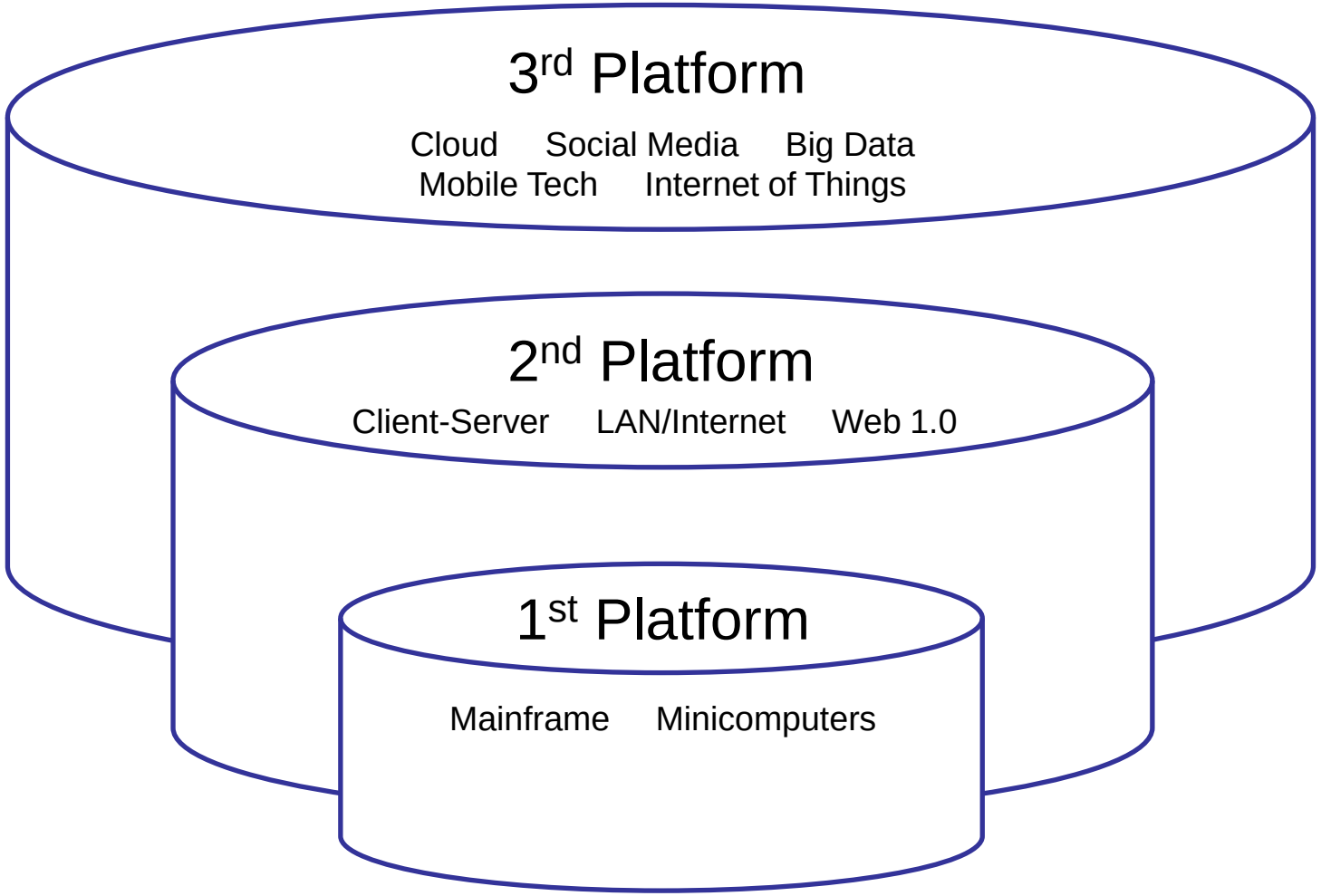
System Data: Origin to Retrieval for Analytics



Advantages to Using DBMS

- Data independence – changes in data definitions can be made with minimal impact on applications using the data
- Concurrency control - multiple users can access/update the same data in a supervised environment to avoid inconsistencies
- Data integrity can be programmatically monitored and enforced
- Reduce redundancy in data storage and in application development
- Backup data and data recovery; data security

Database Technology Evolution



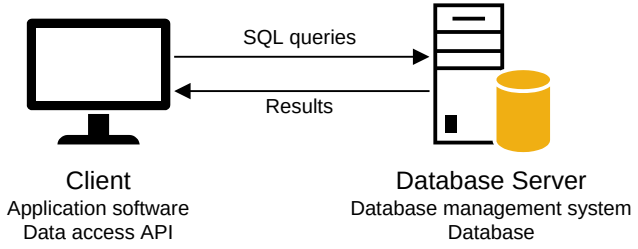
Relational Database
NOSQL
NewSQL
Big Data platforms

Relational Database

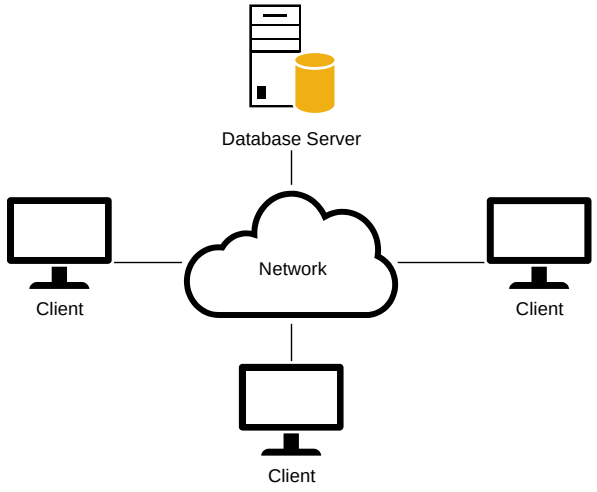
Hierarchical Database

Client-Server Architecture

- Server software
  - Database management system (DBMS)
  - “Back-end” processing

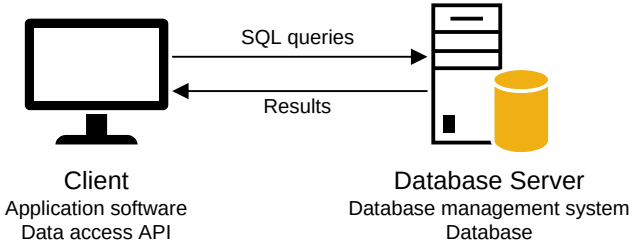


- Client software
  - Application software
  - “Front-end” processing



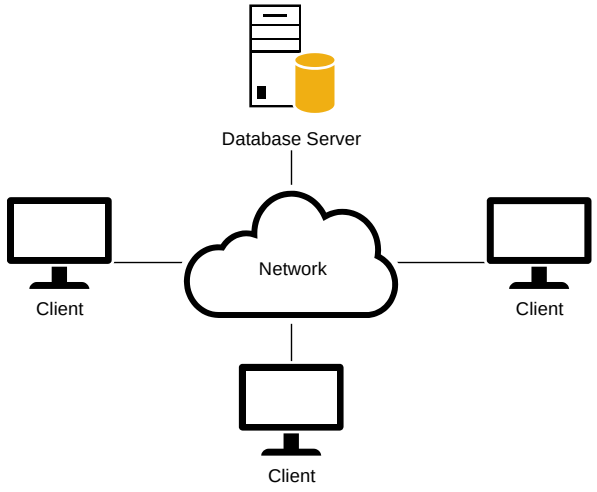
Client-Server Architecture

- Server software
  - Database management system (DBMS)
  - “Back-end” processing



MySQL Server

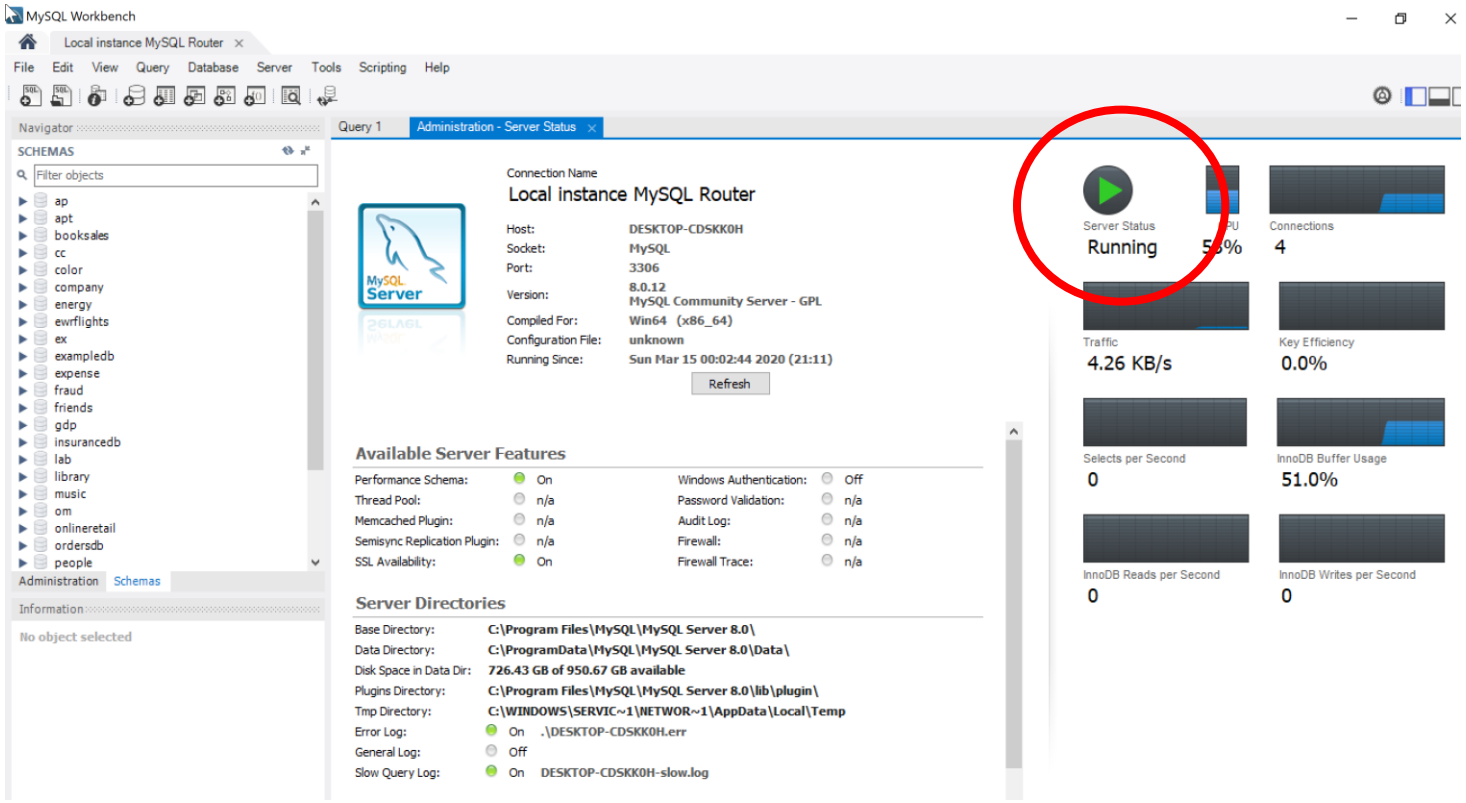
- Client software
  - Application software
  - “Front-end” processing



MySQL Workbench

Note on Accessing MySQL

If you have successfully installed both MySQL Workbench and MySQL Server, you should be able to go to Server > Server Status (in the tool menu) and see a screen like this, that shows the Server Status as “Running”



Scripts to Be Able to Run

- create_databases.sql (on Blackboard under Week 1 or module 1)

Database Terminology in MySQL

- Database
  - a collection of data
- Schema
  - the internal structure (layout) of a database, including tables and the relationships between tables
- Table
  - A matrix of rows and columns containing datapoints
- Column/Row/Cell
  - Components of tables
- Key
  - A column in a table designated as a means of identifying rows in the table, used especially for making connections between tables

Example: Database Schema

Database model: exampledb

student (number, name, address, email)
course (number, name)
building (number, address)

Example: Database

- Database: exampledb

STUDENT			
Number	**Name**	**Address**	**Email**
0165854	Bart Baesens	1040 Market Street, SF	Bart.Baesens@kuleuven.be
0168975	Seppe vanden Broucke	520, Fifth Avenue, NY	Seppe.vandenbroucke@kuleuven.be
0157895	Wilfried Lemahieu	644, Wacker Drive, Chicago	Wilfried.Lemahieu@kuleuven.be

COURSE	
Number	**Name**
D0I69A	Principles of Database Management
D0R04A	Basic Programming
D0T21A	Big Data & Analytics

BUILDING	
Number	**Address**
0600	Naamsestraat 69, Leuven
0365	Naamsestraat 78, Leuven
0589	Tiensestraat 115, Leuven

Example: Table

Primary key **Columns**

vendor_id	vendor_name	vendor_address1	vendor_address2	vendor_city	vendor_st
1	US Postal Service	Attn: Supt. Window Services	PO Box 7005	Madison	WI
2	National Information Data Ctr	PO Box 96621	NULL	Washington	DC
3	Register of Copyrights	Library Of Congress	NULL	Washington	DC
4	Jobtrak	1990 Westwood Blvd Ste 260	NULL	Los Angeles	CA
5	Newbrige Book Clubs	3000 Cindel Drive	NULL	Washington	NJ
6	California Chamber Of Commerce	3255 Ramos Cir	NULL	Sacramento	CA
7	Towne Advertiser's Mailing Svcs	Kevin Minder	3441 W Macarthur Blvd	Santa Ana	CA
8	BFI Industries	PO Box 9369	NULL	Fresno	CA
9	Pacific Gas & Electric	Box 52001	NULL	San Francisco	CA
10	Robbins Mobile Lock And Key	4669 N Fresno	NULL	Fresno	CA
11	Bill Marvin Electric Inc	4583 E Home	NULL	Fresno	CA
12	City Of Fresno	PO Box 2069	NULL	Fresno	CA

Rows

Example: Relationship between Vendors & Invoices tables in AP database

Primary key

vendor_id	vendor_name	vendor_address1	vendor_address2	vendor_city	vendor_s
114	Postmaster	Postage Due Technician	1900 E Street	Fresno	CA
115	Roadway Package System, Inc	Dept La 21095	NULL	Pasadena	CA
116	State of California	Employment Development ...	PO Box 826276	Sacramento	CA
117	Suburban Propane	2874 S Cherry Ave	NULL	Fresno	CA
118	Unocal	P.O. Box 860070	NULL	Pasadena	CA
119	Yesmed, Inc	PO Box 2061	NULL	Fresno	CA
120	Dataforms/West	1617 W. Shaw Avenue	Suite F	Fresno	CA
121	Zylka Design	3467 W Shaw Ave #103	NULL	Fresno	CA
122	United Parcel Service	P.O. Box 505820	NULL	Reno	NV
123	Federal Express Corporation	P.O. Box 1140	Dept A	Memphis	TN

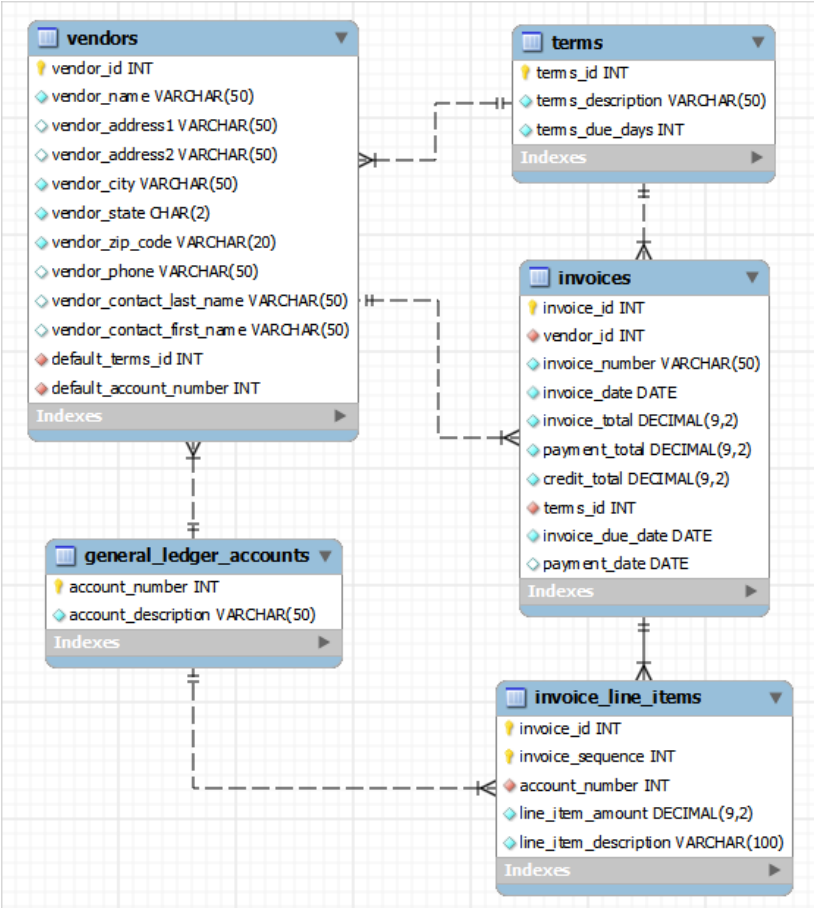
invoice_id	vendor_id	invoice_number	invoice_date	invoice_total	payment_total	credit_total
55	123	963253245	2014-06-10	40.75	40.75	0.00
56	86	367447	2014-06-11	2433.00	2433.00	0.00
57	103	75C-90227	2014-06-11	1367.50	1367.50	0.00
58	123	963253256	2014-06-11	53.25	53.25	0.00
59	123	4-314-3057	2014-06-11	13.75	13.75	0.00
60	122	989319-497	2014-06-12	2312.20	2312.20	0.00
61	115	24946731	2014-06-15	25.67	25.67	0.00
62	123	963253269	2014-06-15	26.75	26.75	0.00
63	122	989319-427	2014-06-16	2115.81	2115.81	0.00
64	123	963253267	2014-06-17	23.50	23.50	0.00

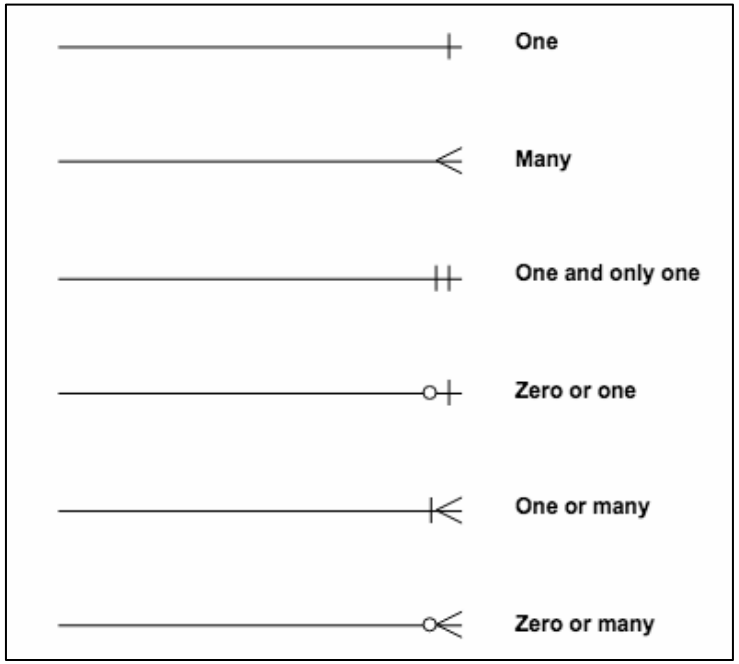
Foreign key

- Primary key
  - Column(s) in a table that uniquely identify each row in that table
- Composite key
  - Keys that consist of two or more columns
- Foreign key
  - Column(s) in a table that refer to a primary key in another table

(Enhanced) Entity-Relationship Diagram

An EER diagram for the AP database





SQL = Structured Query Language

- SQL is both a standard for DDL (Data Definition Language) and DML (Data Manipulation Language)

- Used to write queries, i.e. describe which parts of a database to retrieve

- Basic SQL statement are the same for all “dialects” of SQL.
  - Knowing one version of SQL allows you to easily learn others
  - Some syntax modifications may need to be made if moving to another databse

The Anatomy of a SQL Statement

SELECT statement

```
-- select invices with balances outstanding
SELECT vendor_name, invoice_number, invoice_date,
    line_item_amount, account_description
FROM vendors v
    JOIN invoices i
        ON v.vendor_id = i.vendor_id
    JOIN invoice_line_items li
        ON i.invoice_id = li.invoice_id
    JOIN general_ledger_accounts gl
        ON li.account_number = gl.account_number
WHERE invoice_total - payment_total - credit_total > 0
ORDER BY vendor_name, line_item_amount DESC;
```

KEYWORDS

Identifiers

Semi-colon

Comment

SQL Best Practices

- Use ALL CAPS for keywords
- Use lowercase for other code, including table names and variables
- Use underscores to separate words in table/column names
- Start each clause on a new line
- Break long clauses into multiple lines, using indents
- Comment code for readability
- SQL note: white space (including line breaks, extra spaces and indents) does not affect the operation of the SQL code

SQL Readability

A SELECT statement that's difficult to read

```
select invoice_number, invoice_date, invoice_total,
payment_total, credit_total, invoice_total - payment_total -
credit_total as balance_due from invoices where
invoice_total - payment_total - credit_total > 0 order by
invoice_date
```

A SELECT statement that's coded with a readable style

```
SELECT invoice_number, invoice_date, invoice_total,
    payment_total, credit_total,
    invoice_total - payment_total - credit_total
    AS balance_due
FROM invoices
WHERE invoice_total - payment_total - credit_total > 0
ORDER BY invoice_date
```

White space (including line breaks, extra spaces and indents) does not affect the operation of the SQL code, but it does make the code much more readable

SQL Comment Syntax

A SELECT statement with a block comment

```
/*
Author: Joel Murach
Date: 8/22/2014
*/
SELECT invoice_number, invoice_date, invoice_total,
    invoice_total - payment_total - credit_total
    AS balance_due
FROM invoices
```

A SELECT statement with a single-line comment

```
-- The fourth column calculates the balance due
SELECT invoice_number, invoice_date, invoice_total,
    invoice_total - payment_total - credit_total
    AS balance_due
FROM invoices
```

Comments are lines within the code that are not executed by the system.

Why Data Management for Analytics?

- Any analysis project starts with the question: what do I want to know and what kind of data do I need/what form does the data have to be in?
- Very often, the limitations of what analysis is possible stem from the underlying design of the database.
- Within an organization, Business Analysts are often liaisons between IT and Management. Even when not doing the actual programming/database administration, it is very helpful to “speak the language” and understand the structure and possibilities.