Enterprise Data Modeling

1405 Instructor: Ruiqing (Sam) Cao

DATA…!  What is data ?

Data is a collection of values that convey **factual information**. They contain both useful and irrelevant information, therefore **require further processing** to extract meaningful insights.

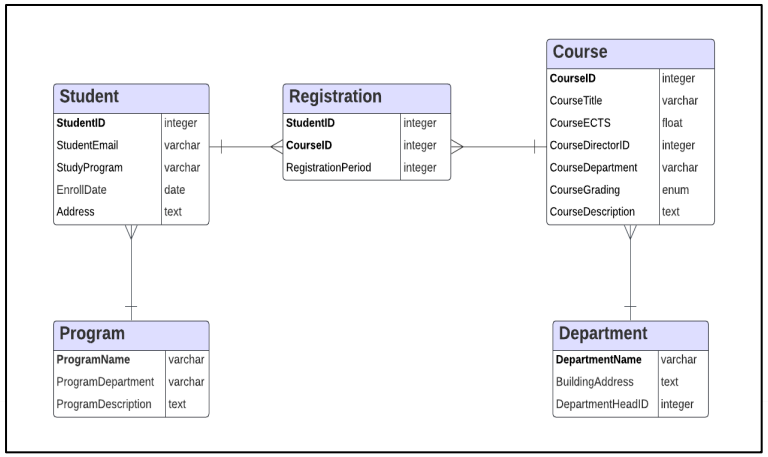
Data can be…

- *produced* from a variety of sources
- *processed* and transmitted in digital form
- *used* as a basis for reasoning, discussion, and computation

Structured, Semi-Structured & Unstructured

Structured Data

	A	B	C	D
1	id	first_name	is_group	date_joined
2	10000126370119684 30	Jimmy	FALSE	2012-10-11T18:06:36
3	10000132326031366 94	John	FALSE	2012-10-11T18:07:47
4	10000188613591043 48	Emilie	FALSE	2012-10-11T18:18:58
5	10000252367011848 12	Kelly	FALSE	2012-10-11T18:31:38
6	10000340615168000 68	Sarah	FALSE	2012-10-11T18:49:10
7	10000348668231680 31	Kyle	FALSE	2012-10-11T18:50:46
8	10000482130984966 61	Samuel	FALSE	2012-10-11T19:17:17
9	10000554608558084 70	Bryan	FALSE	2012-10-11T19:31:41
10	10000577845002245 37	Kyle	FALSE	2012-10-11T19:36:18
11	10000689832919044 92	Madison	FALSE	2012-10-11T19:58:33
12	10000719360819208 58	Brittany	FALSE	2012-10-11T20:04:25
13	10000880925409283 06	Alana	FALSE	2012-10-11T20:36:31
14	10000955919564807 83	Diane	FALSE	2012-10-11T20:51:25
15	10001002979655681 17	Nicole	FALSE	2012-10-11T21:00:46
16	10001199105310724 82	Tim	FALSE	2012-10-11T21:39:44
17	10001409827143682 04	Tien	FALSE	2012-10-11T22:21:36
18	10001444723752968 96	Kayla	FALSE	2012-10-11T22:28:32
19	10001552852910082 02	Brendan	FALSE	2012-10-11T22:50:01
20	10001657291079687 01	Zachary	FALSE	2012-10-11T23:10:46
21	10001889823293450 11	Maxwell	FALSE	2012-10-11T23:56:58



Semi-Structured Data

From: Elon Musk
To: Sam Altman
Subject: Re: question
Date: Monday, May 25, 2015 11:09:22 PM

Probably worth a conversation

```
<html xmlns="http://www.w3.org/1999/xhtml" ng-app="CourseDescriptionApp" ng-
controller="CourseDescriptionCtrl" ng-cloak>
<head>
    <title>
        Stockholm School of Economics
    </title>

    <meta name="viewport" content="initial-scale=1.0, user-scalable=0, minimum-scale=1.0, maximum-
scale=1.0, width=device-width" />

    <link rel="icon" href="/css/img/favicon.ico" type="image/vnd.microsoft.icon" />
    <link rel="apple-touch-icon-precomposed" href="/css/img/apple-touch-icon.png" />
    <link rel="stylesheet" type="text/css" href="/css/bundle.css" />
    <link rel="stylesheet" href="https://use.fontawesome.com/e0806b813c.css">
    <script src="/js/angular.min.js?rev=1" type="text/javascript"></script>
    <script src="/js/Show.js?rev=1" type="text/javascript"></script>


    <meta charset="utf-8" />

    <meta name="description" content="Description of course/module {{courseDescription.CourseNr}} at
the Stockholm School of Economics" />
    <meta property="og:title" content="{{courseDescription.PublicSubHeader}}" />
    <meta property="og:description" content="Description of course/module
{{courseDescription.CourseNr}} at the Stockholm School of Economics" />
</head>
```

Unstructured Data

Structured, Semi-Structured & Unstructured

	Definition	Common Examples
Structured Data	Structured within a well-defined **data model**, and adheres to a pre-defined standardized format	• Tabular data (e.g., CSV files) • Relational (SQL) databases
Unstructured Data	Not structured within any specific data model, and does not adhere to a pre-defined pattern	• Text • Image • Audio, Video
Semi-Structured Data	Has some level of organization but does not conform to the strict schema of structured data	• Email (sender/subject vs. main body) • HTML, JSON

Big Data Management in Organizations

- Penetration of digital technologies in organizations → rapid proliferation of digital data → evolution and innovation of data management systems

- Trend shifts from Hadoop eco-system technologies (~2005) to cloud computing (late 2010s) for managing storage of processing needs of larger data volume, variety, and velocity (***3V's***) in many organizations
- ***Volume***: 2.5 billion GBs (quintillion bytes) of data are generated per day
- ***Variety***: increasingly heterogeneous sources (esp. unstructured data)
- ***Velocity***: continuously generated real-time data gave rise to innovations in stream processing frameworks

Big Data Management in Organizations

- Apart from the IT industry, many **traditional industries** have also increasingly developed diverse data capabilities
  - **Financial services**: large banks, insurance companies
  - **Retail**: especially e-commerce and multi-channel
  - **Healthcare**: emphasis on data protection, cybersecurity, and compliance

- Big data management systems are common in **large organizations**
  - But designs and implementations vary widely, involving ***trade-offs*** based on e.g., business needs, legacy system compatibility, regulatory requirements, and organizational culture, etc.

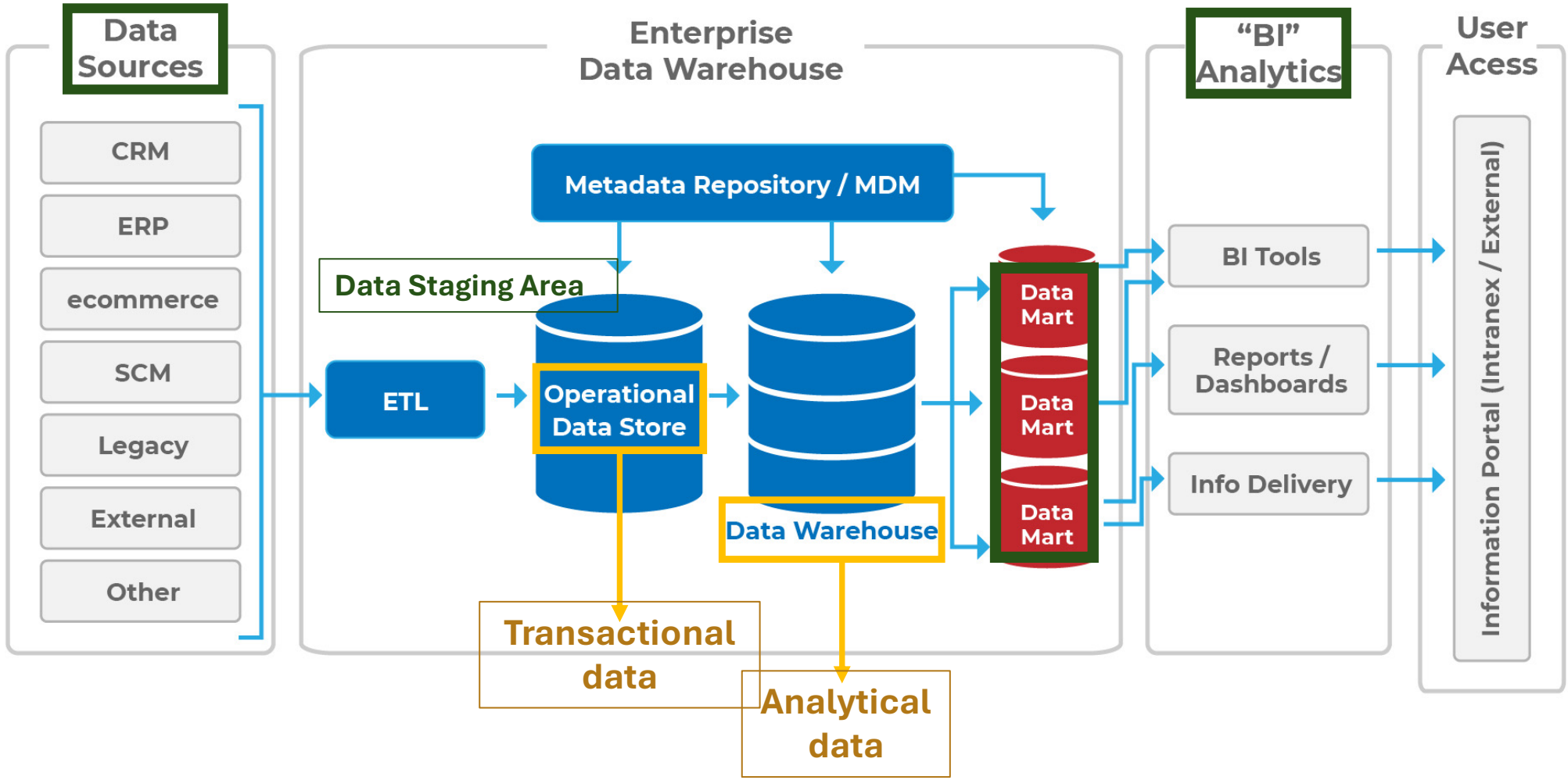
Big Data Management in Organizations

Big data management handles large, complex data beyond the capacity of traditional database management systems (DBMS), supported by a vast ecosystem of related technologies

This class does not focus on big data, but we cover foundational principles applicable to both traditional DBMS and big data systems

- **Data modeling** for designing traditional DBMS
- **Transactional vs. analytical** data processing
- **Structured**, **semi-structured**, and **unstructured** data

(Simplified) Enterprise Data Architecture



Enterprise Data Architecture (Glossary)

- **Data sources**: Raw data generated from a variety of sources, e.g., ERP (enterprise resource planning) systems, sensor networks, customer transactions, point-of-sales, clickstream data, user generated content
- **Data staging areas**: Temporary data storage before loading into a data warehouse
- **Data warehouse**: Centralized repository that loads structured historical data which are integrated from multiple upstream sources
- **Data marts**: Subset of data warehouse that provide domain-specific data for business users to perform analysis
- **Business intelligence (BI)**: Tools and processes for analyzing and visualizing data to deliver insights for strategic decision-making

Transactional vs. Analytical Processing

- **Transactional data** are generated from business operations (e.g., point-of-sales, financial transactions) at *relatively high frequency*, and require *fast processing by simple queries*, high *accuracy and data integrity*, but do not require long retention of historical data

- **Analytical data** provide strategic view of the organization to drive *decision-making*, and involve relatively *large amounts of historical data aggregated from multiple upstream sources* with complex queries, but are less demanding on data integrity and low latency

Data Model

- A **data model** serves as a *blueprint for database design*

- A **data model** provides the framework for defining how data is *organized and stored* in a database, and visually represents the *structure* of the database, including ***data elements***, ***attributes***, ***relationships***, and ***constraints***

- Two distinct modeling approaches for **transactional data (Entity-Relationship Model)** vs. **analytical data (Dimensional Model)**

Data Model: ER vs. Dimensional

The two distinct *data processing paradigms* for organizational data are associated with different *data modeling approaches*

- **Transactional data** by OLTP (Online Transactional Processing)

➔ **Entity-Relationship (ER) data modeling**

- **Analytical data** by OLAP (Online Analytical Processing)

➔ **Dimensional data modeling**

ER Model for Transactional Data

- **Entities**: business objects with data stored about them
  - e.g., a person, a transaction, a product, a department, etc.
- **Relationship**: describes how two entities are associated
  - e.g., a user initiates a transaction, an employee works in a department
- **Attributes**: properties of an entity
  - E.g., a person's email, a transaction's completion date

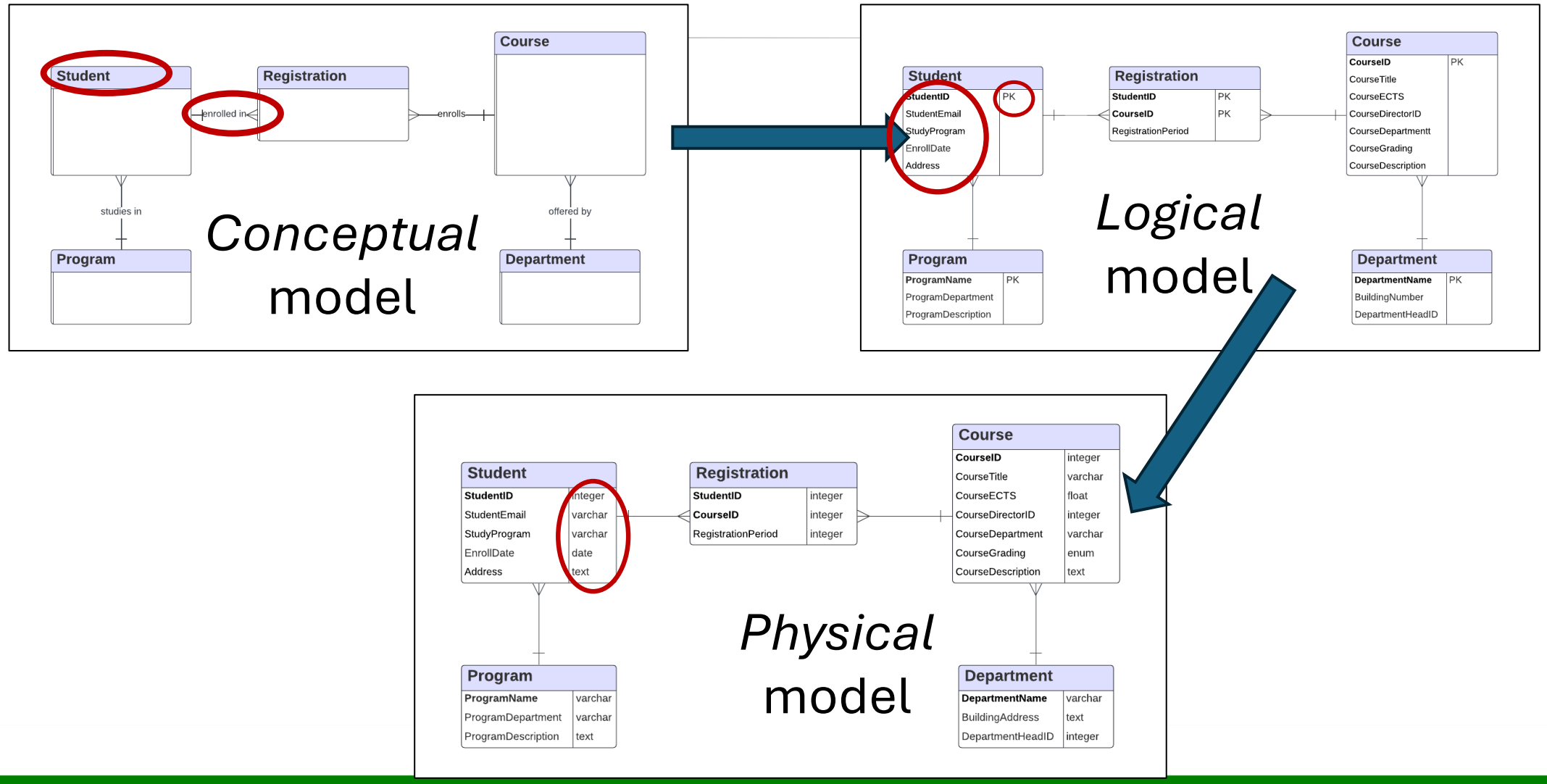
ER Model for Transactional Data

- **Conceptual data model**: High-level representation of organizational data, focusing on ***entities*** and ***relationships*** (including ***cardinalities***) between entities

- **Logical data model**: more detailed blueprint based on the conceptual model, including the ***attributes*** of each entity, which attributes in each entity serve as the ***primary keys***, and ***normalization***.

- **Physical data model**: ***implementation details*** of the logical model in the actual database system, including specific data type of each attribute, and model constraints, storage structures, etc.

ER Model for Transactional Data



ER Model for Transactional Data

In the actual database implementation of the data model:

Entities → Become **Tables** (also called *relations*)

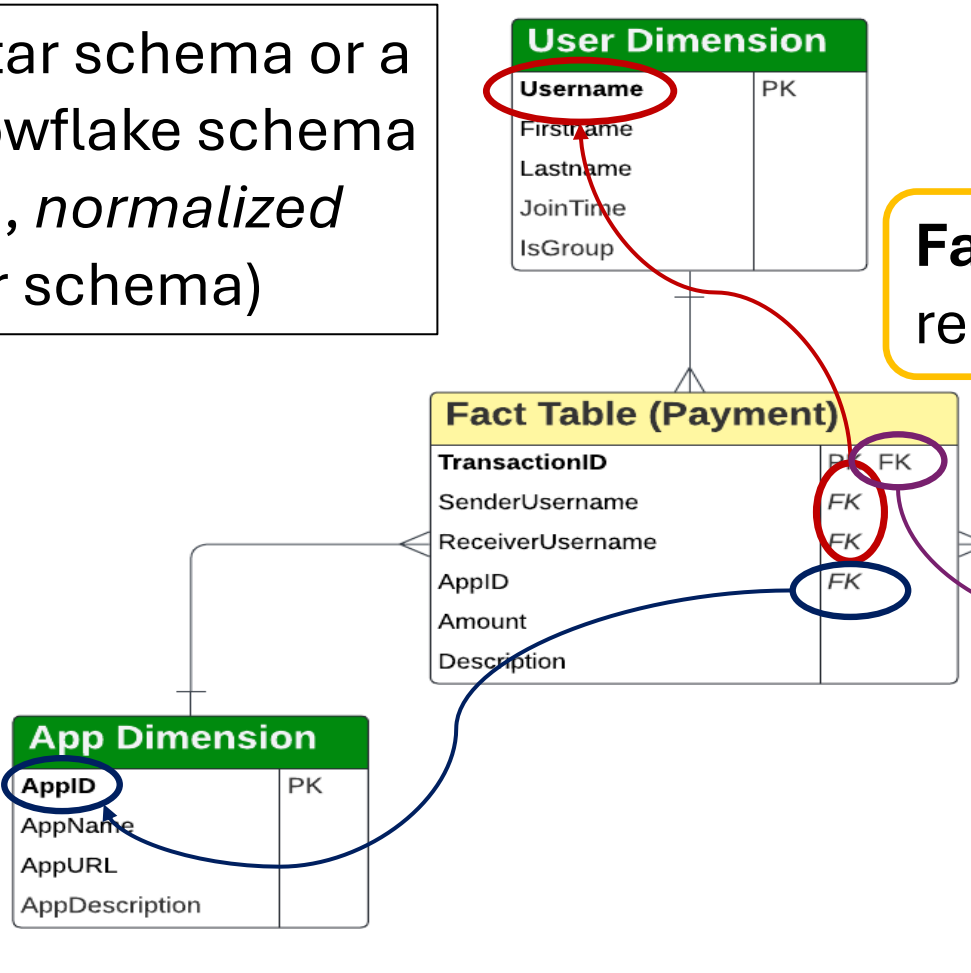
→ Entity instances become **Rows** in a table

Relationship → Is implemented by **foreign key** and **primary key**

Attributes → Become **Columns** in a table

Dimensional Model for Analytical Data

A star schema or a snowflake schema (i.e., *normalized* star schema)



Fact Table: accumulates quantitative metrics resulting from a business event or process

Dimension Tables: provide the reference attributes and relational context for events in the fact table

Exercise: Dimensional Data Model

You are a data consultant working for **Citibank**, a leading global bank, to design a database to analyze customer transactions. Your client needs a system that can answer critical business questions such as:

- What is the total transaction volume each month by customer segment (e.g., regular vs. high-net-worth) and region?
- What are the daily, monthly, and quarterly transaction trends?
- Which types of transactions (e.g., credit card purchases, ATM withdrawals) contribute the most to revenue?

Your task is to design a **dimensional data model** to support these analyses.

Exercise: Dimensional Data Model

Follow these steps to complete the exercise:

1. Identify the fact: Determine the central focus of the data model, and measurable metrics for this fact
2. Identify the dimensions: Surrounding contexts for analyzing the fact (e.g., time, customer, transaction type, region, etc) and attributes
3. Design the Fact Table: Define the structure of the fact table, including the foreign keys that connect to the dimension tables, and measurable metrics
4. Design the Dimension Tables: Specify the attributes for each dimension table that provide contextual details for the fact table
5. Draw the star schema design: A visual representation of your data model that shows the fact table in the center, and how it connects to surrounding dimension tables via foreign keys

ER Modeling vs. Dimensional Modeling

Dimensional modeling	ER modeling
• Analytical data • Star schema and snowflake schema • Organized around facts (business events and processes) • Optimizes database structure for fast analytical queries and business intelligence • Often denormalized to speed up data retrieval	• Transactional data • Conceptual, logical, & physical models • Organized around entities and relationships • Optimizes database structure for data integrity and reliability through ACID properties • Often normalized to reduce redundancy and improve data integrity

Recap: Transactional vs. Analytical data

Paradigm	Transactional data	Analytical data
Purpose	Serve operational business needs	Provide strategic view of the business for reporting and decision-making
Source	Real-time transaction-level business data as they are generated	Multi-dimensional data aggregated from upstream OLTP systems
Storage	Operational data store, smaller covering current non-historical data	Data warehouse, larger to accommodate diverse sources and historical data
Examples	Point-of-sales, financial transactions	Aggregated data for business intelligence
Capabilities	Fast processing of simple queries, high accuracy and integrity (ACID)	Handling complex queries and aggregation of various data sources
Data Model	Entity-Relationship (ER) model	Dimensional model