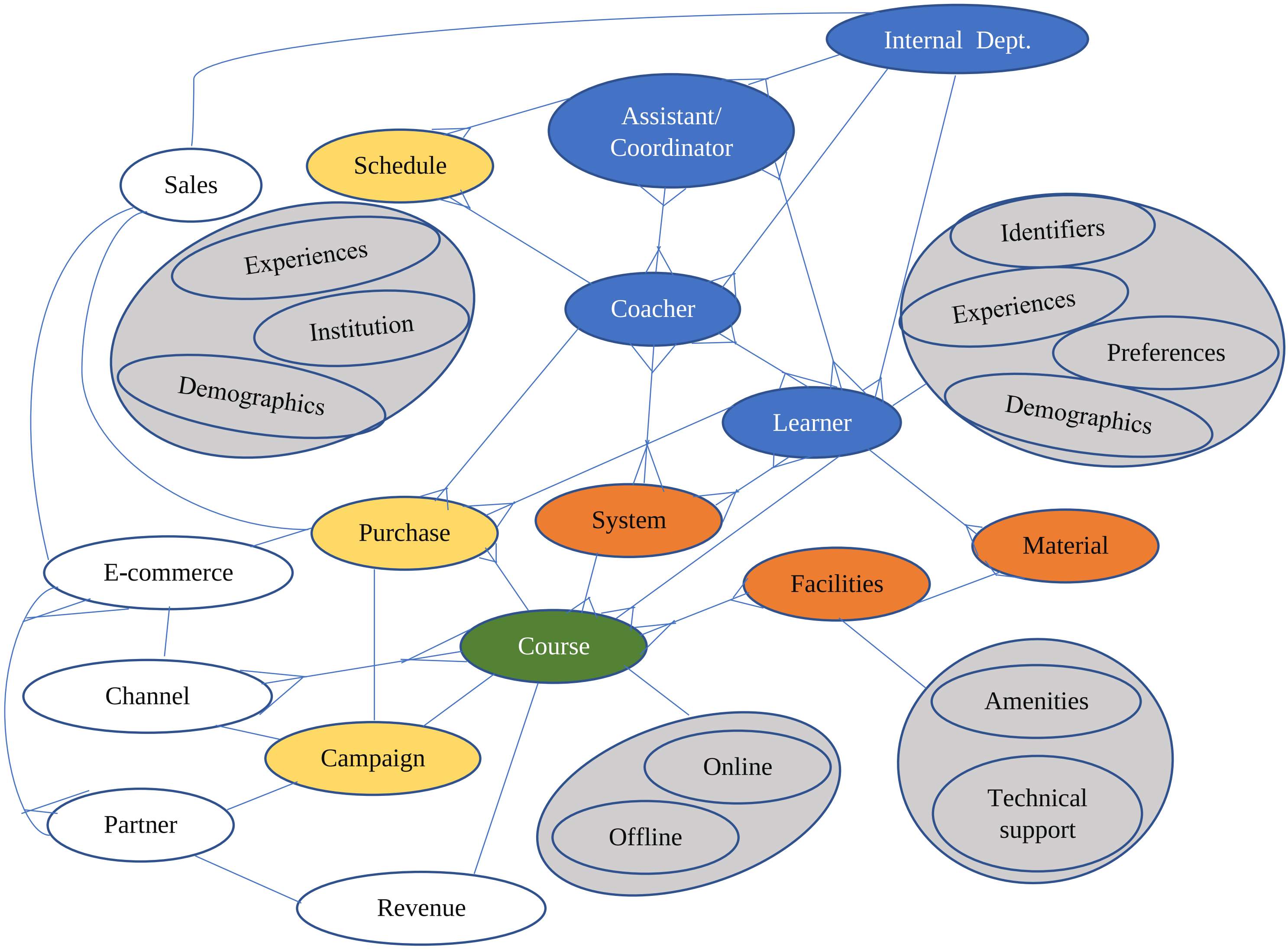
FINAL PROJECT – COURSE ASSIGNMENT

Introduction to Data Analytics for Business

Author: Khanh Ho

1. Part 1: Conceptual business model

Coaching course (outsourcing)



Scenario:

Company sales the coaching courses, which outsources mainly the courses and the coachers from different institutes or with different backgrounds. There are multiple courses advertised on both E-commerce and company site. The learner can choose the coaching courses as his preference to learn, both online and offline. Each course can be taught by one or several coachers, with an assigned schedule. A coacher is responsible for some courses. The coordinator/assistant (company employee) is the direct speaking partner of both coachers and learners. He/she will solve all related problems in the course. A course will have at least one coordinator, and a coordinator can organise some courses. Partner can be the institution, the hotel and restaurant (if in case the proceeded courses are full time and/or in various locations), teaching location (could also outside the company), etc

The facilities and material depends on the course. Some courses can use the same facilities, but some request special equipment (like laboratory, protecting clothes,…). Some of the basic course have the same material/document) but some need the assessment request.

The purchase is proceeded through online or at the office or from salesman. Internal Department is responsible for the billing process of the learners and salary of the coachers. There are some campaign/promotion to increase the sales (contributing to revenue)

All related data about the course is stored in company system.

2. Part 2: Relational data model

Customer.csv

Customer_ID	**Name**	**DOB**	**Location**	**Course**	**Date**	**Experience**	**Preference**
#integer	Full name	Date of birth dd.mm.yyyy	Geograp. location	Course ID	Purchased date dd.mm.yyyy	Customer background	Favorite topics
12345	*Jason Klar*	*03.05.1988*	*Germany*	*CO12345*	*18.03.2022*	*Python beginner*	*Data Science*

Coacher.csv

Coacher_ID	**Name**	**DOB**	**Institute**	**Experience**	**Offer**
#integer	Full name	Date of birth	Current work	background	Course_ID
00001	*Dr. Marian Brigit*	*12.03.1968*	*University of Hamburg*	*Professor in Data Science 10 years*	*GE00005, GE00008*

Partner.csv

Partnert_ID	**Name**	**Location**	**Specialization/Equipment**	**Others**
#integer	Full partner name	location	Offers	Notes
100005	*Intelligence Institute*	*USA*	*Data Science course*	*Support also installations*

Course.csv

Course_ID	**Name**	**Level**	**Location**	**Prerequisite**	**Time**
#course code (2 first digits: general GE or specialization SP)	Full name	#level: 2 digits level beginer-BE, intermediate -IN, advanced-AD)	Location	Some of enrollment requirements	Running period
GE00005	*Data Science in Python*	*BE*	*Germany*	*No*	*01.04.2022 – 31.07.2022*

Purchase.csv

Number	**Date**	**Course**	**Type**	**Customer**	**Base**	**Purchase**	**Campaign**
Based on purchase, course	Purchased date	Course_ID	Private00/ Business01	Customer_ID	Original price in euro	Price in euro	New Customer
1	*18.03.2022*	GE00005	00	*CO12345*	120.00	114.00	5%

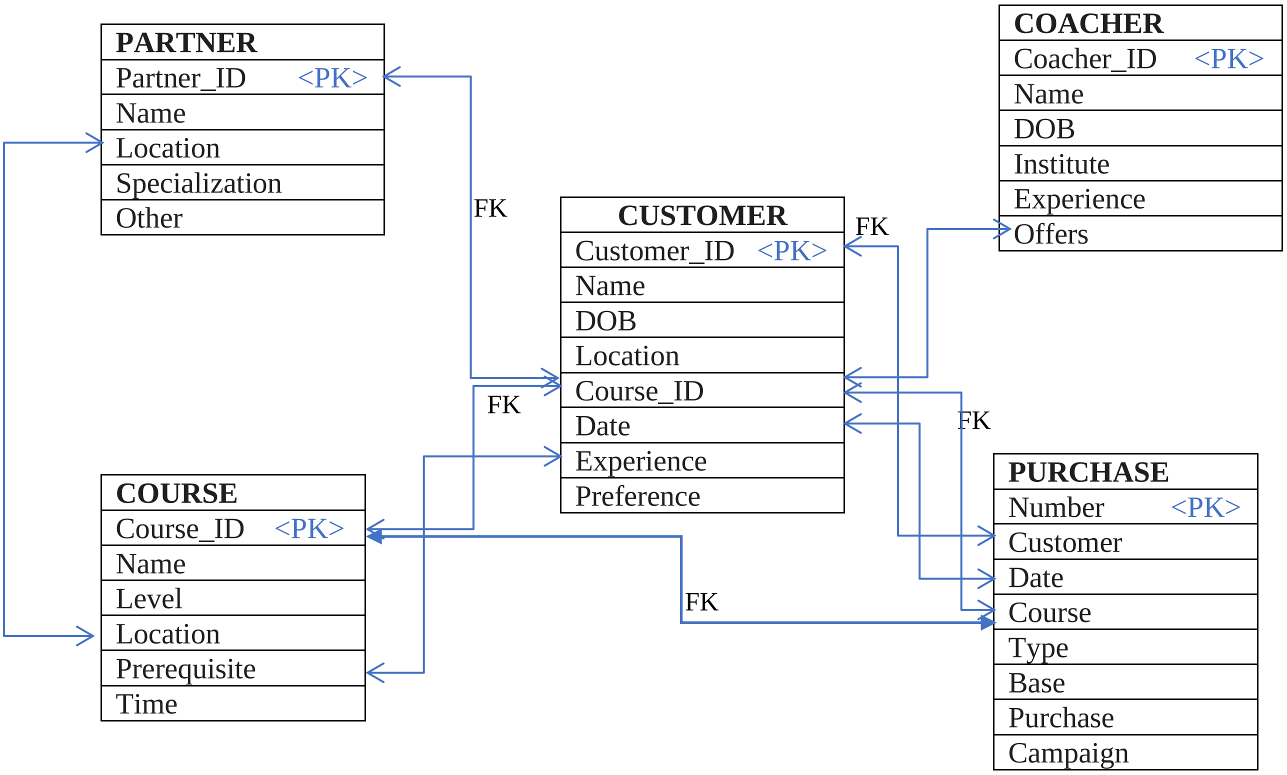


Table	Primary key	Type	Type of system
Customer	Customer_ID	Natural	Customer & People System: CRM, campaign management
Partner	Partner_ID	Natural	External Source System: Partners & Suppliers
Coacher	Coacher_ID	Natural	External Source System: Partners & Suppliers
Course	Course_ID	Composite (according to time and course)	Product & Presence System: Product Management, Web Management & Analytics
Purchase	Number	Surrogate	Core Enterprise: Billing & Invoicing/ ERP

3. Part 3: SQL queries

Querry 1: How many courses are purchased monthly?

- Data: Table Course, and Purchase
- Syntax

```
SELECT Course_ID, COUNT(*) AS Count,
            SUM(Purchase) AS Total_Sales,
            MONTH(Date) AS Month
FROM Purchase
GROUP BY(Month)
ORDER BY Month, Course_ID
```

Querry 2: List of customers in courses

- Data: Table Customer, Coacher, Course, Purchase
- Syntax

```
SELECT A.Customer_ID, A.Course_ID, A.Name, A.Count(Customer_ID) AS A.Amount,
          B.Name,
          C.Location, C.Time
FROM Customer A
LEFT JOIN Coacher B
ON A.Course_ID = B.Offers
LEFT JOIN Course C
ON A.Course_ID = C.Course_ID
GROUP BY A.Course_ID
SORT BY C.Time, A.Course_ID
```

4. Part 4: Sensitive data and data quality issues

a. Fields relate to

- PII - Personal Identifiable Information:
    - Identification: Customer_ID, Coacher_ID, Name, Date of Birth (DOB), Location, Contact number
    - ID card, social security number, driver's license, credit/bank accounts
    - ⇨ Cyber liability insurance policies to protect personal information
    - ⇨ My model: Customer_ID, Coacher_ID, Name, Location, etc
- CFI – Consumer Financial Information:
    - Financial institute/Commercial Banking (authorized): Credit products (loans, cards, accounts used by a customer)
    - With online retailers: show info on previous inquiries (web searches, viewed products, purchases)
    - ⇨ Examine Customer data and protect data from accidental or unknown parties
- CPNI – Customer Proprietary Network Information
    - Call center/telecommunication services: time, phone number, location, duration, problem/issue, cost
    - ⇨ Customer permission of publishing info through sign up/ accept declaration
- PHI – Protect Health Information:
    - Medical healthcare provider, health plan (individual/business), identification number
    - ⇨ Treatment provisions/recommendations

b. What data elements in your model will present the most significant data quality challenges?

- Input data globally and source system
    - can contain errors from the inputs (both customers and employees) such as the name with special letters/signals from
    - the address with the unknown/mistyped postcode may lead to duplicates.
    - Big datapool compares to the old-fashion system/architecture takes time for data preparing and validation
- Data privacy
    - Hacker/malware can attack the company sites to steal the data. This results in data lost/financial issues