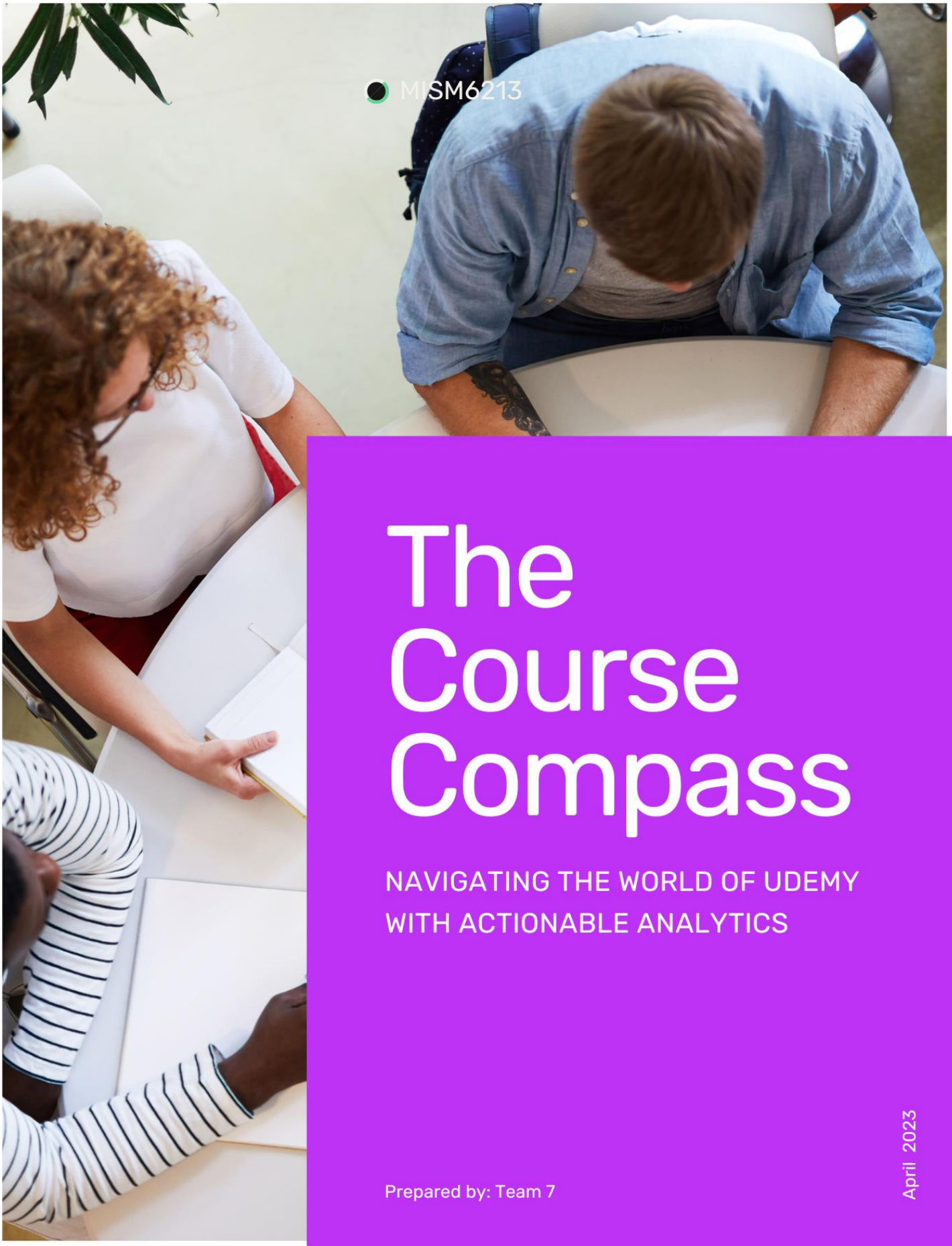


The Course Compass

NAVIGATING THE WORLD OF UDEMY WITH ACTIONABLE ANALYTICS

Prepared by: Team 7

April 2023

Table of Contents

PHASE 1: PROPOSAL

About the company

Udemy is a leading online education platform that offers over 213,000 online video courses with new additions published every month, taught by experts in their respective fields, to over 40 million students worldwide. Since its launch in 2010, Udemy has democratized education, making it accessible and affordable to everyone, regardless of their location or financial background. With a mission to improve lives through learning, Udemy provides an open marketplace where students can choose from a wide range of courses across diverse subjects, including business, technology, health and fitness, and the arts. Through innovative features like self-paced learning and lifetime access to course materials, Udemy has revolutionized online learning, empowering individuals to acquire new skills and advance their careers at their own pace. With a global community of instructors and students, Udemy is committed to transforming the way people learn and thrive in the digital age.

Strategic Information Product Design Plan

The process of designing a strategic information product for Udemy begins with understanding the needs and requirements of the relevant stakeholders, which in this case are Udemy's instructors and course development teams, as well as executives. One of the biggest challenges that Udemy faces is market saturation as the e-learning market is highly competitive, with many platforms offering similar courses to Udemy. Another significant challenge that Udemy faces is retaining learners and keeping them engaged throughout the course. Finally, Udemy must establish a strong brand identity and reputation to succeed in the e-learning market.

Udemy's executives are striving to improve their competitiveness and provide better courses to enhance the learners' experience. To achieve this, optimize its course offerings, and maximize its performance, Udemy has recognized the need for a comprehensive information product: **database management system** (storage type) to keep track of elements such as course titles, number of subscribers, ratings, number of reviews, discount price impact, number of lectures per course, etc. In particular, the company has identified the potential benefits of a DBMS which allows Udemy executives to easily access and manipulate data to identify trends, gaps, and opportunities in course offerings. Moreover, they can also use the DBMS to track the performance of individual courses over time and make data-driven decisions about content, marketing, and pricing strategies.

In terms of **Porter's Five Forces**, this DBMS addresses the Porter's Five Forces of **industry rivalry** and the **threat of new entrants**. By collecting, analyzing, and utilizing course performance data, Udemy gains a competitive advantage over its rivals in the industry such as Coursera, EdX, Khan Academy, etc. This information is used to identify trends, gaps, and opportunities in course offerings and create a more compelling offering for potential customers. Additionally, the DBMS helps Udemy to monitor changes in the competitive landscape and make informed decisions to stay ahead of the competition and retain their market position.

Once that the IP is identified, the first step is the conceptual data design, which includes identifying the entities, the relationships between entities, and the cardinalities of these relationships. This will be captured in an Entity Relationship Diagram (ERD) that serves as a blueprint for the final IP. Next, a logical data design is created that involves normalizing the data to the third normal form, creating a relational schema, and developing entity summary tables for the next phase of the design process. Then, the physical data design phase starts through the creation of tables and data generation. Finally, the IP is finalized with database query examples used to demonstrate the full capabilities of the system.

Quality Information Product Strategy

The information product chosen is a storage type of IP: **Course Performance Analytics Database Management System**. This IP will provide Udemy accurate, up-to-date information on the performance of courses and instructors on their platform. The database will be designed with flexibility and scalability in mind, allowing for a variety of analyses and insights such as course popularity, course quality, instructor engagement, etc.

The basic building blocks of the IP are the entities of Course, Instructor, Traffic, Category and Review, with their corresponding attributes such as course_id, instructor_id, _id, num_subscribers, rating, num_reviews, enrollment_time, completion_time, review_text, etc.

Furthermore, the **data collector** of this IP would be the Udemy platform itself, as it easily collects data on course performance metrics such as ratings, reviews, completion rates, and instructor engagement. The **data custodian**, on the other hand, would be the team responsible for managing and maintaining the DBMS system, ensuring that data is securely stored, backed up, and accessible to authorized users. This team may include database administrators, IT professionals, and data management experts at Udemy. Finally, the **data consumers** in this scenario would be the data analysts, executives, and even instructors and learners at Udemy who use the DBMS to extract, analyze, and visualize data on course performance metrics.

Correspondingly, the **end users** of Udemy's DBMS would primarily be data analysts and executives responsible for making strategic decisions based on course performance data. The DBMS would enable data analysts to perform queries, generate reports, and create data visualizations that identify patterns and trends in course performance. They could also use the DBMS to clean, transform, and merge data from different sources to create a unified view of Udemy's course performance data. In turn, executives at Udemy would be the primary users of the insights generated by the data analysts.

Challenges and Learning

Working with large datasets like the "Business Courses on Udemy" will present various challenges, including data cleaning, data normalization, and data integration. With over 10,000 courses from multiple business domains, one challenge will be ensuring the accuracy and completeness of the data. However, these challenges will provide valuable learning opportunities in data management, data analysis, and data visualization, helping us develop technical and analytical skills that can be applied to other datasets in the future. The insights gained from analyzing this dataset will also provide valuable strategic insights for businesses operating in the online education industry like Udemy. Enlisted below are some major issues:

- Data quality issues
- Technical issues
- Data privacy concerns
- User adoption
- Cost
- Limited scope
- Inflexibility

- Bias
- Data overload
- Competitive pressures

PHASE 2: DATA

Data Selection

The Course Performance Analytics Database has been chosen as the information product for Udemy's strategic plan because it will provide accurate and up-to-date information on the performance of courses on their platform. The current dataset doesn't capture all the required information efficiently, hence the DBMS we suggest would have the entities and attributes mentioned in Phase 3. This information will allow Udemy to make targeted improvements and optimizations to increase student engagement, course completion rates, and overall course quality, which is critical to improving their competitiveness and retaining their customers. The database system also relates to the intensity of competitive rivalry, as Udemy needs to stay competitive with other online learning platforms by providing courses and instruction that meet or exceed the standards of their competitors. Therefore, this IP is critical and important for Udemy's success in the online learning market.

Original Dataset: Business Courses – Udemy (10K courses)[1]

id	title	url	is_paid	num_subscribers	avg_rating	avg_rating_recent	rating	num_reviews	is_wishlisted	num_published_lectures	num_published_practice_tests	created	published_time	discount_price__amount	discount_price__currency	discount_price__price_string	price_detail__amount	price_detail__currency	price_detail__price_string
762616	The Complete SQL Bootcamp 2020: Go from Zero to Hero	/course/the-complete-sql-bootcamp/	TRUE	295509	4.66019	4.67874	4.6787	78006	FALSE	84	0	2016-02-14T22:57:48Z	2016-04-06T05:16:11Z	455	INR	,₹455	8640	INR	,₹8,640
937678	Tableau 2020 A-Z: Hands-On Tableau Training for Data Science	/course/tableau10/	TRUE	209070	4.58956	4.60015	4.6002	54581	FALSE	78	0	2016-08-22T12:10:18Z	2016-08-23T16:59:49Z	455	INR	,₹455	8640	INR	,₹8,640
1E+06	PMP Exam Prep Seminar - PMBOK Guide 6	/course/pmp-pmbok6-35-pdus/	TRUE	155282	4.59491	4.59326	4.5933	52653	FALSE	292	2	2017-09-26T16:32:48Z	2017-11-14T23:58:14Z	455	INR	,₹455	8640	INR	,₹8,640
648826	The Complete Financial Analyst Course 2020	/course/the-complete-financial-analyst-course/	TRUE	245860	4.54407	4.53772	4.5377	46447	FALSE	338	0	2015-10-23T13:34:35Z	2016-01-21T01:38:48Z	455	INR	,₹455	8640	INR	,₹8,640
637930	An Entire MBA in 1 Course:Award Winning Business School Prof	/course/an-entire-mba-in-1-courseaward-winning-business-school-prof/	TRUE	374836	4.4708	4.47173	4.4717	41630	FALSE	83	0	2015-10-12T06:39:46Z	2016-01-11T21:39:33Z	455	INR	,₹455	8640	INR	,₹8,640
1E+06	Microsoft Power BI - A Complete Introduction [2020 EDITION]	/course/powerbi-complete-introduction/	TRUE	124180	4.56228	4.57676	4.5768	38093	FALSE	275	0	2017-05-08T13:03:21Z	2017-05-15T18:48:54Z	455	INR	,₹455	8640	INR	,₹8,640
864146	Agile Crash Course: Agile Project Management; Agile Delivery	/course/agile-crash-course/	TRUE	96207	4.32383	4.29118	4.2912	30470	FALSE	23	0	2016-05-30T22:57:40Z	2016-06-23T17:49:26Z	455	INR	,₹455	8640	INR	,₹8,640
321410	Beginner to Pro in Excel: Financial Modeling and Valuation	/course/beginner-to-pro-in-excel-financial-modeling-and-valuation/	TRUE	127680	4.54034	4.53346	4.5335	28665	FALSE	275	0	2014-10-17T08:39:52Z	2014-11-25T23:00:40Z	455	INR	,₹455	8640	INR	,₹8,640
673654	Become a Product Manager | Learn the Skills & Get the Job	/course/become-a-product-manager-learn-the-skills-get-a-job/	TRUE	112572	4.50386	4.5008	4.5008	27408	FALSE	144	0	2015-11-18T19:35:12Z	2016-03-17T17:04:59Z	455	INR	,₹455	8640	INR	,₹8,640
2E+06	The Business Intelligence Analyst Course 2020	/course/the-business-intelligence-analyst-course-2018/	TRUE	115269	4.50067	4.49575	4.4958	23906	FALSE	413	0	2018-04-19T07:00:09Z	2018-04-25T18:40:55Z	455	INR	,₹455	8640	INR	,₹8,640

The dataset can be found in Kaggle and is a compilation of all business-related courses available on Udemy's website, consisting of 10 thousand courses. The courses are categorized under various domains such as Finance, Entrepreneurship, Communication, Management, Sales, Strategy, Operations, Project Management, Business Law, Data & Analytics, Home Business, Human Resources, and Industry. The data was collected by the author through web scrapping from Udemy's website, and all the data collected is available in the public domain.

[1] Jilkothari. (2021). Business Courses - Udemy (10K Courses). Retrieved from Kaggle website: https://www.kaggle.com/jilkothari/business-courses-udemy-10k-courses

Dataset Utilization

The analysis being done is related to the performance and engagement of these courses and individuals. The tables and their attributes indicate that various metrics related to courses (such as the number of subscribers, ratings, reviews, etc.), instructors (such as engagement rate, the number of courses, etc.), and students (such as enrollment and completion time, ratings, progress, etc.) are being tracked and analyzed.

The data sets could be used for a variety of purposes, including:

- **Course evaluation and improvement:** By analyzing course metrics such as ratings, reviews, and engagement rates, instructors and course developers can identify areas for improvement and make changes to the course content and delivery.
- **Instructor evaluation and improvement:** By analyzing instructor metrics such as engagement rate, ratings, and the number of courses, instructors, and administrators can identify areas for improvement and provide targeted training and support.
- **Student engagement and performance analysis**: By tracking metrics such as enrollment and completion time, progress, ratings, and reviews, instructors and administrators can identify students who may be struggling and provide additional support and resources.
- **Market analysis**: By analyzing course metrics such as the number of subscribers, discounts, and original prices, administrators can identify popular courses and instructors, as well as trends in the online education market.

Overall, these data sets have the potential to be used to evaluate and improve the quality and effectiveness of online courses, as well as to provide insights into the behavior and preferences of students and instructors in the online education space.

The wrangled dataset will drop the highlighted columns:

id	title	url	is_paid	num_subscribers	avg_rating	avg_rating_recent	rating	num_reviews	is_wishlisted	num_published_lectures	num_published_practice_tests	created	published_time	discount_price__amount	discount_price__currency	discount_price__price_string	price_detail__amount	price_detail__currency	price_detail__price_string
762616	The Complete SQL Bootcamp 2020: Go from Zero to Hero	/course/the-complete-sql-bootcamp/	TRUE	295509	4.66019	4.67874	4.6787	78006	FALSE	84	0	2016-02-14T22:57:48Z	2016-04-06T05:16:11Z	455	INR	,₹455	8640	INR	,₹8,640
937678	Tableau 2020 A-Z: Hands-On Tableau Training for Data Science	/course/tableau10/	TRUE	209070	4.58956	4.60015	4.6002	54581	FALSE	78	0	2016-08-22T12:10:18Z	2016-08-23T16:59:49Z	455	INR	,₹455	8640	INR	,₹8,640
1E+06	PMP Exam Prep Seminar - PMBOK Guide 6	/course/pmp-pmbok6-35-pdus/	TRUE	155282	4.59491	4.59326	4.5933	52653	FALSE	292	2	2017-09-26T16:32:48Z	2017-11-14T23:58:14Z	455	INR	,₹455	8640	INR	,₹8,640
648826	The Complete Financial Analyst Course 2020	/course/the-complete-financial-analyst-course/	TRUE	245860	4.54407	4.53772	4.5377	46447	FALSE	338	0	2015-10-23T13:34:35Z	2016-01-21T01:38:48Z	455	INR	,₹455	8640	INR	,₹8,640
637930	An Entire MBA in 1 Course:Award Winning Business School Prof	/course/an-entire-mba-in-1-courseaward-winning-business-school-prof/	TRUE	374836	4.4708	4.47173	4.4717	41630	FALSE	83	0	2015-10-12T06:39:46Z	2016-01-11T21:39:33Z	455	INR	,₹455	8640	INR	,₹8,640
1E+06	Microsoft Power BI - A Complete Introduction [2020 EDITION]	/course/powerbi-complete-introduction/	TRUE	124180	4.56228	4.57676	4.5768	38093	FALSE	275	0	2017-05-08T13:03:21Z	2017-05-15T18:48:54Z	455	INR	,₹455	8640	INR	,₹8,640
864146	Agile Crash Course: Agile Project Management; Agile Delivery	/course/agile-crash-course/	TRUE	96207	4.32383	4.29118	4.2912	30470	FALSE	23	0	2016-05-30T22:57:40Z	2016-06-23T17:49:26Z	455	INR	,₹455	8640	INR	,₹8,640
321410	Beginner to Pro in Excel: Financial Modeling and Valuation	/course/beginner-to-pro-in-excel-financial-modeling-and-valuation/	TRUE	127680	4.54034	4.53346	4.5335	28665	FALSE	275	0	2014-10-17T08:39:52Z	2014-11-25T23:00:40Z	455	INR	,₹455	8640	INR	,₹8,640
673654	Become a Product Manager | Learn the Skills & Get the Job	/course/become-a-product-manager-learn-the-skills-get-a-job/	TRUE	112572	4.50386	4.5008	4.5008	27408	FALSE	144	0	2015-11-18T19:35:12Z	2016-03-17T17:04:59Z	455	INR	,₹455	8640	INR	,₹8,640
2E+06	The Business Intelligence Analyst Course 2020	/course/the-business-intelligence-analyst-course-2018/	TRUE	115269	4.50067	4.49575	4.4958	23906	FALSE	413	0	2018-04-19T07:00:09Z	2018-04-25T18:40:55Z	455	INR	,₹455	8640	INR	,₹8,640

Dataset Quality Assessment

The dataset used for this project is sourced directly from Udemy, although currently it has been obtained from Kaggle. While the datasets are considered authentic and meet the criteria of believability, accessibility, and relevancy, they do not meet all quality metrics as there is a lack of consistency in some areas.
To improve the quality of the dataset, it is recommended to focus on capturing only the necessary and relevant data. Instead of capturing as much data as possible, it is important to prioritize and ensure that only the required data is recorded. This approach will add value to the dataset and make it more relevant for analysis.

Data Quality Improvement Plan and Processes

To improve the quality of the data sets, the first step would be to conduct a thorough data quality analysis. This analysis would involve identifying the sources of the data sets, the methods used to collect the data, and the accuracy and completeness of the data. The data quality analysis would also involve identifying any missing data, duplicates, or outliers that impact the accuracy and reliability of the IP. Once the data quality analysis is complete, data cleansing techniques would be applied to the data sets to remove any inconsistencies and errors. This would involve removing duplicates, filling in missing values, and correcting any data that is inaccurate or inconsistent. After data cleansing, data integration techniques would be used to combine the data sets from different sources into a single, unified data set. This would ensure that the data is consistent and can be used for analysis.
Finally, data validation and verification would be performed to ensure that the data is accurate, complete, and consistent. This would involve comparing the data sets to external sources and performing statistical analysis to ensure that the data is reliable and can be used to make informed decisions.
Overall, the quality of the data sets used to create the Instructor and Course Analytics Database is critical in achieving the desired outcomes of increasing student engagement, course completion rates, and overall course quality. Therefore, a thorough data quality analysis, data cleansing, data integration, and data validation and verification processes will be performed to ensure the highest quality of data is used in the creation of the IP.

PHASE 3: CONCEPTUAL DATA DESIGN

Entities

1. Instructor
2. Course
3. Category
4. Traffic
5. Review
6. User

Relationships Between Entities

- Instructor – Course
- Category – Course
- Traffic – Course
- Review –Course
- User – Review

Cardinality Of Relationships Among Entities

- Instructor (mandatory many) Course (mandatory one)
- Category (mandatory one) Course (optional many)
- Traffic (optional many) Course (mandatory one)
- Review (mandatory one) Course (optional many)
- User (mandatory one)  Review (optional many)

Attributes Of All Entities

CATEGORY
- Category_ID
- Category

INSTRUCTOR
- Instructor_ID
- Instructor

TRAFFIC
- Course_ID
- Date
- Number_of_users
- Duration_of_visits
- Traffic_source

REVIEW
- Review_ID
- Course_ID
- User_ID
- Review_rating
- Review_comment

USER
- User_ID
- Name
- Age
- Gender
- Country
- Education
- Occupation

COURSE
- Course_ID
- Instructor_ID
- Category_ID
- Is_Paid
- Num_of_subscribers
- Num_of _reviews
- Course_rating
- Is_wishlisted
- Num_of_lectures
- Num_of_tests
- Published_date
- Discount_amt
- Price_amt
- Summary
- Language
- Total_duration

Entities And Attributes

CATEGORY	
PK	Category_ID
	Category

INSTRUCTOR	
PK	Instructor_ID
	Instructor

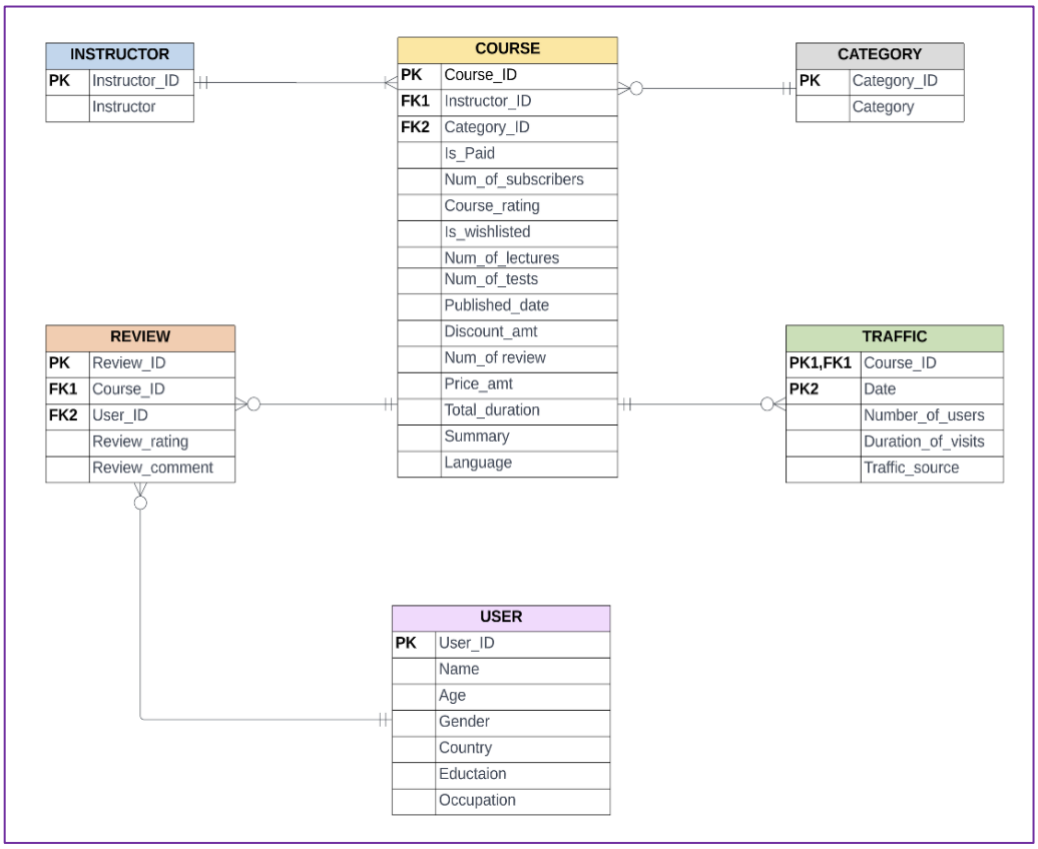
REVIEW	
PK	Review_ID
FK1	Course_ID
FK2	User_ID
	Review_rating
	Review_comment

COURSE	
PK	Course_ID
FK1	Instructor_ID
FK2	Category_ID
	Is_Paid
	Num_of_subscribers
	Course_rating
	Is_wishlisted
	Num_of_lectures
	Num_of_tests
	Published_date
	Discount_amt
	Num_of review
	Price_amt
	Total_duration
	Summary
	Language

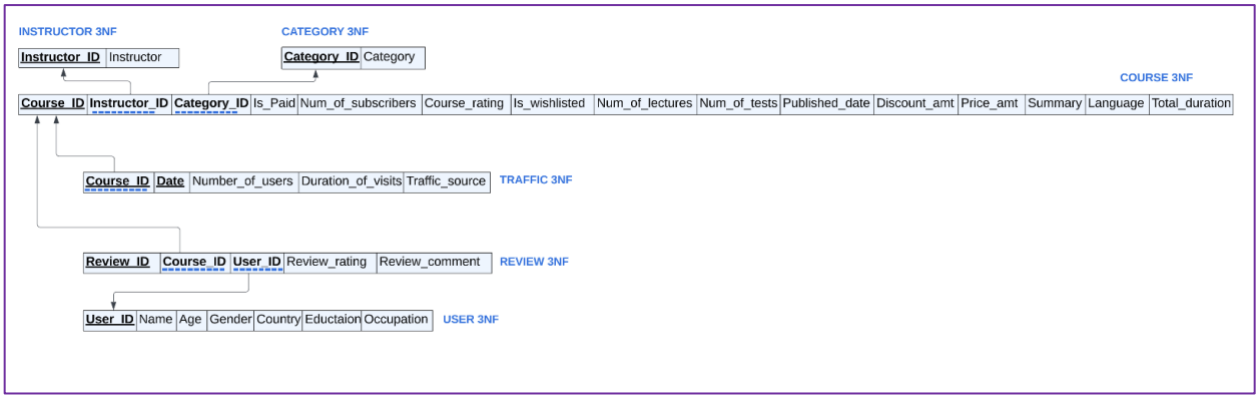
USER	
PK	User_ID
	Name
	Age
	Gender
	Country
	Education
	Occupation

TRAFFIC	
PK1,FK1	Course_ID
PK2	Date
	Number_of_users
	Duration_of_visits
	Traffic_source

Entity Relationship Diagram



Dependency Diagram



3NF Diagram

PHASE 4: FUTURE STEPS

From Data Collection to Dashboard Creation

The significant increase in the amount of data generated by online courses has made it essential for platforms like Udemy to efficiently organize and analyze this data. Creating a dashboard is a logical progression from the DBMS in order to achieve this goal. A dashboard can provide a visual representation of crucial metrics and trends, making it an effective method to do so. The following analyses can be displayed utilizing these dashboards:

- **Course engagement metrics** will display valuable information regarding which courses are popular among users and which ones require improvement. This will help Udemy create better courses that are engaging and can retain users for longer periods, ultimately resulting in increased revenue. To help visualize this data, several visualizations will be used. For instance, a pie chart or **donut chart** could represent the course completion rate, with the completed and incomplete sections of the chart showing the percentage of users who completed or did not complete the course.
- An **analysis of platform traffic** can provide valuable insights for optimizing the user experience and increasing engagement on Udemy. By examining metrics such as the number of daily, weekly, and monthly users, duration of visits, and traffic sources, the platform can identify which sources are driving the most traffic, which pages are most visited, and which features are most used. This

information can be used to make data-driven decisions on optimizing the platform, resulting in better user experiences and increased revenue for Udemy. To better visualize this data, several visualizations can be utilized. For instance, a **line chart** can show the trend in the number of users over time, with the x-axis displaying time intervals and the y-axis displaying the number of users. Additionally, a **histogram** can display the distribution of visit durations, with the x-axis showing time intervals and the y-axis displaying the number of visits.

- **Instructor performance analysis** will display crucial information about which instructors are popular among users and which instructors need improvement in terms of content, delivery, or engagement. This information can be analyzed by examining metrics such as course completion rate, course ratings, and revenue generated. The visualizations used for this analysis will include a line chart or bar chart to represent the completion rate for each course by an instructor. The x-axis will show the course ID, and the y-axis will show the completion rate percentage.

Notes:

The **USER table** provides executives with a comprehensive analysis of Udemy's user base, including demographics such as age, gender, location, education level, and occupation. This information can be used for targeted marketing and course development, ensuring that courses are tailored to specific demographics.

Additional charts *outside* the mentioned analyses will include the following visualizations:

- Heatmaps or geographic maps: Displaying the number of users or subscribers by country or region, to help stakeholders understand where their user base is concentrated.
- Stacked bar charts: Displaying the number of courses in each category broken down by whether they are paid or free, to help stakeholders understand how pricing affects course popularity.
- Word clouds or sentiment analysis: Analyzing course reviews to reveal common themes or sentiments expressed by users, providing stakeholders with valuable insights into the user experience.

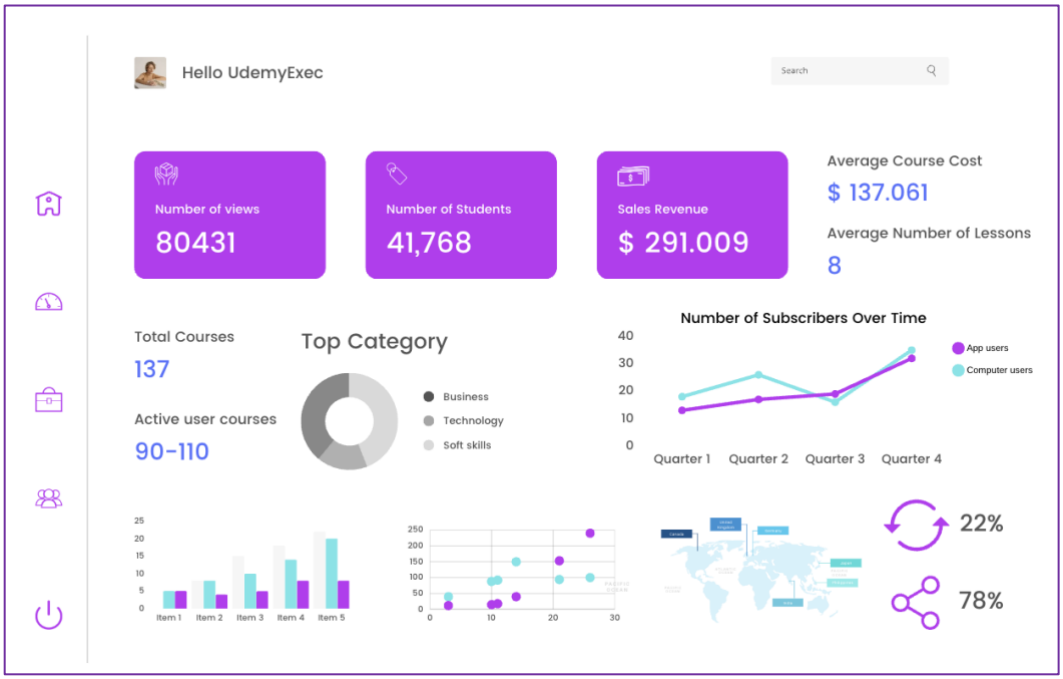


Figure 1 Prototype

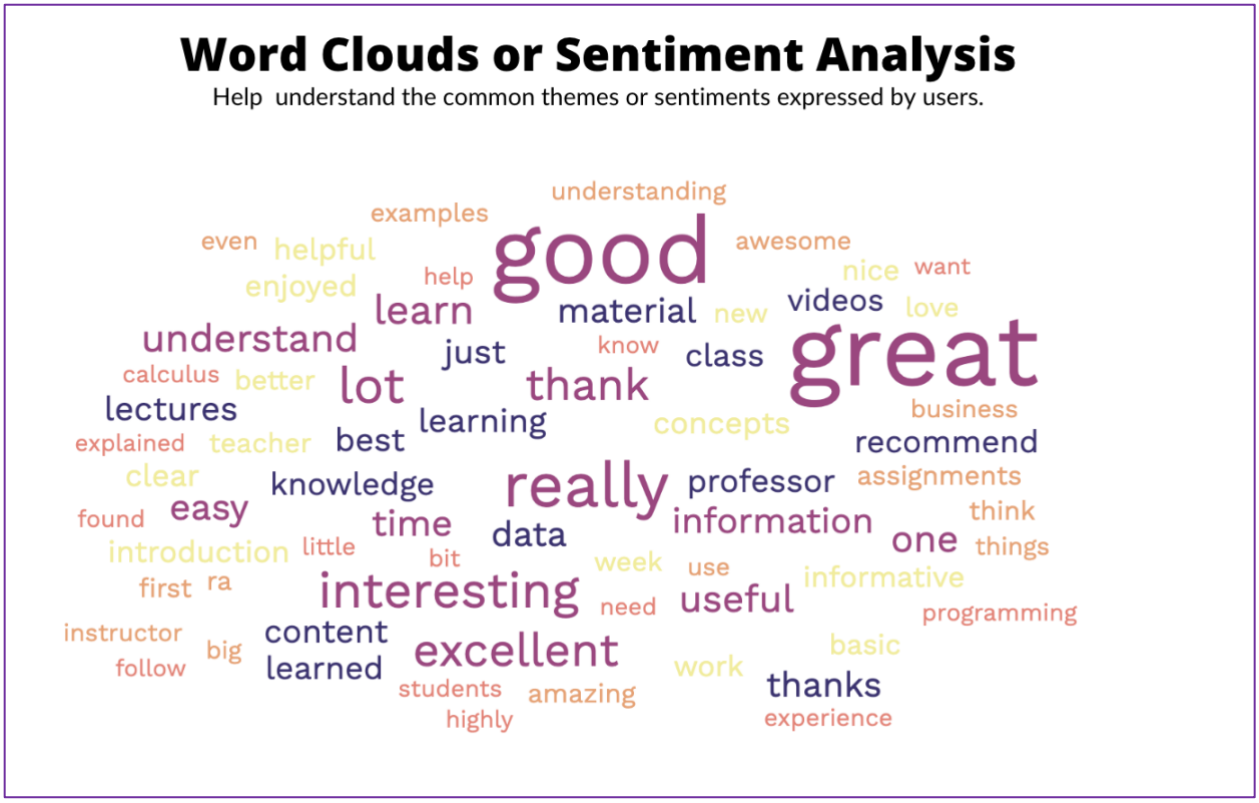


Figure 2 Word Cloud

Exhibits

Exhibit 1: Data Dictionary

Id	The course ID of that particular course.
Title	Shows the unique names of the courses available under the development category on Udemy.
url	Gives the URL of the course
is_paid	Returns a boolean value displaying true if the course is paid and false if otherwise.
num_subscribers	Shows the number of people who have subscribed that course.
avg_rating	Shows the average rating of the course.
avg rating recent	Reflects the recent changes in the average rating
num_reviews	Gives us an idea related to the number of ratings that a course has received.
num_ published_lectures	Shows the number of lectures the course offers.
num_ published_ practice_tests	Gives an idea of the number of practice tests that a course offers.
created	The time of creation of the course
published_time	Time of publishing the course.
discounted_ price_amount	The discounted price which a certain course is being offered at.
discounted_ price_currency	The currency corresponding to the discounted price which a certain course is being offered at.
price_ detail_amount	The original price of a particular course.
price_ detail_currency	The currency corresponding to the price detail amount for a course