A Data-centric Approach to Tracking Student Academic Performance and Progression

Jorge E. Ibarra-Esquer, *Member, IAENG,* Brenda L. Flores-Rios, Maria A. Astorga-Vargas, Maria L. González-Ramírez, Araceli C. Justo-López, and Gloria E. Chávez-Valenzuela

***Abstract*—Tracking student performance individually and as a group is a crucial activity for educational institutions. It serves as an indicator of success and provides valuable data for self-assessment and decision-making. Some important metrics and statistics can be used to understand student performance and progression. Examples include the number of students in a cohort or academic period, dropout and graduation rates, and failure and success rates for specific courses. These indicators can provide a comprehensive overview of how students are doing overall. However, the process of dealing with high volumes of data from different sources that are needed to calculate, present, explain, analyze, and visualize these indicators is not always a streamlined one. As part of a continuous improvement strategy in an engineering college at a public university in Mexico, a new role was defined to manage historical and current student data. Its primary objective was to establish a consistent and flexible system for collecting, organizing, processing, analyzing, and sharing student performance indicators from institutional data. The goal was to create an approach for tracking students' progression at college, program, and individual levels. This paper describes the data approach that guided this process, highlighting the dynamic reports, visualizations, and tools created to enhance data and improve decision-making.**

***Index Terms*—Data mining, Data and knowledge visualization, Student progression, Academic analytics, Academic performance indicators.**

I. Introduction

ACADEMIC analytics is a concept first proposed in [1] to describe the intersection of technology, information, organizational culture, and the application of data analytics to manage an institution [2]. As stated in [3], academic analytics provide overall information about what is happening in a specific program and how to address performance challenges, reflecting the role of data analysis at an institutional level. This is accomplished by combining large datasets with statistical techniques and predictive modeling to improve decision making, providing data that administrators can use to support the strategic decision-making process as well as a method for benchmarking in comparison with other institutions [3]. The type of academic analytics we are interested in is referred to as "institutional analytics", generally used to understand factors related to running the business of the university [4].

Results of a survey applied to higher education institutions show that they use multiple components to integrate technology platforms for academic analytics [1]. Among the most used are:

- Data warehouses (DW): These are data repositories that store large amounts of structured, filtered, and processed data collected from different sources [5].
- Data marts: They can be seen as implementations of DW that are used for very particular areas or applications [6].
- Extract-Transform-Load (ETL) tools: These are tools used to conduct the ETL process, which is used in creating a DW [7].
- Operational data stores: These are databases that integrate data from multiple operational sources, apply some operations on them, and act as a source for the DW [8].
- Vendor-supplied reporting solutions.

Additionally, several methods and tools are used to access the information. Examples of these are multiple types of reports, queries, dashboards, alerts, and data extracts to offline tools.

The work described in this paper takes elements of academic analytics to support self-assessment and accreditation processes in an engineering college (EC) of a public university in Mexico, particularly focusing on the processes related to student cohorts' performance. To gain knowledge about student cohorts, our approach emphasizes understanding data and their relationships. This will help in creating new data and tools to share both data and information, providing a different perspective on the use and exploit of these data.

The rest of the paper is structured as follows. Section II presents related work about experiences with academic analytics in higher education. Section III describes the use of a data mining methodology for our data-centric approach. Section IV presents the results in terms of datasets, reports, and tools. Section V reflects on the achieved and potential benefits of this work, and Section VI states the conclusions and some actions needed to improve our approach.

II. Related Work

The impact of academic analytics has been explored and exploited with different objectives. In [9], authors discuss

Manuscript received November 8, 2023; revised August 13, 2024. Financial support for part of the activities described was provided by Universidad Autónoma de Baja California's General Coordination of Research and Postgraduate Studies under grant 105/6/C/25/4.

J. Ibarra-Esquer is a professor of Computer Engineering at the College of Engineering of Universidad Autónoma de Baja California, Mexicali, B.C., México. (e-mail: jorge.ibarra@uabc.edu.mx).

B. Flores-Rios is a researcher at the Institute of Engineering of Universidad Autónoma de Baja California, Mexicali, B.C., México. (e-mail: brenda.flores@uabc.edu.mx).

M. Astorga-Vargas is a professor of Computer Systems at the College of Engineering of Universidad Autónoma de Baja California, Mexicali, B.C., México. (e-mail: angelicaastorga@uabc.edu.mx).

M. González-Ramírez is a professor of Computer Engineering at the College of Engineering of Universidad Autónoma de Baja California, Mexicali, B.C., México. (e-mail: maria.gonzalez@uabc.edu.mx).

A. Justo-López is a professor of Computer Systems at the College of Engineering of Universidad Autónoma de Baja California, Mexicali, B.C., México. (e-mail: araceli.justo@uabc.edu.mx).

G. Chávez-Valenzuela is a professor of Computer Engineering at the College of Engineering of Universidad Autónoma de Baja California, Mexicali, B.C., México. (e-mail: gloria_chavez@uabc.edu.mx).

the use of big data technologies to extract information from modern DW in educational systems. They explain how these technologies can simplify data analysis and assist top-level managers in decision-making. The system they use is based on a DW that extracts structured and unstructured data from heterogeneous sources to create customizable charts and reports.

Relying on a DW, the work in [10] describes the implementation of a student progression system at a university. Authors present a thorough description of the DW architecture and the ETL process that was followed. Reports generated from the DW are used in the analysis of student performance. Data is gathered from different information systems where a variety of formats is used, and data redundancy is found. In their architecture, data from these systems goes through an ETL process and into the DW to be used in the creation of reports, analysis, and mining. The researchers in this study list some precautions that we have also found important for our implementation: guarantee a standardized format for stored data; guarantee accuracy of the accessed data; choose the right time and frequency to produce and store data into the DW.

The study in [11] discusses how building a unified DW from operational data in the university's information systems can benefit users by providing them with information. It also adds support for data analysis and mapping, as well as the capability to create customizable reports, charts, and graphics. Access to data in the DW also facilitates the use of data mining techniques to analyze academic data of students in various programs. Users can use their DW to create reports that meet their specific needs, making the process quicker than the previous system. This differs from our current approach, where a single department creates reports to ensure a standardized format and accurate data, as suggested by [10].

According to [12], a DW is a suitable tool for analyzing historical data from university repositories. Authors make use of students' behavior data from different private universities in Colombia that offer the Industrial engineering major and use artificial neural networks to predict future trends. They conclude on the difficulty added from the creation of specific applications and systems to retrieve data from heterogeneous sources. In our case, although we minimized the number of data sources, the data still had heterogeneous formats and required a similar process.

While describing an experience on the integration of a distributed academic DW, the work of [13] elaborates on the complexity of the ETL process. To leverage it, the first step was to identify and analyze the business process, which led to a proper selection of source data and proper testing of the results of the ETL process. The authors of the study emphasize the importance of not overlooking the data cleaning and conforming steps within this process when working with heterogeneous data.

In [14], authors present a detailed description of the architectural design of an educational warehouse. They depict a modular and flexible architecture comprising ETL, storage, data intelligence, data visualization, and data access modules. Its architecture aims to fulfill data requirements resulting from academic management and teaching-learning processes. The proposed design is based on the notion that analytical approaches to educational data must cater to the needs of both these processes.

The work in [15] describes the use of academic analytics to support the practice of academic advising. This process relies on the implementation and use of an academic analytics tool identified as the Student Success System. Results are based on perception from the academic advisors. While it mentions little improvement in the outcomes of the process, the study reports a positive attitude towards the use of these types of tools and draws attention to the fact that their use brought important value to assisting with student success. Our project aims to create tools that enhance academic processes using data from the DW. Academic advising is one of the initial processes we've focused on.

Several other works use data extracted from DWs to predict student performance. In [16], they use a dataset of the grades of first-year subjects to predict the performance of undergraduate computer science students (aptitude tests, grades, and awards). The study reported in [17] offers an alternative viewpoint, where the goal is to predict student dropout and success. In this case, they create a dataset that includes academic, demographic, socioeconomic, and macroeconomic data of students from 17 undergraduate programs from different fields of study. The work of [18] describes the data processes that allow creating datasets from educational administration systems and their use in predicting student academic achievement. They claim that using these data could guide college administrators, elevate teaching levels, and enhance schools' quality.

III. Data approach to Tracking Student Progression

As part of the ongoing process to accredit nine programs at the EC, a position was created to track student progress. This action was a response to the latest revisions in the accreditation organizations' reference frameworks, which required focusing on students' accomplishments and performance. These modifications, coupled with suggestions from our internal quality management system, indicated the need for new methods to track student progress and compute performance indicators.

The initial priority was to develop a system for calculating and disseminating student performance indicators. There was a previous experience with a legacy homegrown system that processed student academic records to calculate statistics pertaining to recent cohorts. Results were presented in tables that summarized the information from each student academic record, as shown in Fig. 1. A second section contained a brief of the behavior of the students in each of the engineering programs (Fig. 2).

While it helped with calculating indicators, this system had limitations, highlighting the need for better ways to process data. Interpreting the outcomes required a substantial amount of effort, as some results were presented as text descriptions instead of numerical or categorical data that could be readily subjected to further calculations. The update process was slow, as the system had to analyze the entire dataset, which adds over 20,000 new records after each semester. According to the individual responsible for utilizing the system, this process would take several hours to complete. Finally, the results did not include all the significant data from the

dataset, making it impossible to track individual student progress or relate a specific student to a cohort.

Our strategy involves merging data from various sources to generate comprehensive reports that add value to institutional information. These reports will have global data on EC performance and specific details about its programs. They will provide accurate and timely access to relevant information about the individual progression of each EC student. The aim is not to replace existing information, but to expand and enhance it. Thus, the resulting information would be beneficial for tasks like advising and tutoring, which impact student performance.

As the plan was on designing and developing a system able to impact different processes using the same student data, the first decision was to adopt a methodology that supported the identification of relevant data, the discovery of information through them, and ultimately, the generation of institutional knowledge for decision-making. Cross-Industry Standard Process for Data Mining (CRISP-DM) was chosen for the first stage of our approach, as it provides a framework for developing data mining projects that is independent of the field and technology [19]. This methodology consists of six phases in an iterative process (Fig. 3) and was used to direct the activities and decisions as described next.

A. Business understanding phase

The business understanding phase includes defining the university context, understanding student progression, and reviewing the requirements outlined in the relevant reference frameworks of the AO. At this stage, the scope was delimited to engineering programs and kept under institutional definitions and regulations. The principle of unification of concepts was employed to establish a common understanding between the AO and the CE. The system's scope was established to include all cohorts of students in the current curriculum for each program offered in the CE. Curriculum modifications were underway when the project started, and they would be incorporated once in effect.

In this same phase, the need to create strategies for knowledge management (KM) is observed. KM is defined as a systematic organizational process to acquire, organize, and communicate the tacit and explicit knowledge of the members of an organization, so that others can use it to increase their efficiency and productivity [20]. According to [21] from the definitions proposed in [22], [23], explicit knowledge is objective, rational, and is codified in different media and formats, while tacit knowledge stems from personal experience, reflection, internalization, or talent of an individual. In other words, explicit knowledge is acquired through documents, while tacit knowledge is gained through interactions with people.

B. Data understanding phase

In this phase, we gather the data and conduct an exhaustive analysis to comprehend their meaning and structure. We analyzed multiple data sources and ultimately chose two that have all the necessary information for tracking student progress and calculating performance indicators.

- Information system of the institution's student services area: This system provides access to reports on the student population, course performance, academic advisor assignments, and individual academic records of current and former students of the EC. The system offers extra reports that are not needed for updating our data, but can be useful for handling outliers and missing data. All these reports are generated from institutional databases, but direct access to the databases is not granted because of privacy and security policies.
- EC graduation department: Generates reports that include lists of graduate students, the modality by which they got their engineering degree, and the graduation date.

The work in [14] suggests that the data should address the educational context, academic performance, and behavior to achieve better outcomes, as supported by these two sources. The information system of the student services area is the primary data source. Its data allow us to reconstruct the academic record and pathway of any student enrolled in any program in the EC in the last 30 years. These data permit measuring, comparing, analyzing, and forecasting performance indicators globally or individually. One important feature of this system is that it experiences a period of intensive data updates from the end of one semester to the end of the first month of the next. After that, it remains largely the same. This behavior allows us to set a frequency for acquiring usable data and schedule all the data activities and processes.

C. Data preparation phase

This phase includes the activities necessary to make the data usable. The graduation department reports are delivered as spreadsheets that can be searched directly through formulas within the same sheets or using custom-made computer applications. The graduation department produces approximately three to five reports every semester, each corresponding to a graduation event where multiple graduates participate after meeting the university's degree requirements. These events do not occur on a set schedule, which explains why the number of events per semester may vary.

Even though data is well-structured, graduates' information is entered manually, and it is usually necessary to perform additional verification to identify and correct errors. These occur typically as incorrect or non-existent student IDs, or inconsistencies in the names used to refer to the graduation modalities.

The largest volume of data is obtained from the information system of the student services area. Reports from this system are downloaded as readable but not processable format, such as PDF documents, or as spreadsheets where the data are not necessarily found in consecutive columns or rows and require some pre-processing and formatting to be interpreted correctly.

The academic history dataset is the largest in volume (The September 2023 update comprised 14 files, totaling 598MB of data). This dataset can be downloaded as a set of CSV files - one file for each major. It contains each record related to each course completed by every student who has been enrolled in any major in the EC since circa 1995. These files were the input that the legacy system used to calculate indicators and statistics.

Although the format in which the academic history is obtained and the structure of the data allow each record

STUDENT ID	PROGRAM	CURRICULA	TRANSFER STUDENT	GPA	FAILED COURSES	STATUS	SUMMARY
1234567	ELECTRICAL ENGINEERING	20092	NO	82.21	2	2	Dropped out after semester 2015-2
1235678	ELECTRICAL ENGINEERING	20092	NO	91.55	0	5	Graduated in 2017-2
1236789	ELECTRICAL ENGINEERING	20092	NO	84.66	0	5	Graduated in 2017-1
1237900	ELECTRICAL ENGINEERING	20092	NO	86.35	0	5	Graduated in 2017-2
1239011	ELECTRICAL ENGINEERING	20092	NO	76.91	6	0	98% of credits earned
1240122	ELECTRICAL ENGINEERING	20092	NO	86.16	1	5	Graduated in 2017-2
1241233	ELECTRICAL ENGINEERING	20092	NO	83.67	6	0	93% of credits earned
1242344	MECHANICAL ENGINEERING	20092	YES	80.62	8	0	89% of credits earned
1243455	MECHANICAL ENGINEERING	20092	YES	83.24	7	0	73% of credits earned
1244566	MECHANICAL ENGINEERING	20092	NO	96.02	0	5	Graduated in 2017-1
1245677	MECHANICAL ENGINEERING	20092	NO	82.09	7	0	40% of credits earned
1246788	MECHANICAL ENGINEERING	20092	NO	88.88	0	5	Graduated in 2017-1
1247899	MECHANICAL ENGINEERING	20092	NO	83.7	5	0	75% of credits earned
1249010	COMPUTER ENGINEERING	20071	NO	87.66	4	0	71% of credits earned
1250121	COMPUTER ENGINEERING	20071	NO	95.08	0	5	Graduated in 2017-1
1251232	COMPUTER ENGINEERING	20071	YES	83.76	1	5	Graduated in 2017-2
1252343	CIVIL ENGINEERING	20082	NO	94.53	0	5	Graduated in 2017-1
1253454	CIVIL ENGINEERING	20082	NO	91.86	4	2	Dropped out after semester 2014-2
1254565	CIVIL ENGINEERING	20082	NO	91.31	0	0	99% of credits earned
1255676	CIVIL ENGINEERING	20082	NO	95.12	0	5	Graduated in 2017-1

Fig. 1. Sample of report structure of the students' academic records in the legacy system

PROGRAM: MECHANICAL ENGINEERING										
Cohort	**2013-1**	**2013-2**	**2014-1**	**2014-2**	**2015-1**	**2015-2**	**2016-1**	**2016-2**	**2017-1**	**2017-2**
Total students	43	47	40	45	38	46	44	43	48	46
Active	11	20	30	38	35	39	40	39	45	46
Dropout	8	8	5	6	3	2	3	4	3	0
Academic dismiss	3	1	2	0	0	3	0	0	0	0
Changed major	1	3	0	1	0	2	1	0	0	0
Graduated	20	15	3	0	0	0	0	0	0	0

Fig. 2. Report of program performance indicators in the legacy system

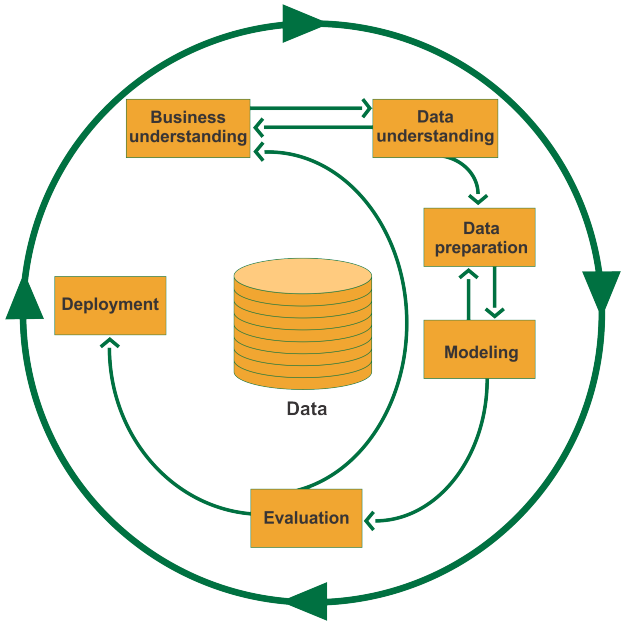


Fig. 3. Phases of the CRISP-DM methodology

to be used immediately, it contains many redundant data across records and some variables within single records that can be inferred from others. Each file contains 54 variables, but some of them are a combination of others and were discarded. For instance, there is one variable for the total compulsory courses in the curriculum, another for passed courses, and one for remaining courses, where passed courses can also be calculated by counting the number of rows with an approval grade. Selected variables are categorized and described in Table I. One drawback is that variables are unnamed in the dataset; names and descriptions were taken from a document elaborated by the designer of the legacy system. Pre-processing this dataset is simpler than other reports because of the standard structure guaranteed by the CSV format. However, it's still necessary to reduce data size and improve processing speed for analysis.

Through this phase, data tasks heavily relied on the use of spreadsheets and formulas within them. While spreadsheets may not be the most efficient, they make it simple to analyze data and discover connections between variables.

D. Modeling phase

The data comprehension, preparation and modeling phases were the subject of several iterations. Using the outcomes from these phases, we found how the data and reports are related, and created new ways to store student progression data; developed mechanisms to calculate indicators by combining and converting data according to business rules; and established the formats and schemes for the presentation and publication of data and indicators. New knowledge was simultaneously developed for different EC areas involved in tracking students' academic performance.

The first iterations of the methodology provided knowledge about data, business rules, and data resources. As mentioned in the data preparation phase, spreadsheets were

TABLE I
DESCRIPTION OF THE ACADEMIC HISTORY DATASET

Category	Variables
Student identification	Record ID, student ID, campus ID, student index, name, last name. Record ID acts as the identifier for a student-major-curriculum enrollment; if a student switches to a new curriculum or enrolls in a different major, a new record ID is created while keeping the same student ID. Student ID is the concatenation of campus ID and student index. All these variables are repeated in every row of the dataset that belongs to the same student.
Program identification	College ID, program ID, curriculum as year-period, curriculum as numeric value, and total courses and credits in the curriculum, separated as compulsory, elective, and professional practice. For practical uses, a program is assigned an ID formed by the concatenation of college and program ID from this dataset. All these variables are repeated in every row of the dataset that belongs to the same student.
Academic status	Status key, GPA, and credits earned, separated as compulsory, elective, and professional practice. Status key has one of the values: 1 for active students, 2 for dropouts, 3 for students that finished all their credits, 4 for students that fulfilled all their graduation requirements, and 7 for academic dismissals. All these variables are repeated in every row of the dataset that belongs to the same student.
Course record	Course ID, course name, type of evaluation (ID and name), date, semester ID (year and period), grade, course credits, and college ID. There are also a few variables that are only used for courses credited to transfer students, students switching curriculum, or students changing major. Each row in the dataset is a record of a course where a student was graded, either as fail or approval. College ID is needed because they can take elective courses at different colleges, e.g. arts and sports.

used extensively to understand and prepare data. Institutional reports were formatted and copied to tables in spreadsheets; the reports were used to create new tables, where formulas were applied to associate and validate data. This process also involved combining the data from the reports to generate new variables for the required datasets. As a result, a historical file was created to monitor students' progression. The file had a summary of academic records for 11,757 students and was used as the starting point for automated extraction of information, calculation, and presentation of indicators. The dataset included variables such as cohort, student ID, name, major, most recent semester, percentage of credits earned, current academic status, and modality of graduation.

The modeling phase ultimately resulted in the design of a customized database tailored to support all aspects of student pathways and performance indicators and statistics. Its relationships summarize the core information extracted from the different institutional reports, enhanced by applying knowledge rules from the process. The database (STrEs-DB) consists of 16 tables (Fig. 4), which contain the general characteristics of each major in terms of compulsory and elective credits and the history of curriculum modifications. It also stores data about each student or graduate, linking them to their major, course records, and current status (e.g., active student, graduate, dropout, etc.) [24].

Adaptation of the ETL process was necessary for database integration. Instead of using spreadsheets, a set of programs and scripts were created to extract and transform the relevant data from the academic history. These programs also performed validations and applied knowledge rules before loading the data into the database. This process is depicted in Fig. 5 [25].

E. Evaluation phase

In this phase, we compared the indicators and statistics from the modeled dataset to those from the legacy system and the institution's published statistics. Using these data as a reference, we validated the scheme defined for the calculation of indicators and the creation of new data. Furthermore, it was necessary to iterate on the previous phases of the methodology to adjust it and correct any errors or differences that were found. We also ensured that the required datasets for calculating indicators and statistics could be generated from the database.

F. Deployment phase

The last step in the process is to make models, datasets, and reports to be used for predictions, analysis, and decision making. Initially, we chose to provide the datasets as tables in spreadsheets. We included formulas to calculate the indicators and added visualizations to make the information easier to understand and use. Our aim was to democratize the data by publishing it and enabling users to generate their own reports, graphs, and visualizations based on their specific needs. Additionally, it would allow full access to data and reports with no Internet connection. This decision became significant when Covid-19 prevented access to school facilities, as firewalls limit access to these systems from outside the university's network.

Initially, we used a Google Drive repository for each program to share the datasets and reports. This ensured that only authorized individuals could access the files, keeping the information safe and confidential. In this repository, there is a private section that holds all the data and work documents for the student progression tracking system that are used to calculate, maintain, and update indicators and information.

The database contains a set of customized stored procedures to create the datasets. After created, they are copied to spreadsheet tables. Indicators are then recalculated by adjusting variables that impact the formulas (e.g., current semester identifier). The following section elaborates on the deployment phase, expanding on details about the datasets, reports, and tools.

IV. DATASETS, REPORTS, AND TOOLS

As mentioned above, stored procedures create the datasets required for calculations and reports. There are three basic

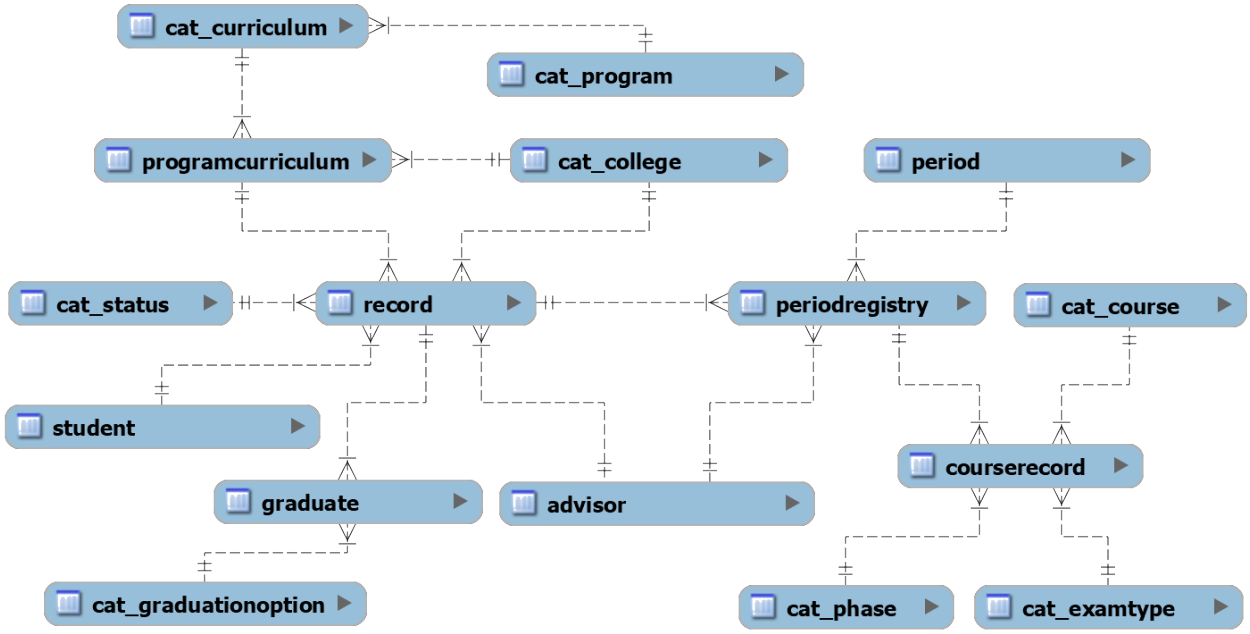


Fig. 4. Database model for STrEs-DB

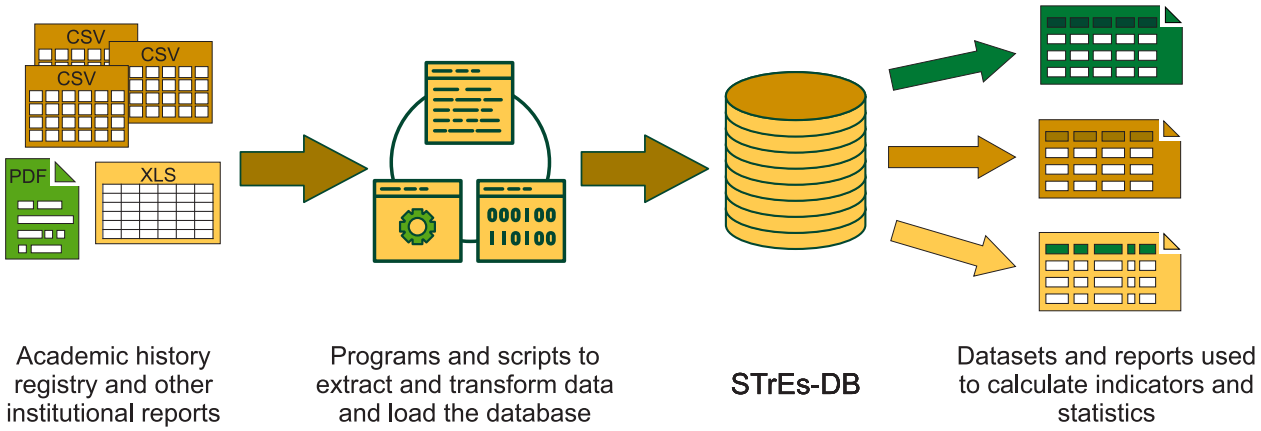


Fig. 5. Extract-Transform-Load process (Based on [25])

integrated reports that summarize cohorts' behavior and performance. An additional tool is used to track the individual performance of students as part of academic advisement processes.

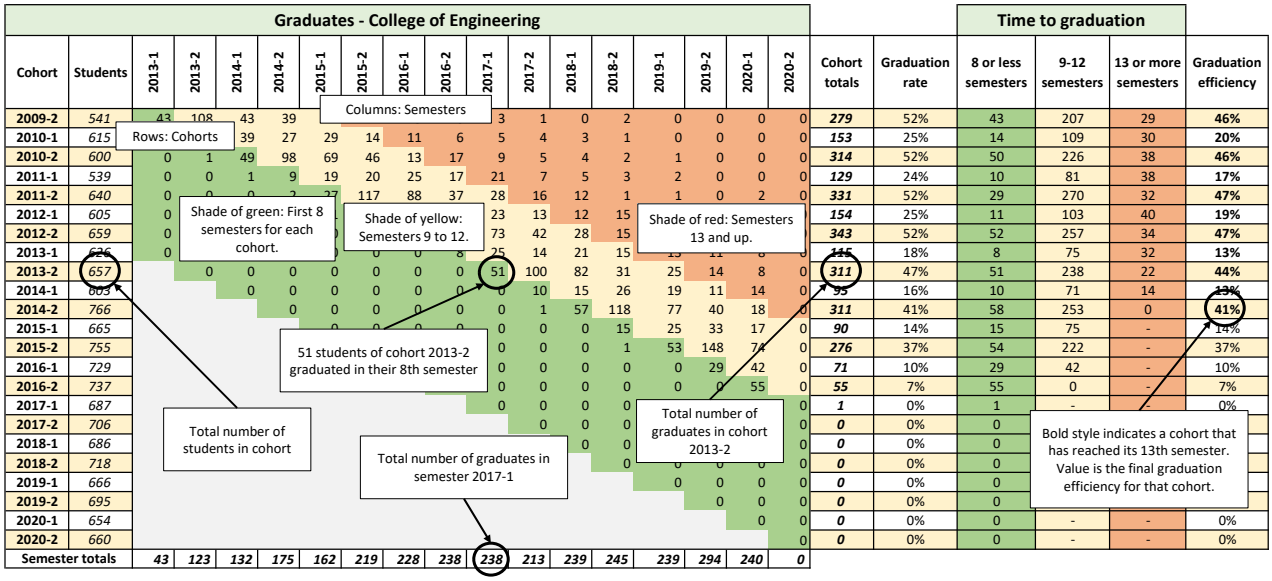
A. Tracking cohorts' behavior

The last part of the ETL process shown in Fig. 5 is the creation of datasets from the database. The stored procedures take data from the database tables and format it. This formatted data is then copied to spreadsheets where additional rules are applied to verify and validate the data. This process creates new data and information from the existing datasets. Finally, indicators are calculated, and reports and visualizations are generated. These additional steps improve the data and help identify changes or new rules for the transformation scripts.

The enhanced and verified datasets are then loaded to the deployment spreadsheets, where there are formulas that calculate the indicators and create visualizations according to selectable parameters or filters. The relevant data for each program is included in separate spreadsheets, and copies of these spreadsheets are made for every program. Finally, these copies are shared on the Google Drive repository.

1) Initial deployment: The early iterations of the process were mainly focused on the business and data understanding phases. Source data were extracted and transformed directly on spreadsheets through a sequence of tables and filters until the usable dataset was created. This dataset was copied to a table and used to create visualizations like the one in Fig. 6. A chart was created for each performance indicator, using colored stripes to represent time ranges for different student outcomes: green shades for students who complete their studies within the curriculum's specified timeframe; yellow for those who finish within the maximum period deemed appropriate by the OA; red for those who exceed this timeframe. The data was organized by cohorts and semesters, making it easy to see how students progressed.

Updating and using the charts became a concern because they grew larger with each update cycle, and the dataset preparation process was time-consuming. To accommodate



Cohort	Students	2013-1	2013-2	2014-1	2014-2	2015-1	2015-2	2016-1	2016-2	2017-1	2017-2	2018-1	2018-2	2019-1	2019-2	2020-1	2020-2	Cohort totals	Graduation rate	8 or less semesters	9-12 semesters	13 or more semesters	Graduation efficiency
2009-2	541	43	108	43	39					3	1	0	2	0	0	0	0	279	52%	43	207	29	46%
2010-1	615			39	27	29	14	11	6	5	4	3	1	0	0	0	0	153	25%	14	109	30	20%
2010-2	600	0	1	49	98	69	46	13	17	9	5	4	2	1	0	0	0	314	52%	50	226	38	46%
2011-1	539	0	0	1	9	19	20	25	17	21	7	5	3	2	0	0	0	129	24%	10	81	38	17%
2011-2	640	0	0	0	2	27	117	88	37	28	16	12	1	1	0	2	0	331	52%	29	270	32	47%
2012-1	605	0	0	0	0	0	0	0	0	23	13	12	15	0	0	0	0	154	25%	11	103	40	19%
2012-2	659	0	0	0	0	0	0	0	0	73	42	28	15	0	0	0	0	343	52%	52	257	34	47%
2013-1	626	0	0	0	0	0	0	0	0	25	14	21	15	13	11	8	0	115	18%	8	75	32	13%
2013-2	657		0	0	0	0	0	0	0	51	100	82	31	25	14	8	0	311	47%	51	238	22	44%
2014-1	665			0	0	0	0	0	0	0	10	15	26	19	11	14	0	95	16%	10	71	14	13%
2014-2	766				0	0	0	0	0	0	1	57	118	77	40	18	0	311	41%	58	253	0	41%
2015-1	665					0	0	0	0	0	0	0	15	25	33	17	0	90	14%	15	75	-	14%
2015-2	755						0	0	0	0	0	0	1	53	148	74	0	276	37%	54	222	-	37%
2016-1	729							0	0	0	0	0	0	0	29	42	0	71	10%	29	42	-	10%
2016-2	737								0	0	0	0	0	0	0	55	0	55	7%	55	0	-	7%
2017-1	687									0	0	0	0	0	0	0	0	1	0%	1	-	-	0%
2017-2	706										0	0	0	0	0	0	0	0	0%	0	-	-	0%
2018-1	686											0	0	0	0	0	0	0	0%	0	-	-	0%
2018-2	718												0	0	0	0	0	0	0%	0	-	-	0%
2019-1	666													0	0	0	0	0	0%	0	-	-	0%
2019-2	695														0	0	0	0	0%	0	-	-	0%
2020-1	654															0	0	0	0%	0	-	-	0%
2020-2	660																0	0	0%	0	-	-	0%
Semester totals		43	123	132	175	162	219	228	238	238	213	239	245	239	294	240	0						

Fig. 6. Visualization of program performance indicators in the initial deployment

the latest cohort, we had to add a new row along with a column for the new semester. Formulas needed only minor modifications, both for calculations and formatting rules.

These charts were designed not only for displaying indicators but also to observe and analyze cohort behavior through time. Moving from left to right in a single row described the latter. However, as data from new cohorts and semesters were added to the chart, empty space formed at the bottom left, and the top right had only zeros when older students graduated or dropped out.

2) Dataset creation: When the database was deployed, creation of datasets became an easier and fastest task. As a result, we were able to focus our efforts on redesigning and revamping the reports.

A stored procedure retrieves 19 fields from the database to create the dataset. These fields summarize a student's record, including their ID, program, curriculum credits, earned credits, first enrollment semester, recent academic period, and current status. Data includes every student that has enrolled in a major at the EC since 2009. A listing of first semester students is obtained from a separate report and added to the results from the database, as these students are included in the academic history dataset, and hence added to the database, after they finish their first semester and their first grades are officially recorded by the student services area. The most recent update accounted for a total of 20,055 records.

These records are copied into a table in a spreadsheet where a formula identifies students with multiple records. One of the improvements we made over the academic history is a more detailed specification of the student academic status and the association of multiple student's records in the following cases:

- Records marked as dropouts for cases where the student transferred to another program or requested to join an updated curriculum in their same program.
- Records of students who enrolled in a second program after successfully finishing their first, and that requested to transfer credits from the previous program.

The scripts in the second step of the ETL process identify most of these cases and set the correct status to the records in the database. However, there may be instances where the scripts fail, hence the need for verification. Usually, these situations become additions to the knowledge rules and are later programmed into the scripts.

Another table contains formulas that calculate a student's enrollment duration, percentage of earned credits, and generate values for semester identification in reports. In this table, a dataset from a second stored procedure is used to check if a student who is marked as a dropout has any academic restrictions for re-enrollment. A final formula tags all the records that were modified, so changes can be applied back into the database.

The resulting dataset has 20 variables and it's copied to a deployment spreadsheet. This dataset is used to calculate and format the reports described in the following sections. An anonymized excerpt from the dataset is shown in Fig. 7. The last three variables are not shown in the figure; these are numeric values needed for calculations that make an adjustment for students whose latest academic term was an inter-semester.

3) Integrated report of cohort performance indicators: First, all the cohort-related indicators were computed and consolidated into a single table (i.e., size of the cohort, total students still registered, total dropouts, total graduates, and rates for each indicator). Some additional values, like average time to finish studies, were also included to aid in analyzing performance. Fig. 8 shows the redesigned report of cohort performance indicators. In the updated deployment, all indicators for each cohort are shown together in one row, eliminating the need to scroll through multiple charts or tables. This makes it easier to track and compare cohorts. In addition, updating the report only requires copying the new data into a table in the file and adding a new row at the bottom of the report.

On top of the chart, three lists were included and used as filters. The chart can calculate and show data based on criteria from various accreditation boards. It can also filter

Record ID	Cohort	Student ID	Name	Program	Curricula	Status key	Status	Latest period	GPA	Credits earned	Finished CC	Total semesters	CC semesters	Program semesters	Transfer	School
289725	20111	1113452	RUBEN LOPEZ	Civil Engineering	2009-2	2	Dropout	20121	87.04	45%	20111	3	0	3	1	140
153606	20142	1118883	ALMA ROJAS	Mechanical Engineering	2009-2	4	Graduate	20182	81.84	100%	20152	9	3	6	0	140
151824	20142	1117922	ROBERTO SANCHEZ	Engineering Common Core	2009-2	2	Dropout	20151	87	5%	0	2	2	0	0	140
445998	20182	1153111	ROSARIO DIAZ	Mechatronics Engineering	2009-2	6	Switched curriculum	20222	76.2	61%	20191	9	2	7	0	140
355779	20182	1153111	ROSARIO DIAZ	Mechatronics Engineering	2019-2	1	Active	20231	76.65	59%	20191	10	2	8	0	140
107423	20112	1100234	BEATRIZ LLAMAS	Computer Engineering	2009-2	2	Dropout	20132	76.37	47%	20112	5	0	5	1	140
465708	20192	1121223	TEODORO GARCIA	Aerospace Engineering	2020-1	1	Active	20231	88	78%	20201	8	2	6	0	140
498662	20202	1127895	MARIANO SOTO	Renewable Energy Engineering	2020-1	1	Active	20231	85.55	53%	20211	6	2	4	0	140
316121	20212	1137654	PEDRO JUAREZ	Civil Engineering	2020-1	1	Active	20231	75.31	22%	20222	4	3	1	0	140
127099	20131	1109343	DAVID MENDEZ	Aerospace Engineering	2009-2	5	Changed major	20141	76.82	31%	20132	3	2	1	0	140
171248	20142	1109343	DAVID MENDEZ	Mechanical Engineering	2009-2	4	Graduate	20172	80	100%	0	7	0	7	1	140
152412	20151	1120466	JORGE RODRIGUEZ	Mechatronics Engineering	2009-2	7	Academic dismiss	20191	78.71	42%	20162	9	4	5	0	140
246237	20182	1187632	ALICIA MEZA	Mechanical Engineering	2009-2	2	Dropout	20201	75.06	30%	20192	4	3	1	0	140
248613	20191	1190043	MOISES ALVAREZ	Mechanical Engineering	2020-1	2	Dropout	20202	71.27	25%	20202	4	4	0	0	290
365563	20142	1110128	EDGAR ARANDA	Electrical Engineering	2009-2	4	Graduate	20201	79.6	100%	20152	12	3	9	0	290
374178	20161	1132987	FERNANDO RUIZ	Aerospace Engineering	2009-2	4	Graduate	20212	80.46	100%	20182	12	6	6	0	140

Fig. 7. Dataset structure and sample data

data based on curriculum changes and include or exclude transfer students.

4) Cohort behavior report: The behavior of the cohort, that in the initial deployment could be observed by moving horizontally in the charts (see Fig. 6), was now converted into an independent report that is generated from the same dataset. An example of this report is shown in Fig. 9.

The chart shows one cohort at a time, selected by typing its identifier in a cell on top of the chart. Color stripes have the same use as in the initial deployment. The first column contains the semester identifier, starting with the top row representing the first cohort and ending with the current semester at the bottom. The second column displays the number of students in the cohort at the start of each semester, with the first row representing the initial cohort size. Moving left shows us the number of students who graduated, dropped out, or transferred, and the percentage of students still in the cohort at the end of the semester.

Similar to the previous report, data is now shown together instead of needing to scroll through several tables to get a full picture of a cohort's behavior. Adding filters instead of showing all the cohorts at once allows one to focus on a specific cohort to analyze its performance.

5) Program timeline: The bottom row in the charts of the initial deployment (Fig. 6) was an aggregate of the corresponding indicator for all the cohorts in every semester. By combining all these data together, we gained insight on the program behavior through time. Fig. 10 presents an integrated view that acts as a program timeline. In this case, each row corresponds to a semester and columns show the value for each indicator in such semester.

As with the previous reports, the same dataset is used to calculate data in this table. The dataset and the three reports are integrated into a single file, each in a separate sheet. Reports are first created as college-level, and then a report for each program is obtained by using the corresponding subset of the data.

B. Tools for individualized tracking

The dataset mentioned earlier has information on students' identities, academic status, and outcomes. However, it doesn't provide details on how they achieved those outcomes or if any actions are necessary for academic success. A tool for academic advisement was first designed to list all currently enrolled students and tag them as normal, risk, or critical condition, using a formula that takes as parameters credits earned and the number of semesters since each student first enrolled at the university. The tool highlights each student in the list with green, yellow, or red, depending on the assigned tag. Hence, the tool is identified as the Academic Advisement Semaphore (AAS) (Fig. 11).

Two stored procedures from the database are used to create the dataset. This dataset resembles the one in the previous section, but it only contains academic records of students from the current and previous semesters, as well as those who switched curriculums since the last update. The list of academic advisors is downloaded from an institutional repository and formatted to be used in the tool. The AAS [26] is given to all academic advisors in the college. They should use it according to criteria set by the advisement department.

A second tool that complements the AAS was developed to provide the advisor with a higher level of detail about a student's academic progression. It uses data from the same dataset as the AAS, along with another dataset that includes summaries of each student's academic periods. There are approximately 26,000 records included in this supplementary dataset. An institutional report helped us create a dataset that lists which courses each student is registered for in the ongoing semester. Including this information is important for the advisors, as there is not a widely available system that allows them to visualize the courses their students are taking. Because of its origin and relationship with the AAS, this tool (Fig. 12) is identified as the Individual Advisement Semaphore (IAS) [27].

The IAS offers an interface that simplifies the advisor's task by requiring only the input of a student ID. For every student, the IAS displays the same details as the AAS. The list of courses the student is taking is shown at the bottom of the page. Finally, the IAS gives a condensed timeline of the student's academic behavior for each academic period.

V. Discussion

The ETL process has matured enough to become fully repeatable and to adapt to the requirements for new datasets and reports that support tracking student academic performance and progression. Integrating the procedures into the database simplifies creating personalized datasets and updating reports in a more efficient manner.

Include transfer	No	Curricula	All	Standard	CACEI	Overall graduation efficiency	33.0%

Updated March 7 2023

Time to graduation

Cohort	Students	Transfer students	Active	Graduates	Dropouts	Changed major	Switched curriculum	Academic dismiss	Retention rate	Graduation rate	Minimum defined by the curricula	Max for graduation efficiency	Exceeds time for efficiency	Graduation efficiency	Dropout rate	Academic dismiss rate
2013-2	654	31	1	332	230	11	0	80	50.9%	50.8%	50	250	32	**45.9%**	36.9%	12.2%
2014-1	602	26	0	118	361	5	0	118	19.6%	19.6%	10	76	32	**14.3%**	60.8%	19.6%
2014-2	747	28	0	366	249	19	0	113	49.0%	49.0%	62	261	43	**43.2%**	35.9%	15.1%
2015-1	664	12	0	155	378	2	0	129	23.3%	23.3%	17	100	38	**17.6%**	57.2%	19.4%
2015-2	749	24	4	393	253	9	4	90	53.0%	52.5%	59	310	24	**49.3%**	35.0%	12.0%
2016-1	728	35	12	189	394	7	5	126	27.6%	26.0%	30	131	28	**22.1%**	55.1%	17.3%
2016-2	731	14	17	385	247	15	5	67	55.0%	52.7%	57	313	15	**50.6%**	35.8%	9.2%
2017-1	680	10	33	126	402	5	19	114	23.4%	18.5%	12	114	0	**18.5%**	59.9%	16.8%
2017-2	701	12	56	324	250	11	36	60	54.2%	46.2%	58	266	0	46.2%	37.2%	8.6%
2018-1	687	12	108	146	335	9	53	89	37.0%	21.3%	18	128	0	21.3%	50.1%	13.0%
2018-2	717	17	174	211	264	9	104	59	53.7%	29.4%	56	155	0	29.4%	38.1%	8.2%
2019-1	659	17	222	26	335	13	43	63	37.6%	3.9%	26	0	0	3.9%	52.8%	9.6%
2019-2	694	12	456	0	173	13	0	52	65.7%	0.0%	0	0	0	0.0%	26.8%	7.5%
2020-1	676	15	322	0	258	14	0	82	47.6%	0.0%	0	0	0	0.0%	40.2%	12.1%
2020-2	689	7	425	0	190	24	0	50	61.7%	0.0%	0	0	0	0.0%	31.1%	7.3%
2021-1	673	7	322	0	271	4	0	76	47.8%	0.0%	0	0	0	0.0%	40.9%	11.3%
2021-2	735	22	530	0	172	12	0	21	72.1%	0.0%	0	0	0	0.0%	25.0%	2.9%
2022-1	729	27	445	0	283	1	0	0	61.0%	0.0%	0	0	0	0.0%	39.0%	0.0%
2022-2	797	18	677	0	120	0	0	0	84.9%	0.0%	0	0	0	0.0%	15.1%	0.0%
2023-1	731	23	731	0	0	0	0	0	100.0%	0.0%	0	0	0	0.0%	0.0%	0.0%

Fig. 8. Interface of the integrated report of cohort performance indicators

Cohort	2016-2
Include transfer	No
Curricula	All

Semester	Active students	Graduates	Dropouts	Changed major	Academic dismiss	Switched curriculum	Retention rate
2016-2	731	0	32	0	0	0	95.6%
2017-1	699	0	85	1	0	0	83.9%
2017-2	613		36	5	9	0	77.0%
2018-1	563		29	4	20	0	69.8%
2018-2	510		12	3	3	0	67.3%
2019-1	492		5	1	10	0	65.1%
2019-2	476	0	12	1	11	0	61.8%
2020-1	452	57	2	0	1	0	61.4%
2020-2	392	142	7		6	0	59.6%
2021-1	237	104	6		1	0	58.7%
2021-2	126	39	11		3	1	56.8%
2022-1	73	28	7	0			55.4%
2022-2	35	15	3	0			55.0%
2023-1	17	0	0	0			55.0%

Shade of green: First 8 semesters for the selected cohort.

Shade of yellow: Semesters 9 to 12.

Shade of red: Semesters 13 and up.

Fig. 9. Interface of the cohort behavior report

Reports and tools listed in the previous section have been used in three successful accreditation processes. The first of them was carried by the Accreditation Council for Engineering Education (CACEI), which is the largest Mexican AO on engineering programs. This experience involved 10 programs offered in the EC, acting as an initial validation of our approach and a testing scenario for functionality and usefulness of earlier versions of the tools. The second time was in a process conducted by the National Council for Accreditation in Computing and Computer (CONAIC). Their information requirements resulted in enhancements to the indicators included in some reports. The third time was an international experience where three engineering programs worked towards meeting the standards set by the EUR-ACE Framework for Accreditation. ANECA, the National Agency for Quality Assessment and Accreditation, supervised this process. ANECA requested specific indicators for the programs under accreditation. Upon conclusion of the process, they were included in the reports for all EC programs. Additionally, the information from this process helps with internal decision-making for program organization and management. It is also valuable for yearly reports presented by the EC's Principal.

VI. Conclusion

We employ a data-centric strategy that combines academic analytics to monitor student academic performance and support assessment and accreditation processes. With the help of a data mining methodology, we discovered connections between data sources and extracted valuable information about business processes. This information was then used in an ETL process to gather data from different reports and store them in a database. The database contains details about students, their academic history, curriculum, and academic advisors. This database can generate multiple datasets through stored procedures. These datasets can be used for various purposes, such as creating reports and tools. They not only serve the initial purpose of calculating and updating indicators for program accreditation, but also provide information for other processes like academic advisement.

Focusing on data processes over defining and specifying requirements for developing an application presented clear

Include transfer	No	Curricula	All

Cohort	Students	Transfer students	Graduates	Dropouts	Changed major	Switched curriculum	Academic dismiss
2013-2	654	31	124	222	9	0	54
2014-1	602	26	133	361	13	0	105
2014-2	747	28	173	198	10	0	63
2015-1	664	12	164	377	10	0	109
2015-2	749	24	225	188	15	0	100
2016-1	728	35	225	337	10	0	112
2016-2	731	14	242	233	6	0	106
2017-1	680	10	233	373	8	0	124
2017-2	701	12	245	304	11	0	84
2018-1	687	12	259	478	8	0	175
2018-2	717	17	252	257	11	0	10
2019-1	659	17	263	282	9	1	129
2019-2	694	12	306	217	10	7	119
2020-1	676	15	258	309	3	17	51
2020-2	689	7	274	228	6	77	25
2021-1	673	7	326	339	19	75	61
2021-2	735	22	294	272	27	45	99
2022-1	729	27	283	422	12	34	113
2022-2	797	18	344	365	23	13	146
2023-1	731	23	0	0	0	0	0

Fig. 10. Interface of the program timeline

Cohort	Student ID	Name	Program	Curricula	Credits earned	GPA	Semesters	Semesters skipped	Finished CC	Advisor ID	Advisor	Transfer student	Estimated time to graduation	Status
20212	1162187	CARLOS CRUZ	Aerospace Engineering	2020-1	57%	92.74	3	0	20212	18452	ALEJANDRO ROBLE		mesters	Normal
20201	1162464	JOHAN MARTINEZ	Civil Engineering	2020-1	33%	86.1	6	3	20221	19377	ROSARIO MARQUE		mesters	Risk
20201	1162530	LORENA SALAS	Mechanical Engineering	2020-1	53%	81.16	6	0	20202	11367	MIGUEL ENCINAS	NO	12 semesters	Risk
20201	1162844	LUIS ROMO	Mechanical Engineering	2020-1	60%	82.25	6	0	20202	11367	MIGUEL ENCINAS		mesters	Risk
20201	1162869	RANDY DELGADO	Civil Engineering	2020-1	73%	85.66	6	0	20202	19377	ROSARIO MARQU		mesters	Risk
20192	1163071	RAMON CASTRO	Mechatronics Engineering	2019-2	40%	74.35	7	0	20202	16832	JESUS HEREDIA	NO	18 semesters	Critical
20201	1163108	SUSANA GIL	Renewable Energy Engineering	2020-1	61%	89.11	6	0	20202	17656	EDGAR ALVAREZ	NO	10 semesters	Risk
20192	1163783	MIRNA MENDEZ	Aerospace Engineering	2020-1	57%	74.26	7	0	20201	18452	ALEJANDRO ROBLES	NO	13 semesters	Critical
20192	1163877	VICENTE RUIZ	Mechanical Engineering	2020-1	51%	73.73	7	0	20201	11367	MIGUEL ENCINAS		mesters	Critical
20192	1163940	JOSE ZAVALA	Mechanical Engineering	2020-1	53%	80.84	7	0	20211	11367	MIGUEL ENCINAS		mesters	Critical
20222	1163956	GABRIELA SANTOS	Mechatronics Engineering	2019-2	35%	89.95	1	0		16832	JESUS HEREDIA	YES	3 semesters	Normal
20202	1164915	ALFREDO MONTES	Aerospace Engineering	2020-1	53%	88.38	5	0	20211	18452	ALEJANDRO ROBLES	NO	10 semesters	Risk
20221	1164992	CRISTINA PEREZ	Renewable Energy Engineering	2020-1	54%	92.81	2	0		17656	EDGAR ALVAREZ	YES	4 semesters	Normal

Normal status highlighted green

Risk status highlighted yellow

Critical status highlighted red

Fig. 11. Academic Advisement Semaphore

Student ID	**1174325**		
	Type the student ID in the space above		
Name	DAVID VALENZUELA		
Program	Industrial Engineering		
Curricula	2007-1		
First semester	20182		
Total semesters	9		
Skipped semesters	0		
Credits	93%		
GPA	83.47		
	Credits	**Required**	
Compulsory	263	266	
Elective	55	66	
Professional practice	0	10	
Estimated time to graduation and status	10 semesters	**ACTIVE**	
Advisor	DIANA JIMENEZ		

Semester	Courses	Completed courses	Courses completed in supplementary exam	Courses completed after third opportunity	Transfer courses and special exams	Credits earned	Total credits	Status at beginning of semester
2018-2	6	6	0	0	0	42	42	First semester
2018-5	-	-	-	-	-	-		
2019-1	6	5	1	0	0	49	91	Risk
2019-4	-	-	-	-	-	-		
2019-2 cc	5	5	0	0	0	40	131	Normal
2019-5	-	-	-	-	-	-		
2020-1	5	5	0	0	0	32	163	Normal
2020-4	-	-	-	-	-	-		
2020-2	5	5	0	0	0	34	197	Risk
2020-5	-	-	-	-	-	-		
2021-1	7	6	1	0	0	42	239	Risk
2021-4	-	-	-	-	-	-		
2021-2	6	5	0	0	0	24	263	Risk
2021-5	1	1	0	0	0	4	267	
2022-1	7	7	0	0	0	30	297	Risk
2022-4	-	-	-	-	-	-		
2022-2	7	7	0	0	0	21	318	Risk
2022-5	-	-	-	-	-	-		

Current semester courses: 2023-1 (6)
14532 - Workshop on professional practice
14639 - Entrepreneurship
14687 - Production systems
14690 - Business management
14700 - Strategic planning
14701 - Process enginnering

Fig. 12. Individual Advisement Semaphore

benefits. By thoroughly analyzing academic history records, we created a reliable data model. This model helps us create datasets, extract information, and match the needs of various processes that depend on student data. Having a strong link between data structure and knowledge rules helps in incorporating new rules or variations into the process.

Using spreadsheets to display data and statistics with customized visualizations served the purpose of accessing these data without the need of an Internet connection, unlike Web-based systems. This decision was important for us to keep using the resources during the Covid-19 pandemic, since our university's systems and servers have limited access outside of our network. The files were distributed via Google Drive, so people could download them for offline use. Access to the files was controlled using the platform's sharing features. Ensuring that the data is up-to-date and fully available, in relation to a recently concluded school period, turns out to be a critical factor in our strategy's success.

In technical terms, the next step in our approach is to transition to a more manageable system. Just as the initial deployment got to a point where updating the indicators and visualizations was too complex, it is expected that a similar situation will happen with the current reports and tools. There is however a vast amount of knowledge embedded in these reports and documented guides for creating and updating them that can be refactored into both functional and non-functional software requirements that drive the design and development of a more robust and consolidated system in the near future, built on the existing data structure.

As the authorities of the EC have shown interest in this approach, and other areas within the EC have also inquired about how to deal with data in their respective processes, it is important to identify the status of the data activities and projects. In [28], authors propose a maturity model for companies adopting big data strategies. Although our approach may not necessarily deal with big data, their assessment framework includes characteristics that can apply to our data strategies. Further evaluation is needed to properly assess the EC in their framework, as we currently match some descriptions of levels 2 and 3 of their 5-level maturity model.

ACKNOWLEDGMENT

Actions and results reported in this paper were performed at the College of Engineering of Universidad Autonoma de Baja California (UABC) as part of projects 105/6/C/25/4 and 105/2868.

REFERENCES

[1] P. J. Goldstein and R. N. Katz, *Academic analytics: The uses of management information and technology in higher education*. Educause, 2005, vol. 8, no. 1.

[2] A. Nguyen, L. Gardner, and D. Sheridan, "Data analytics in higher education: An integrated view," *Journal of Information Systems Education*, vol. 31, no. 1, p. 61, 2020.

[3] B. Daniel, "Big data and analytics in higher education: Opportunities and challenges," *British Journal of Educational Technology*, vol. 46, no. 5, pp. 904–920, 2015. [Online]. Available: https://bera-journals.onlinelibrary.wiley.com/doi/abs/10.1111/bjet.12230

[4] D. Dennehy, K. Conboy, J. Babu, J. Schneider, J. Handali, J. vom Brocke, B. Hoffmeister, and A. Stein, "Adopting learning analytics to inform postgraduate curriculum design," in *Re-imagining Diffusion and Adoption of Information Technology and Systems: A Continuing Conversation*, S. K. Sharma, Y. K. Dwivedi, B. Metri, and N. P. Rana, Eds. Cham: Springer International Publishing, 2020, pp. 218–230.

[5] A. Nambiar and D. Mundra, "An overview of data warehouse and data lake in modern enterprise data management," *Big Data and Cognitive Computing*, vol. 6, no. 4, 2022. [Online]. Available: https://www.mdpi.com/2504-2289/6/4/132

[6] I. A. Najm, J. M. Dahr, A. K. Hamoud, A. S. Alasady, W. A. Awadh, M. B. Kamel, and A. M. Humadi, "OLAP mining with educational data mart to predict students' performance," *Informatica*, vol. 46, no. 5, 2022.

[7] M. Patel and D. B. Patel, "Progressive growth of ETL tools: A literature review of past to equip future," in *Rising Threats in Expert Applications and Solutions*, V. S. Rathore, N. Dey, V. Piuri, R. Babo, Z. Polkowski, and J. M. R. S. Tavares, Eds. Singapore: Springer Singapore, 2021, pp. 389–398.

[8] W. Inmon and D. Linstedt, "The operational data store," in *Data Architecture: a Primer for the Data Scientist*, W. Inmon and D. Linstedt, Eds. Boston: Morgan Kaufmann, 2015, pp. 121–126. [Online]. Available: https://doi.org/10.1016/B978-0-12-802044-9.00019-2

[9] L. W. Santoso and Yulia, "Data warehouse with big data technology for higher education," *Procedia Computer Science*, vol. 124, pp. 93–99, 2017, 4th Information Systems International Conference 2017, ISICO 2017, 6-8 November 2017, Bali, Indonesia. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050917329022

[10] R. P. Singh and K. Singh, "Design and research of data analysis system for student education improvement (case study: Student progression system in university)," in *2016 International Conference on Micro-Electronics and Telecommunication Engineering (ICMETE)*, 2016, pp. 508–512.

[11] P. Edastama, A. Dudhat, and G. Maulani, "Use of data warehouse and data mining for academic data : A case study at a national university," *International Journal of Cyber and IT Service Management*, vol. 1, no. 2, p. 206–215, Oct. 2021. [Online]. Available: https://www.iiast-journal.org/ijcitsm/index.php/IJCITSM/article/view/55

[12] J. Silva, L. Hernández, N. Varela, O. B. Pineda Lezama, J. T. Cabrera, B. R. L. León Castro, O. Redondo Bilbao, and L. Pérez Coronel, "Intelligent and distributed data warehouse for student's academic performance analysis," in *Advances in Neural Networks – ISNN 2019*, H. Lu, H. Tang, and Z. Wang, Eds. Cham: Springer International Publishing, 2019, pp. 190–199.

[13] A. A. Yulianto, "Extract transform load (ETL) process in distributed database academic data warehouse," *APTIKOM Journal on Computer Science and Information Technologies*, vol. 4, no. 2, pp. 61–68, 2019.

[14] D. Amo, P. Gómez, L. Hernández-Ibáñez, and D. Fonseca, "Educational warehouse: Modular, private and secure cloudable architecture system for educational data storage, analysis and access," *Applied Sciences*, vol. 11, no. 2, 2021. [Online]. Available: https://doi.org/10.3390/app11020806

[15] K. C. Dumont, *Understanding Academic Advising Practice: Academic Advisors' Perceived Impact of an Academic Analytics Tool on the Practice of Academic Advising*. Michigan State University, 2021.

[16] SA Bogle, and KM Black, "Classifiers for Predicting Undergraduate Computer Science Performance," Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering 2018, 4-6 July, 2018, London, U.K., pp228-231.

[17] V. Realinho, J. Machado, L. Baptista, and M. V. Martins, "Predicting student dropout and academic success," *Data*, vol. 7, no. 11, 2022. [Online]. Available: https://doi.org/1010.3390/data7110146

[18] C. Liu, H. Wang, and Z. Yuan, "A method for predicting the academic performances of college students based on education system data," *Mathematics*, vol. 10, no. 20, 2022. [Online]. Available: https://doi.org/10.3390/math10203737

[19] R. Wirth and J. Hipp, "CRISP-DM: Towards a standard process model for data mining," in *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, vol. 1. Manchester, 2000, pp. 29–39.

[20] M. Alavi and D. Leidner, "Knowledge management systems: issues, challenges, and benefits," *Communications of the Association for Information systems*, vol. 1, no. 1, p. 7, 1999.

[21] M. A. Astorga-Vargas, B. L. Flores-Rios, G. Licea-Sandoval, and F. F. Gonzalez-Navarro, "Explicit and tacit knowledge conversion effects, in software engineering undergraduate students," *Knowledge Management Research & Practice*, vol. 15, pp. 336–345, 2017.

[22] I. Nonaka, "A dynamic theory of organizational knowledge creation," *Organization Science*, vol. 5, no. 1, pp. 14–37, 1994. [Online]. Available: https://doi.org/10.1287/orsc.5.1.14

[23] L. N. Hau and F. Evangelista, "Acquiring tacit and explicit marketing knowledge from foreign partners in IJVs," *Journal of Business Research*, vol. 60, no. 11, pp. 1152–1165, 2007. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S014829630700118X

[24] J. E. Ibarra-Esquer, M. L. González-Ramírez, B. L. Flores-Rios, C. M. Curlango-Rosas, L. E. Arredondo-Acosta, J. A. Armenta-García, and

C. A. Bernal-Cira, “Student Career Tracking System Database - Base de Datos del Sistema de Seguimiento a la Trayectoria Estudiantil (STrEs),” INDAUTOR, México, 03-2021-040814112900-01, 2021.

[25] J. E. Ibarra-Esquer, C. M. Curlango-Rosas, M. L. González-Ramírez, J. P. García-Vázquez, G. E. Chávez-Valenzuela, L. E. Arredondo-Acosta, B. L. Flores-Rios, and L. R. Bravo-Ramírez, “Perspective of the School Progression of Female Students in Undergraduate Information Technology Programs - Perspectiva de la Trayectoria Escolar de Estudiantes Mujeres en Programas Educativos de Tecnologías de Información,” in *Trabajos Científicos en México. Tomo VI: Ingeniería*, M. E. Sánchez-Morales, G. V. Vázquez-García, A. Martínez-García, C. E. Solano-Sosa, and E. L. Ramos-Guerrero, Eds. Guanajuato, México: Centro de Investigaciones en Óptica, A.C., 2022, pp. 164–175. [Online]. Available: https://cio.mx/archivos/trabajos_cientificos_mexico_2021/tomo_6.pdf

[26] J. E. Ibarra-Esquer, M. L. González-Ramírez, W. E. Aguilar-Salinas, M. Angulo-Bernal, D. Hernández-Balbuena, A. Mungaray-Moctezuma, and J. A. Suástegui-Macías, “Semaphore for the classification and visualization of academic progress for monitoring and tutoring processes of undergraduate students - Semáforo para la clasificación y visualización de avance escolar para procesos de seguimiento y tutoría de estudiantes de licenciatura,” INDAUTOR, México, 03-2019-082611262200-01, 2019.

[27] J. E. Ibarra-Esquer, A. D. Martínez-Molina, M. L. González-Ramírez, M. Angulo-Bernal, D. M. Álvarez Sandez, L. E. Arredondo-Acosta, C. M. Curlango-Rosas, A. C. Justo-López, and G. E. Chávez-Valenzuela, “Individual Advisement Semaphore (STrEs-IAS) - Semáforo de Tutorías Individual (STrEs-SETI),” INDAUTOR, México, 03-2021-080912170400-01, 2021.

[28] Soukaina Mouhib, Houda Anoun, Mohammed Ridouani, and Larbi Hassouni, ”Global Big Data Maturity Model and its Corresponding Assessment Framework Results,” IAENG International Journal of Applied Mathematics, vol. 53, no.1, pp393-404, 2023.

Jorge E. Ibarra-Esquer (M’12) holds a MSc degree in Computer Science from the Center for Scientific Research and Higher Education at Ensenada (CICESE), Mexico, obtained in 2001, and a Doctorate in Computer Science from the Universidad Autónoma de Baja California (UABC), Mexico, obtained in 2019. Since 2002, he has worked as a professor and researcher at the College of Engineering of UABC where he is currently in charge of student progression analysis. His research interests include Data Mining and Science applied to academic processes, taking part in research projects related to improving the learning processes for students in the Computer Engineering major. He is a Level 1 National Researcher of Excellence (SNI), a member of the Mexican Academy of Computation (AMEXCOMP), an IAENG member, and an IEEE member.

Brenda L. Flores-Rios is an Engineer in Computational Systems from the Technological Institute of La Paz, Mexico. She obtained her master’s degree in Computer Science from the Center for Scientific Research and Higher Education of Ensenada (CICESE) and her Doctorate in Sciences from the Universidad Autónoma de Baja California (UABC), Mexico. She is currently the director of the Software Engineering Laboratory at the Instituto de Ingeniería of UABC. Her research interests are Software Process Improvement and quality in small companies and the impact of Knowledge Engineering on Software Engineering. She is a Level 1 National Researcher of Excellence (SNI 1), and member of the Mexican Academy of Computation (AMEXCOMP). She is also a member of the Mexican Thematic Network of Software Engineering (REDMIS).

Maria A. Astorga-Vargas got her M.S. and Ph.D. in Computer Science from the Universidad Autónoma de Baja California (UABC). She is currently a professor at the Computer Systems program from UABC. Her research interests are Software Engineering with a focus in Software Process Improvement in small companies, and the effectiveness in software development teams. She is a Level 1 National Researcher of Excellence (SNI 1) by the National Council of Science and Technology (CONACYT) and a member of the Mexican Thematic Network of Software Engineering (REDMIS), and a member of the Mexican Academy of Computation (AMEXCOMP). She has participated as a consultant in the implementation of Software Process Improvement initiatives and an Appraisal Team Member in CMMI-DEV SCAMPI A levels 2 and 3.

Maria L. González-Ramírez is a Professor of Algorithms and data structures and Mobile devices programming at College of Engineering of the Universidad Autónoma de Baja California (UABC) in Mexicali. She is a PhD Candidate in Computer Science; her research interests are in information technology for education, and the use of technology in the caring of mental health. She is currently head of the Computer Engineering major.

Araceli C. Justo-López has a PhD in Engineering from Universidad Autónoma de Baja California (UABC). She is currently a Full Professor and Director of the College of Engineering of UABC. Her research interest is focused on the use and development of educational technologies in Engineering.

Gloria E. Chávez-Valenzuela received her master’s degree in network technologies and informatics at Centro de Enseñanza Técnica y Superior. She is currently a Full Professor at the College of Engineering of Universidad Autónoma de Baja California. Her research interest includes databases and assessment of educational outcomes at undergraduate levels.