Building an Enterprise Data Warehouse (EDW) for Colleges and Universities of Higher Education

Prepared by
Lou Galli
Chuck Mears
Tammy Cowden
Quantum Group
quantumgllc.com

INTRODUCTION

Our Quantum Team has over 100 years of experience in Pharmaceutical Commercial Operations sector. Two years ago, we were approached with an interesting opportunity that we could not turn down.

A Medical School in Atlanta reached out to us and asked us to design and build them an Enterprise Data Warehouse (EDW). The purpose of the warehouse was to consume data from over 75 various applications that different university business groups currently utilize. This data would be sourced from 90% external SaaS applications and 10% internal/home grown sources. The business units that would be the end consumers of the EDW worked in many different groups, including the Registrar's Office, Finance, Admissions, Educational Advancement, Clinical, Research, Community and Educational Research. The EDW would also be used for state and federal reporting purposes.

Our group was eager to take on this challenge to design, develop and deploy what we believed to be the optimal solution for the university. Like with all projects, the roadmap, timeline and cost were of the utmost importance in making the project a success.

REFERENCE ARCHITECTURE

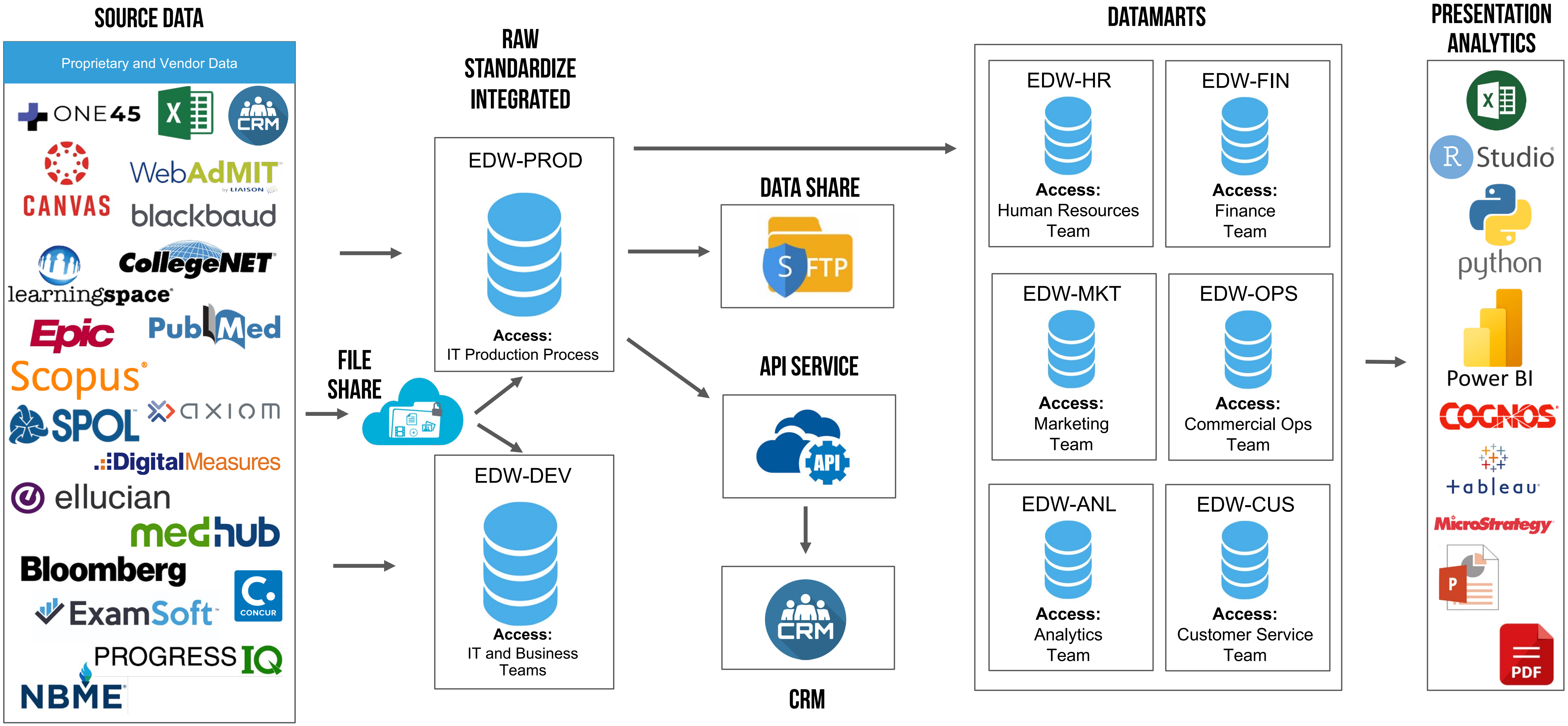
When you work in IT, you constantly feel the pressure to push forward and use the newest technology. “Shiny Penny Syndrome” is something we must constantly be weary of falling victim to. When you work for a Tech company, your business is to be cutting edge and write software and processes that push your company forward. But the vast majority of IT professionals do not work for Tech companies. We work in the Finance, Insurance, Education, Pharmaceutical, Manufacturing and Hospitality industries. We are facilitators. We are tasked with building simple, manageable and cost-effective solutions for business units. Going over budget and lengthening timelines to deploy the latest and greatest is not in the best interest of the business if the added time and cost does not offer added value.

TECH STACK EXAMPLE

It is highly recommended that groups standardize around a specific cloud provider. Although some companies use multiple solutions, this can be ineffective due to cost, integration and full time/contract resources. Below is an example of an Azure centric tech stack.

- Cloud Provider: Microsoft Azure
- RDMS : Azure SQL Server
- ETL/Pipeline : Azure Data Factory
- Files Share/Storage: Azure Storage
- Business Intelligence: PowerBI
- Application Development: .NET
- Datafile Types: XLSX / CSV / XML / JSON

REFERENCE ARCHITECTURE – CLOUD AGNOSTIC



SOURCE DATA

Proprietary and Vendor Data

FILE SHARE

RAW
STANDARDIZE
INTEGRATED

EDW-PROD

Access:
IT Production Process

EDW-DEV

Access:
IT and Business Teams

DATA SHARE

API SERVICE

CRM

DATAMARTS

EDW-HR

Access:
Human Resources Team

EDW-FIN

Access:
Finance Team

EDW-MKT

Access:
Marketing Team

EDW-OPS

Access:
Commercial Ops Team

EDW-ANL

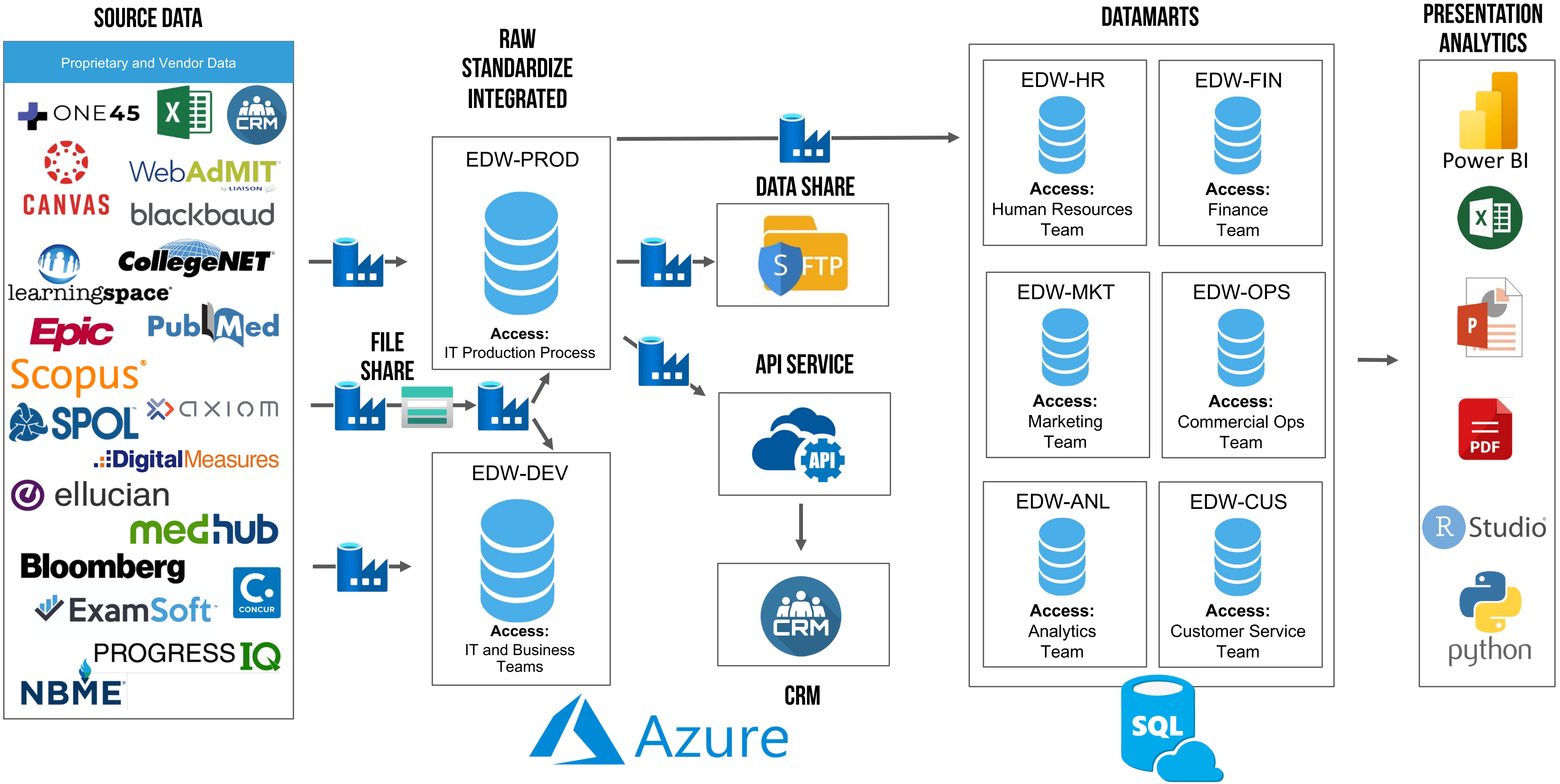
Access:
Analytics Team

EDW-CUS

Access:
Customer Service Team

PRESENTATION ANALYTICS

REFERENCE ARCHITECTURE – MICROSOFT AZURE



SOURCE DATA

Proprietary and Vendor Data

RAW
STANDARDIZE
INTEGRATED

EDW-PROD

Access:
IT Production Process

FILE SHARE

EDW-DEV

Access:
IT and Business Teams

DATA SHARE

API SERVICE

CRM

DATAMARTS

EDW-HR

Access:
Human Resources Team

EDW-FIN

Access:
Finance Team

EDW-MKT

Access:
Marketing Team

EDW-OPS

Access:
Commercial Ops Team

EDW-ANL

Access:
Analytics Team

EDW-CUS

Access:
Customer Service Team

PRESENTATION ANALYTICS

Power BI

DATA INGESTION

Ingestion Process - Data Sources should be ingested into the landing area of your EDW in the native data format. Often, these sources are dirty and denormalized. Regardless of the state of the data, transformations should not be done at this step

Ingestion Methods - How are you going to receive the data from the vendors and internal customers? API, DB to DB linked service, Manual ETL, email, file watcher, etc. Many times, this is beyond our control, but we should always seek out the method which leads to the most productionized solution to reduce issues with your production data refreshes and maintenance

Cleanse - Requirements sessions will determine the type of cleanse and standardization rules required. Data profiling is extremely important at this stage and IT should not make assumptions as to what should be done. We usually lack the subject area expertise to understand how the data is collected and used for deliverables

DATA INGESTION (CONTINUED)

Normalization - Data sources may be received in a denormalized state. Requirements gathering will determine whether your ETL process should normalize the data or if you should go back to the vendor to ask for different cuts of data. As an example, we consumed data from a publication vendor that sent us a single denormalized view that did not meet the needs of the consumer. The vendor joined articles, authors and author types into a single record. We were able to have the vendor build normalized and deduped extracts for Articles, Authors and Author types. We loaded the data and then created our own Report Ready master view based on the requirements from our customer. There are many examples where you can work directly with vendors to meet the needs of your end user.

Integration - What data sources need to be integrated or linked? Are you trying to link admission records to the grading system? Do you want to see how your matriculated students are doing with the patient logs and duty hours? All these questions need to be asked so you can build your XREF tables to view data across sources.

Agnostic Design - Datamarts should not be designed for consumption by a specific tool or device. If those tools change in the future, your marts will need to be redesigned. Even if you are creating mobile apps for iOS or Android, you can create an API service that both device applications can consume. Always strive to be device and application agnostic

DATA STORAGE LAYERS

Raw/Landing - Data is first pulled into this layer as is. There are no data transformations and all attributes are ingested without data types, cleanse of structure being applied

Cleanse/Standardize - Data is assigned data types and cleansed. Standardization rules are applied for attributes such as name, address, phone, email etc.

Integrated/Master - Data is matched and merged to avoid redundant dimension data. Customers, accounts, facilities are deduped

DATA STORAGE LAYERS (CONTINUED)

Integrated - Sources are integrated together using XREF/Crosswalk tables to join across various dimension and fact data

Canonical - Providing a narrow focus of the data needs for specific domains and groups. Data in this layer can bae normalized and denormalized

Datamarts / Data Stores - Cuts of data that are created for specific customers and teams. These teams will use Dashboards, Spreadsheets and BI Tools to extract the data from these Datamarts to perform analysis. General and specific reporting will also be run out of these marts.

DATAMART TYPES

For all industries, datamart creation is based on the specific needs of the entities. Higher Education is no different. We have identified eight specific marts that are required for these universities, but others may be required based on the need of the school.

HR Datamart - Needed for the Human Resources Department. This data will be highly sensitive in nature and privacy should be of the utmost importance. Faculty and Student attributes needs to be masked for most users

Finance Datamart - Needed for the Finance and Financial Aid teams. Thus data is also sensitive since it contains details pertaining to salary, government aid and scholarship details

Facilities Datamart - Required for facilities management. Building assignments, classroom utilization, parking lot and parking pass assignments, utilities tracking, maintenance of the school grounds.

DATAMART TYPES (CONTINUED)

Outreach Datamart - Alumni outreach, donations and grants tracking, community contact, email marketing.

Analytics Datamart - Analysis, dashboard and reporting to school leadership, required State and Federal reporting, predictive modeling.

Research Datamart - Tracking all research that is being performed at the school, status of research, analysis and details of research related finding, research reporting to state and federal governments

Student Datamart - Tracking student achievement through quiz, test and exam grades, student retention rates, at-risk student indicators, student field duty hours.

Faculty Datamart - Tracking the success of the staff and faculty based on student success, surveys and compensation dashboards.

UNIQUE DATA SOURCES

Every industry has unique data sources that can make ingestion into an EDW challenging. Higher Education is no different. There are four widely used applications in the Medical Education space that I want to detail. Each provide their own challenges.

- **Banner** (Ellucian) - Enterprise Resource Planning
- **Canvas** (Instructure) - Course and Grade Management
- **CollegeNet** - Admissions and Interview Scheduling Management
- **One45** (Altus) - Field Activity and Assessment Management
- **WebAdMIT** (Liason) - Application Management

UNIQUE DATA SOURCES (CONTINUED)

Banner - This ERP is used by many universities, not just medical schools. The UI is a bit cumbersome and the backend is not easy to access via API. It appears they developed their data model many years ago and have not updated it. Their naming conventions of their schemas and tables is very cryptic. Since we had access to their Oracle back end, we created a linked service within Azure to consume the data into our staging layer.

Canvas - This tool allows you to access some of its data through the API. Other data within their UI is not so easily accessible. The GradeBooks for example is only available as a CSV download and a manual process needs to be utilized to extract data elements that are required.

CollegeNet - This SaaS application is one of the standard bearers for application management. Their data is available via API. In addition to application management, they also offer interview scheduling services. The one drawback is that there is no linkage to the ERP system once an applicant matriculates. An XREF table would have to be generated OR the application management business team could append a students ERP ID to a miscellaneous field

UNIQUE DATA SOURCES (CONTINUED)

One45 - There is an API to consume Professor Assessments which is fairly simple to use. The interface to pull the Resident Patient Logs and Duty Hours is quite problematic however. You either have to manually pull the data through their UI or use a web scraping procedure (Python/Beautiful Soup) to get the data you need. Even after pulling the data, we still needed to write an app (We used VBA/Excel) to combine the data into a tabular format to be loaded to the EDW.

WebAdMIT - This SaaS application is one of the standard bearers for application management. Their data is available via API. Like CollegeNet, there is no linkage to the ERP system once an applicant matriculates. An XREF table would have to be generated OR the application management business team could append a students ERP ID to a miscellaneous field

DATA CATEGORIES / MISSION AREAS / PII

Schools of Higher Education will have various categories of data. This data will pertain to specific functional areas including Administration, Finance, Marketing, Alumni Outreach, Human Resources, Student/Registrar, etc. Each of these functional areas has unique use cases for their data and often times will desire analytics from the same data sources to be cut in various ways. When architecting your data repository, you must keep these factors in mind. Including Personal Identifiable Information (PII) or sensitive information that is not necessary for specific Mission Areaa should be avoided.

As an example, the Student/Registrar (Education) Mission Area will need analytics and reports based on students, professors, assignments and grades. The Human Resources and Finance teams will also need this information. Sensitive data pertaining to professor's salaries or addresses may be required by Finance and HR, but the Student/Registrar Datamart should not be loaded with this sensitive data.

THANK YOU