Academic Analytics Landscape at the University of Phoenix

Mike Sharkey, Director of Academic Analytics, University of Phoenix

Abstract. The University of Phoenix understands that in order to serve its large population of non-traditional students, it needs to rely on data. We have created a strong foundation with an integrated data repository that connects data from all parts of the organization. With this repository in place, we can now undertake a variety of analytics projects. One such project is an attempt to predict a student’s persistence in their program using available data indicators such as schedule, grades, content usage, and demographics.

Keywords: Learning analytics, academic data, integrated data, data modeling, predictive analytics, Hadoop

1 Introduction to the University of Phoenix

The University of Phoenix is a regionally accredited degree-granting institution founded in 1976 by Dr. John Sperling. Based in Phoenix, Arizona and with over 200 locations throughout the United States and a strong online campus, the University of Phoenix is the largest private university in North America. As of May 2010, over 470,000 students were enrolled at the University of Phoenix.

1.1 History

In 1976, Dr. John Sperling, a Cambridge-educated economic historian and professor, founded University of Phoenix on an innovative idea: making higher education accessible for working adults.

In the early 1970s, while a tenured professor at San Jose State University in California, Dr. Sperling and several associates conducted field-based research on new teaching and learning systems for working adult students. From this research, Dr. Sperling realized that the convergence of technological, economic, and demographic forces would herald the return of working adults to higher education. He saw a growing need for institutions that are sensitive to the learning requirements, life situations, and responsibilities of working adults. These beliefs resulted in the creation of University of Phoenix.¹

¹ Taken verbatim from our institution’s website at: http://www.phoenix.edu/about_us/media_relations/just-the-facts.html. For a detailed history of Dr. Sperling and the University of Phoenix, refer to the book Rebel With a Cause (John Wiley & Sons, 2000).
1.2 Students Served

In the past, the University of Phoenix focused on degree completion for non-traditional adult learners. Over the years, changing demographics have seen students with little to no credits starting at the university with the goal of completing their entire programs. The university also expanded in 2006 when it started offering associate’s degrees in addition to bachelor’s, master’s, and doctoral degrees.

The demographics of the non-traditional learners at the University of Phoenix differ from students at traditional post-secondary institutions. Non-traditional students tend to be older, largely female, and tend to come from more diverse socioeconomic and racial backgrounds. Many non-traditional learners also have jobs and family obligations as opposed to 18-22 year-old residential students at a traditional institution. This different demographic changes the motivators and drivers behind the students’ actions.

1.3 Structure

Key aspects of the university’s organizational structure differentiate it from traditional degree-granting institutions and community colleges. These aspects have a significant effect on determining the direction of technological initiatives.

Academically, the university has a central administrative unit. All programmatic and curricular decisions are made by the central Academic Affairs unit and then distributed throughout each of the campus locations. This process holds true for the creation of new programs, the modification of existing programs, and the development of course curricula.

Courses are taught by practitioner faculty; the faculty members have experience working in the field that is related to their course. Curriculum is centrally designed by faculty members, content experts, and instructional designers. University of Phoenix faculty members have the academic freedom to enhance the standard curriculum with their expertise, content and theoretical knowledge, and the practical experience they gain as a result of working in the fields in which they teach.

Each campus is an independent entity and receives operational support from the centralized organization (curriculum, marketing guidelines, operational requirements).

Over two thirds of the students take courses online while the remaining students attend in a face-to-face modality. Because many students take one hundred percent of their courses online, there is a campus dedicated solely to online students. This is the largest of the University of Phoenix campuses and it is also based in Phoenix, Arizona. Students enrolled at physical campuses can choose to take their courses in local classrooms or they can attend online classes as long as the programmatic requirements are met.

One last point is that the university’s academic calendar does not follow a semester, trimester, or quarter model. For the most part, courses are taken serially throughout a student’s entire program. At the associate level, students take two complementary nine-week courses concurrently. At the undergraduate and graduate levels, students take one five-week or six-week course in serial (respectively). A
student may continue to take one course after another until the program is complete, or a student might choose to take time off between courses if there are scheduling conflicts due to personal or work issues.

1.4 Unique Needs of the Institution

Before diving into the details of how data are utilized by the University of Phoenix, it is important to emphasize how the structure of the organization lends itself toward a set of needs that differ from traditional universities.

One common thread throughout the institution is scale. While the university maintains a class size in the mid-teens, there are hundreds of courses in progress affecting thousands of students on any given day. Students have scheduling flexibility because there is a good chance that the desired course is starting on any week. We need to think in terms of tens of thousands for course design, content, facilitation, tutoring, licensing, data collection, or any other variable function (or terabytes when it comes to data).

Another aspect of our institution is operational efficiency. The sheer size of the business combined with the number of campuses and the centralized administration means services must be simple and as efficient as possible. This is common when it comes to areas such as applications and registrar functions. However, it must also be viewed from an academic perspective. A good example is in the licensing of software. Some academic software products require an instructor to upload a course roster to allow students to log into the product. Due to the flexibility of University of Phoenix scheduling, course rosters might, and frequently do, change right up to the time class starts. Instructors do not usually have the roster information in advance. In this case, administrators work with the content partners to provide a single sign-on feature that allows students to automatically login from our learning management system to the vendor system.

2 Our Approach to Data

As an institution, we are very aware of the fact that data will make or break our ability to educate our students effectively in the future. Although we have the advantage of a proprietary online learning system, we realize that we have not come close to tapping the potential of the data stored in our systems.

To that end, we have spent a significant amount of time and effort over the past few years to ensure that data has its place in the foundation of the organization.

2.1 The Move Toward a Data Driven Culture

While it might sound trite, it is vital that any change starts with the people who make up the organization. We started a concerted effort to stress the importance of data at every rung of the organizational ladder.
One basic step we took involved messaging. After a restructuring of the product and engineering groups in 2009, our new management focused on three areas of performance:

- Site Up
- Data
- User Experience

The fact that data was one of the three focal areas of the group was a testament to our commitment to a data-driven culture. We followed up on the messaging with key hires in the data arena. This included bolstering our technical capacity and bringing on board staff with experience in analytics, cognitive science, and data-driven consumer products businesses. The investment in staff who can move, align, and interpret data is something that will pay dividends in the future.

### 2.2 Applications of the Data

Data are different from information. Data are atomic units; they set the foundation for capabilities that can have a deep impact on learning and business. We ask ourselves, “What can we do with the data once this foundation is set?” The following diagram illustrates our answer to that question:

![Pyramid showing the different applications of our data foundation](image)

Reporting and business intelligence is the base of the pyramid. Although commonplace, we do not want to underestimate the impact of basic business intelligence. A good, integrated data structure can provide simple answers to many questions. Predictive analytics goes to the next level. It allows us to answer the tougher business questions and use data to look ahead. Data-driven learning is where
we apply the data not only to business/operational questions but to the core activity of our institution - learning. These areas will be discussed in more detail in Section 4.

2.3 Data as a Strategic Advantage

The University of Phoenix is in a unique position as compared to traditional universities and colleges. Because we are a for-profit entity, we need to address business and financial implications in addition to the implications of learning.

One of the largest advantages we have in the higher education space is our size. With over 400,000 students, we have the ability to use data and analytics that would produce accurate and reproducible results. We are not limited to testing a new learning tool on a class of 25 students. We can test with hundreds or thousands of students, so long as the trials do not negatively impact the students’ ability to learn. From a data standpoint, that means that we have more than enough data points to support the efficacy of the tool.

3 Topology

It should not be surprising that we have data strewn in different databases across the entire student lifecycle. This scattering of data reduces the efficacy of our analytic capabilities. To combat that, we set up a replication system where all data flows into a single integrated repository (see Figure 2).

Fig. 2. The flow of data to our integrated data repository
3.1 Source Systems

The first step in the workflow is to replicate all source systems. We use a commercial replication tool called Golden Gate to accomplish this (Golden Gate was acquired by Oracle in 2009). Golden Gate is used on any of the source systems we want to move to the integrated repository – these include both Oracle and SQL databases.

One of the benefits of the replication is that we alleviate the problem of destructive data. Normally, if a field in one of the source systems is overwritten, we lose the older data forever. With replication, we effective-date the tables so that any overwrites are saved. This helps with older systems that inadvertently destroy data due to a poor/obtunded design.

The table below is a listing of some of the source applications that we replicate to our staging systems. It is not an exhaustive list.

Table 1. A partial listing of source data systems replicated to a single staging area.

<table>
<thead>
<tr>
<th>Source system</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aptinet</td>
<td>Lead/marketing data</td>
</tr>
<tr>
<td>Apply web</td>
<td>Student’s application to the institution</td>
</tr>
<tr>
<td>Osiris</td>
<td>Students, courses, schedules, course grades (SIS)</td>
</tr>
<tr>
<td>Galaxy</td>
<td>Student contact data (CRM)</td>
</tr>
<tr>
<td>Gradebook</td>
<td>Assignment grades</td>
</tr>
<tr>
<td>OLS</td>
<td>Online Learning System (discussion forums)</td>
</tr>
<tr>
<td>Application Log</td>
<td>Content usage tracking</td>
</tr>
<tr>
<td>HEAT</td>
<td>Tech support issue tracking system</td>
</tr>
</tbody>
</table>

3.2 Integrated Data Repository

The key to our data foundation is the integrated data repository (IDR). The IDR is a unified, normalized data structure of all data elements across the entire company. Replication copies each source table to a staging area. The next step is to travel from staging to the IDR. To do this, we needed to rationalize every field that we moved over. As an example, we looked at the student’s home zip code. We may have collected that zip code when the student first contacted the university, we may have collected it on an application, and the campus may also have collected it in the course registration process. We now have three instances of the zip code in our systems and regardless of whether they are the same or not, a student should only have one current zip code on file. This is where the data modeling comes into play.

**Modeling.** We have a team of data modelers who work to create a normalized physical model of all of the data elements. In our zip code example, the first thing the modelers do is create the proper data schema. Instead of having a Marketing Person, Application Person, and Registrar Person table, we only have a single Person table. The next step is to determine which source table has the right zip code. We may determine that the zip code stored in the registrar’s database is the proper one to use, so we designate that field for transport.
**Extract, Transform, Load (ETL).** After the entire integrated schema is mapped out, the next step is to populate it. The ETL team writes jobs to move the data from staging into the proper place in the IDR. A significant part of the ETL process is quality control. As data are moved over, we check for inconsistencies and errors and do what we can to address them. The IDR is known as the *single source of truth*, so consistency and quality are vital characteristics.

**Data Marts.** One other facet of the architecture is a data mart. The IDR is large and normalized; this is not a good combination for efficient querying. In order to have a data structure that is built for fast querying of complex data, we need to de-normalize and index the data. There may be multiple data marts in existence. One may be a series of tables dealing with students' progression throughout their programs, while another may focus on financials and accounting.

### 3.3 Hadoop

Due to the size of our institution, we know that we would be running into issues with the sheer volume of data in some of the tables. For example, the discussion forums databases have a record for each post for every student and faculty in every class. If you multiply the posts by students (and faculty) by week and by course, we can see millions of records in a week or even a day. Mining tables of this size in an efficient manner calls for a different solution.

The solution we have adopted is an open source product called Hadoop. Hadoop was inspired by work at Google and extended through usage at Yahoo!, Facebook, and other prominent Internet companies. Hadoop addresses the problem of large datasets by using distributed parallel processing. A Hadoop cluster is made up of many commodity server nodes - the benefit is to use a large number of cheap servers instead of a small number of expensive ones. The University of Phoenix product group built a 40-node cluster in 2009, and we are continuing to develop its capabilities.

Hadoop is used to solve specific problems with our data. It is most applicable in two cases. First, it helps to digest large datasets. Whether it is the discussion forum tables or raw web usage logs, Hadoop can process the data, create summarized tables, and send the summaries back to the IDR. Second, Hadoop can help with detailed data analysis on non-fielded data. The actual discussion forum posting is a block of text. In a traditional database, that block of text is lumped into one field and that makes it hard to mine the text beyond the use of simple query statements or regular expressions. With Hadoop, we can run cycles of queries or code against the non-fielded data, continually reducing the problem into smaller chunks. When we have derived the information we are looking for, we can write that summarized information back to the IDR for use with traditional queries.
3.4 Analytics Tools

The ‘last mile’ of analytics includes any data reporting, analysis, or visualization tool used to turn data into information. Following the mantra of ‘using the right tool for the right job’, there are a number of tools being used within our institution. A tool like Microsoft Excel is always available as a failsafe, but we rely on other products for more specialized needs.

**Tableau.** Tableau is a commercial data visualization tool whose strength lies in its ability to figuratively paint many different kinds of pictures. Unlike traditional visualization tools where one might start by selecting the desired type of visualization (e.g. scatter plot, bar graph), Tableau lets the user start by just adding measures and dimensions to a palette. As the user adds fields, one can either try different visualizations to see how it looks or use suggested types from Tableau. The product helps the user paint the picture that will tell the desired story in the best way possible.

**R.** If the need is to perform statistical calculations or correlations, R is the right tool for the job. It is a powerful open source software product that can complete a myriad of statistical functions.

**PL/SQL.** Many times, the need is to simply explore the data in order to zero in on whatever answer, question, or anomaly one is looking for. Our IDR is an Oracle database and a simple SQL querying tool such as PL/SQL Developer will often be the right tool for the job.

4 Analytics Applications

The data foundation described in this overview is just that – a foundation. In and of itself, it has no value. One must apply the data towards an end goal such as answering a business question. The pyramid in Figure 1 shows three levels by which we can categorize the application of data across our organization.

4.1 Business Intelligence (BI)

The BI team works like many traditional reporting teams. The goal is to provide reporting services to the areas of the company where it is needed. The kinds of services provided depend on the needs and capabilities of the requestor. At its most basic level, we have the reporting tool. We use the commercial Business Objects tool to provide reporting to all areas of the business. The departments might author reports on their own or they may put in a request and have a central reporting team develop the report on their behalf.

Another variation on reporting is dashboards. Our development group can create simplified visual dashboards that answer a few key business questions in an easy-to-
understand manner. If reports are good for departments that need to make operational
decisions, dashboards might be better for high-level overviews of a business process.

4.2 Predictive Analytics

The analytics team at the University of Phoenix is set up to focus in on the more
difficult questions that cannot be easily answered with a single report. We just
changed the curriculum in a certain course…is it a change for the better? Is one
campus location doing a better/worse job than another in its ability to deliver
instruction? How many weeks does it take for an MBA student to get to their fifth
course? These are complex questions that require complex analysis.

There is more than just answering the question, though. We want to be able to use
data to predict future outcomes so we can stay ahead of the impending trends.
Predictive analytics can be used to predict different outcomes including student
learning, student success, marketing channel efficacy, or financial outcomes.

Student Persistence. The University of Phoenix is currently looking at one specific
predictive channel focused on persistence in a program. The approach is similar to an
actuarial table, but instead of predicting how long an insured person will live, we
want to look at how long a student will progress through their program. As an
institution serving non-traditional learners with competing factors like a job or a
family, we know that external factors can hinder a student’s ability to stay in the
program. We may not be able to avoid these factors, but if we see signs of them
coming, we can help the student handle the change in a more productive manner.

The goal with student persistence is to include as many factors as possible in a
correlation model. By analyzing past data, we might be able to determine what
factors indicate a high probability that the student is preparing to withdraw from the
program. The IDR contains static information such as demographics and active
information such as schedules and grades. It is our belief that some of these factors
will have a high correlation with a student’s intention to withdraw. Therefore, we
will be able to rate the withdrawal risk and give some sort of a persistence score.

The obvious next question is, “So what do we do with this information?” If we are
able to predict persistence/withdrawal with some level of accuracy, then we can
proactively help the students with their decision-making. All students have an
academic advisor who has the job of assisting the student throughout their program.
It is our intent to provide the advisors with up-to-date persistence scores so that the
advisors can intervene and help the student find the best course of action. We do not
know if the best course of action is remediation, taking a break in scheduling, or some
other solution. To that end, our intent is to focus on human intervention (with the
advisors) instead of some automated remediation path.

The student persistence analytics project started in October 2010 and we hope to
share results as the project matures.
4.3 Adaptive Learning Engine

The top level of the data application pyramid is an adaptive learning engine. This is a longer-term project aimed at the heart of our institution. Our goal as a university is to help students achieve the learning outcomes as specified by the program. A project such as student persistence might help to keep the student enrolled in the program, but it does not address the learning.

We have the desire to leverage all of the student data we have to help students traverse that optimal learning path. Through a combination of data analysis, learner profiling, and a learning platform that supports multiple paths to achieving the same outcome, we believe we can guide students down the path that best fits their individual needs as a learner.

5 Conclusion – the Current State of Analytics

It has taken the University of Phoenix many months to establish and populate the IDR, our foundation for analytics. As of this writing, the IDR is still not complete and new tables are continually being added. However, we are not waiting for it to be one hundred percent complete. There is enough data to initiate reporting and analytics projects that can both provide value to the company and set the stage for future research. We are fortunate to be able to dedicate multiple teams to different aspects of the analytics and we will continue to share outcomes with communities such as the Learning Analytics and Knowledge group as results become available.