NATIONAL STUDENT CLEARINGHOUSE CASE STUDY

Committed Partner. Creating Results.

The National Student Clearinghouse, or NSC, won a multi-million dollar grant from the Bill and Melinda Gates Foundation to improve the quality of longitudinal studies on high school students. The ultimate objective was to help high schools access NSC's vast data warehouse of student information and thereby reliably measure the outcomes of educational efforts.

NSC decided to use part of the grant to improve their data collection and warehousing system called Student Tracker for High Schools, or STHS. STHS is the central system that manages all the data used for longitudinal studies. According to NSC's projections, a surge of demand for high school request records would occur in tandem with the roll-out of this new system, reflecting an increase of nearly 2,000% from previous levels to over 47 million annual requests.

Faced with the need to support this much greater access to their data warehouse of over a billion records, and with tight deadlines to meet fall enrollment data requests, NSC hired CC Pace to help deliver the new system.
Several challenges faced the effort. Each challenge was successfully met by the CC Pace team.

Data Quality

Data entering the system were of varying quality. Often, input files were formatted incorrectly, missing data fields. Sometimes, data fields were present and in the right format, but not meaningful; for example, data entry mistakes showing high school students over 100 years old. Correcting these mistakes on the old system was a time-consuming task, requiring the intervention of internal IT staff to operate directly on copies of submitted files.

To address this issue, CC Pace designed and built a data intake workflow system. Colleges and high schools can drop their data files off at a secure FTP site. The system automatically ingests the files, processing several files in parallel if necessary, and runs batteries of complex validation rules to ensure data quality. Two kinds of validations run: syntactic validations ensure that the data formats are correct, and semantic validations ensure that the data make sense. NSC data intake specialists view the validation results at a web site that features innovative "Web 2.0" interactivity. If there are errors, the specialists fix them and re-submit the data. Once the data are clean, the specialists import them into the central NSC data warehouse.

Data that pass through this gauntlet of validations and corrections are much cleaner and more consistent than previously. In turn, research based on these data has lower margins of error.

Student Matching

NSC uses complex algorithms to identify students as they move through the various stages of the longitudinal study. For example, the algorithms must figure out whether Jane Smith who graduated from a high school in Idaho is the same Jane Smith who was admitted into a college in Iowa. These techniques collectively go by the name of student matching algorithms.

Research specialists need to constantly tune and tweak algorithms to improve both frequency and quality of matches. However, the legacy system used "hard wired" matching algorithms: the application code itself contained the various strategies used to assemble data from the warehouse and the rules applied to those data. Only skilled programmers could modify the rules, and the system had to go through a release process when these rules changed. The result was that the specialists' need for better algorithms couldn't be quickly met.

As the volume of data in the warehouse grew to over a billion records, the legacy system also faced performance problems. Matching took an ever increasing chunk of time as the algorithms had to sift through huge volumes of student data.

CC Pace built a matching engine capable of running complex student matching algorithms against the warehouse. The engine does not contain any matching algorithms within its code. Instead, it lets analysts design complex algorithms and, using simple web interfaces, load them into the engine as XML files. Multiple algorithms, each supporting a different research objective, can run simultaneously. Researchers can now modify their matching algorithms without any programmer intervention and while the system remains up and running.

We used our knowledge of parallel computing to achieve matching rates of several hundred students per minute. Matching that used to take many hours or days to run, now runs in minutes, allowing NSC to handle much larger data sets.

NSC needs this matching capability in more than one of their systems, so we built the engine as a service available to any system that needs it. The service has two interfaces: a simple web service interface for situations that need only single matching transactions, and a messaging bus interface for situations that need to process large batches of data in the background.

Reporting

For high schools and districts, the goal of the system is to see how their students progressed through the postsecondary school system. Upon entering and matching their student data, they are presented with several reporting features:

1. A standard PDF report packet containing various bar and line charts that represent student educational outcomes by attribute. For example, students who earned a degree or certification broken out by ethnicity.
2. A csv file of raw student data and their individual postsecondary status.
3. An ad-hoc reporting web interface that allows the user to select customized PDF and Excel reports based on various attributes and outcomes

We designed the reports using a commercially available report server and designer. This allowed us to build the reports envisioned by the NSC research department, including the addition of more student attributes, such as test scores, ethnicity, and English as a second language.

Since reports for thousands of participating high schools and districts were processed in parallel, both the database and report server had to be designed to scale. We accomplished this by offloading the report processing to dedicated report servers. This allowed NSC to scale horizontally by simply adding another report server node and registering that node with the application.

We also ran complex data aggregations and stored the results in the data warehouse. By performing the aggregations once for each participating organization, subsequent requests for reports using the ad-hoc tool ran much faster.

Respecting student privacy was a primary concern to NSC and they adhered to the FERPA law and its rules governing disclosure of student records. For example, if the number of students with a specific attribute is fewer than 10, the report will hide the outcome. Also, if a student requests a FERPA block of their postsecondary records, then it would be suppressed in the raw data report.

System Performance and Cost

The legacy system ran on IBM AS/400 midrange computers, using a variety of different software technologies, including Unix shell scripts and C language programs. It didn't use modern techniques of parallelism and multithreading, so the only way to increase performance was "vertically", by beefing up the AS/400 systems. This was getting very expensive, and the skills to maintain the software were getting harder to find.

NSC decided to address these problems by adopting Intel-powered servers running Linux as a new hardware platform, and Oracle's database and Java Enterprise Edition as its new software platform.

CC Pace used our experience with Java Enterprise Edition and Oracle tools to build the entire system on NSC's new platform. NSC can now run STHS on cheaper and faster hardware, using widely-supported mainstream technology. By designing the software to use multithreading, asynchronous queues, and other parallel computing techniques, we enabled "horizontal scaling", or adding more power by adding more servers to a cluster.

Today

NSC's business goals for STHS have been met. The improved intake workflow and modern user experience sharply reduces the time data analysts spend on cleaning up data. The high performance matching engine improves the quality of longitudinal studies and reduces the time that researchers spend resolving ambiguous matches. In peak season, data ingestion is many times faster than it used to be, freeing the organization to work on more research projects from the same data.

High schools also benefit from the new system's capabilities. They too get higher quality research made possible by the improved and more flexible matching algorithms. The reports they receive from the system are more visually appealing, but also contain a lot more information than previously. All of this allows the schools to concentrate on using the reported data to improve their own educational outcomes.