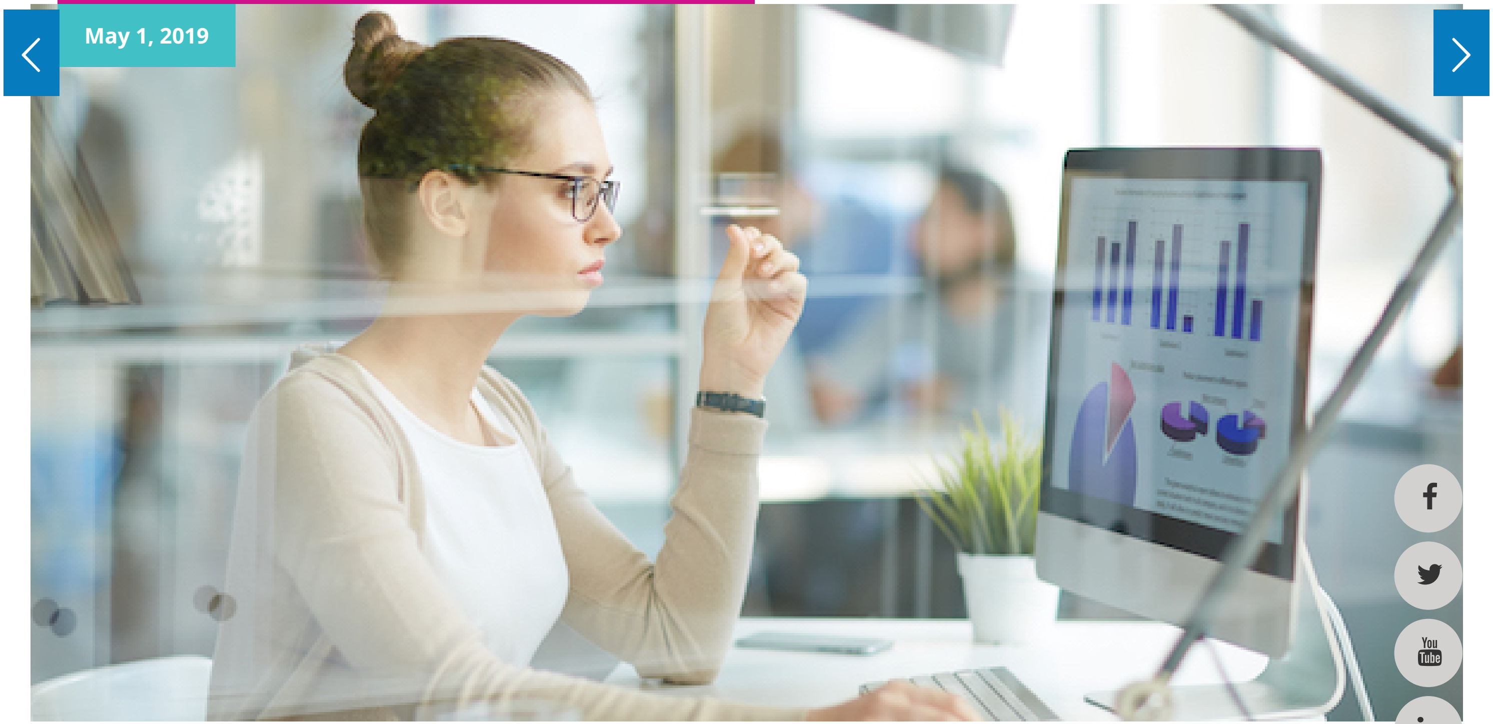
Challenges With Building an Educational Reporting Data Warehouse

By Alla Gorina In data, data warehouse, education

May 1, 2019



When designing data warehouses to fit particular business needs, similar challenges may present themselves. Sometimes, one tool or technology will not cover all requirements. In this blog, we will look at an educational test reporting system that we designed, the data challenges our delivery team faced, and how we chose to solve them.

Business Challenge

The client business domain includes collecting results of K-12 student testing. The business goal is to provide reporting and analysis tool for parents, teachers, and school/district/state administration.

Key data challenges

1. Collect and link the data between the test authoring and test delivery systems, and over several years of testing. Five years of data must be maintained.
2. Ensure data consistency and security, which is critical as students' data is protected by privacy laws.
3. Near real-time response time for a variety of report types. For example: a parent is only interested in their child, a teacher wants to view their their class's results, while a state administrator would like to have a multi-level aggregation report.

Let's look at each challenge in detail.

Collect and link the data between systems, over several years

The data is submitted as XML or JSON documents. A NoSQL document-based storage may seem like a good fit but we know we will have more complex reports involving document relation and aggregation. There are shiny new solutions that support ad-hoc queries against documents (Presto, Druid) but they don't provide other requirements like data updates, consistency, and security.

In our case, the number of XML or JSON data elements is limited and well defined. This means we can present them in the relational tables using an Online Transaction Processing (OLTP) Relational Databas anagement System (RDBMS).

OLTP databases are characterized and measured by the ability to effectively process a large number of short online transactions (INSERT, UPDATE, DELETE) and queries (SELECT) while maintaining data integrity in multi-access environments. The exam data flow fits into this category: the exam datasets are small and delivered one at a time. Due to the nature of the exams administration, the data flow is spikey. The majority of our data load will be INSERTs with the occasional UPDATE and DELETE. Our estimated data volume is about 300-500 million exams over five years. Any modern RDBMS is capable of handling this data flow and size.

There other bonus benefits that come with OLTP:

- Relational databases are well-documented and mature technologies. This means better tools, bet support, and more stable systems.
- OLTP DBs use Structured Query Language (SQL) for manipulating the data. SQL standards are wel defined and commonly accepted.
- Most cloud vendors support different RDBMSes. This makes it convenient and flexible to support cloud deployments without depending on a specific cloud provider.
- A large pool of qualified developers have experience with SQL and RDBMS.

An OLTP RDBMS seemed like a great fit for this challenge. Let's look at the next one.

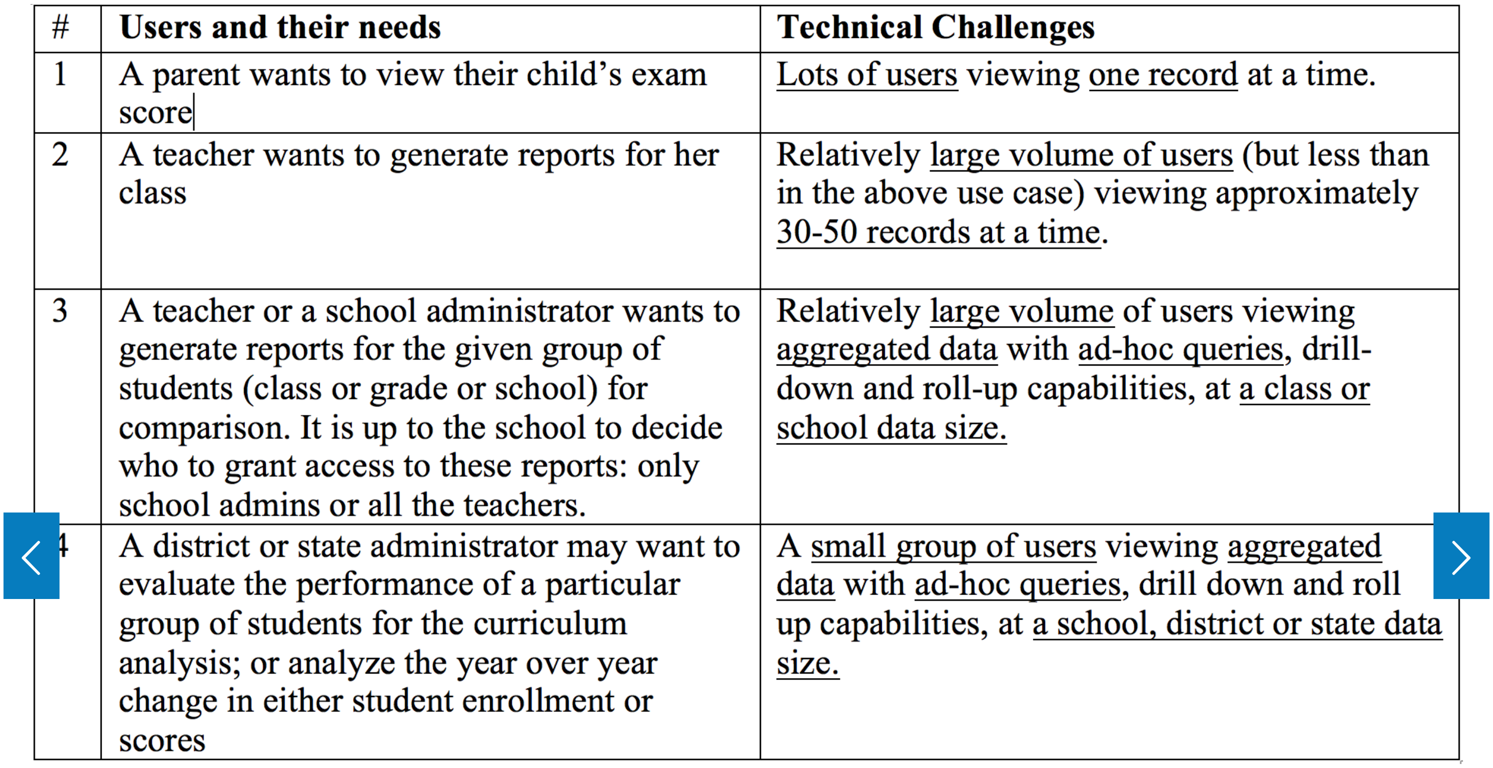
Data consistency and security requirement

Relational Database Management Systems (RDBMSes) address this requirement out of the box:

- OLTP RDBMSes are ACID compliant, which guarantees validity even in the event of errors, failures, etc.
- OLTP solutions have built-in sophisticated security support.
- OLTP vendors support data encryption.

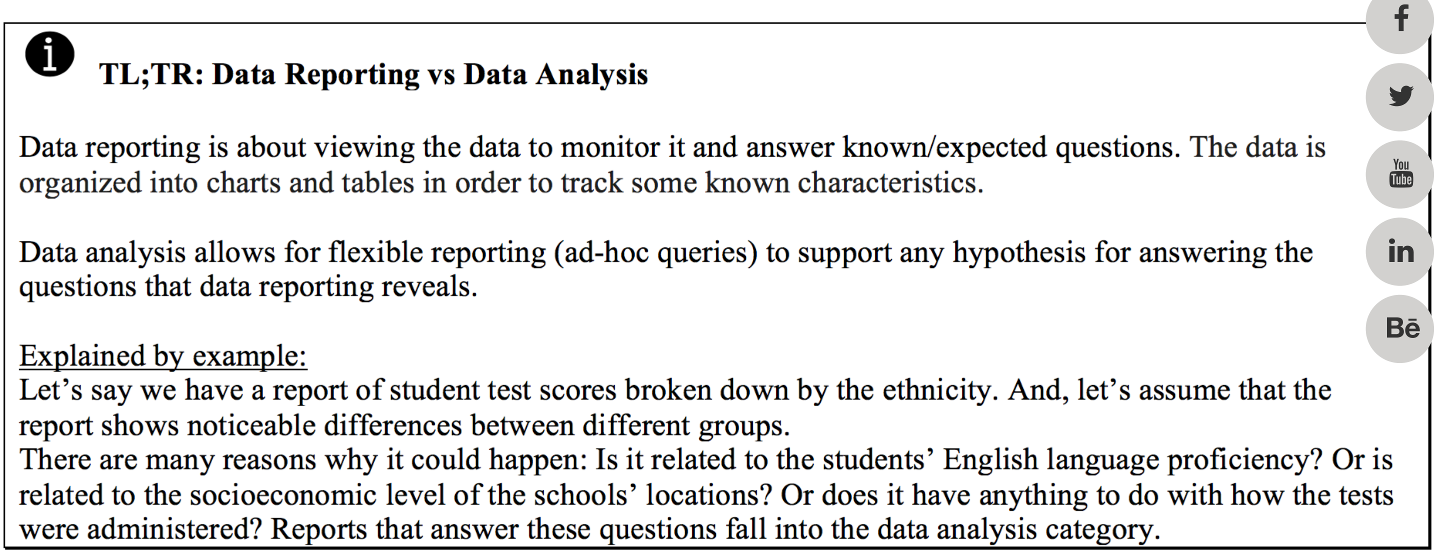
This leaves us with the last problem to solve:

Different Types of Users With Different Reporting Needs



#	Users and their needs	Technical Challenges
1	A parent wants to view their child's exam score	Lots of users viewing one record at a time.
2	A teacher wants to generate reports for her class	Relatively large volume of users (but less than in the above use case) viewing approximately 30-50 records at a time.
3	A teacher or a school administrator wants to generate reports for the given group of students (class or grade or school) for comparison. It is up to the school to decide who to grant access to these reports: only school admins or all the teachers.	Relatively large volume of users viewing aggregated data with ad-hoc queries, drill-down and roll-up capabilities, at a class or school data size.
4	A district or state administrator may want to evaluate the performance of a particular group of students for the curriculum analysis; or analyze the year over year change in either student enrollment or scores	A small group of users viewing aggregated data with ad-hoc queries, drill down and roll up capabilities, at a school, district or state data size.

The first two use cases are data reporting problems, while the other two fall into the category of data analysis. The difference is subtle but is important to understand.



TL;TR: Data Reporting vs Data Analysis

Data reporting is about viewing the data to monitor it and answer known/expected questions. The data is organized into charts and tables in order to track some known characteristics.

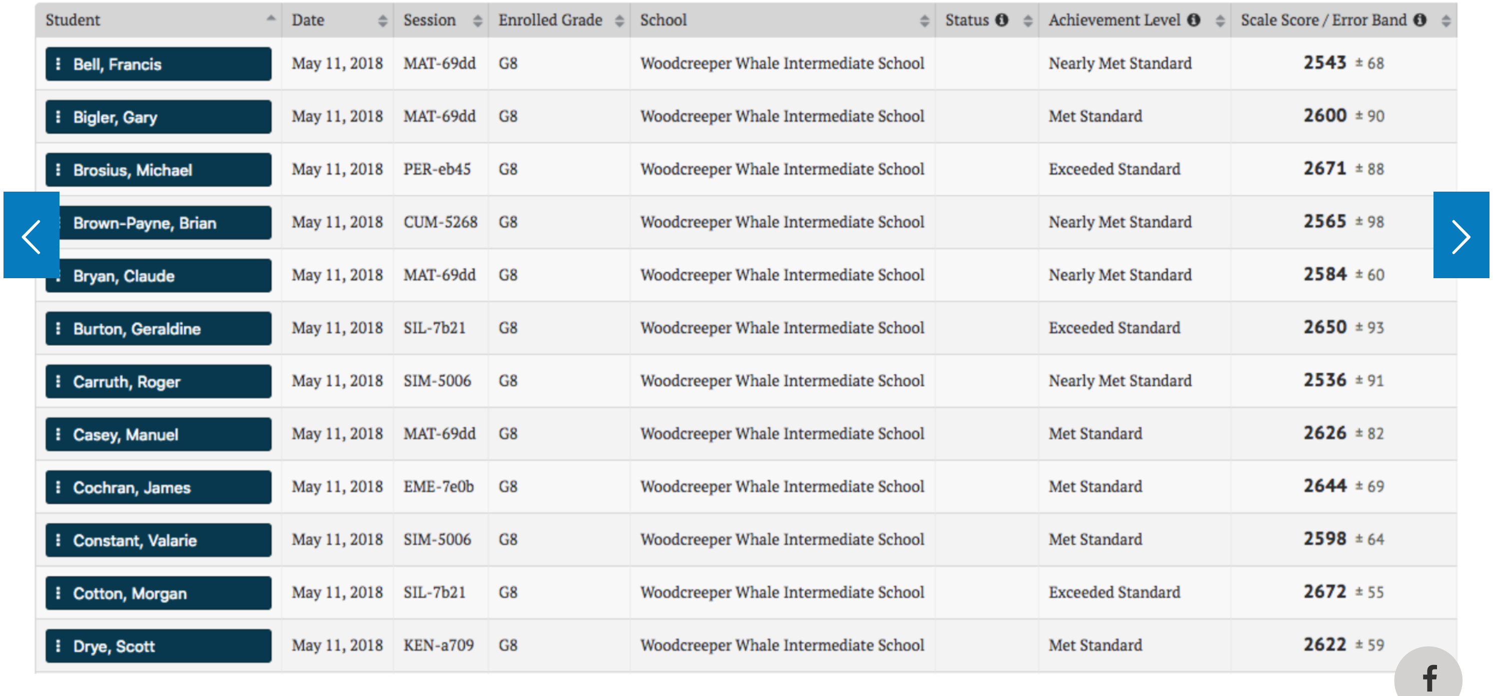
Data analysis allows for flexible reporting (ad-hoc queries) to support any hypothesis for answering the questions that data reporting reveals.

Explained by example:
Let's say we have a report of student test scores broken down by the ethnicity. And, let's assume that the report shows noticeable differences between different groups.
There are many reasons why it could happen: Is it related to the students' English language proficiency? Or is related to the socioeconomic level of the schools' locations? Or does it have anything to do with how the tests were administered? Reports that answer these questions fall into the data analysis category.

Data Reporting : Parents or Teachers viewing their students results

A parent or teacher viewing their student results fits right into OLTP's strength, but raises the concern of managing a large number of concurrent end users. Last time my daughter took the SAT exam, all the results were released to parents on the same date and hour. My attempt to view my child's report ended up with the message: "*Sorry, our system is experiencing a high volume of users. Please check later.*" How could we do better than the SAT site?

In general, OLTP RDBMSes are built with vertical scalability in mind (by adding resources within the same logical unit to increase capacity). With the number of expected users, this would be very expensive even if it's possible. Scaling horizontally (increase capacity by adding nodes to a system) is a more flexible, cheaper solution. Most relational DBs support sharding and replications, but this is not trivial to set up. However, with managed solutions, it has been significantly simplified. For example, setting up write/read replicas on Amazon Aurora is a straightforward task and allows for horizontal scalability. So far so good: OLTP is still a winner for solving this organization's challenge.

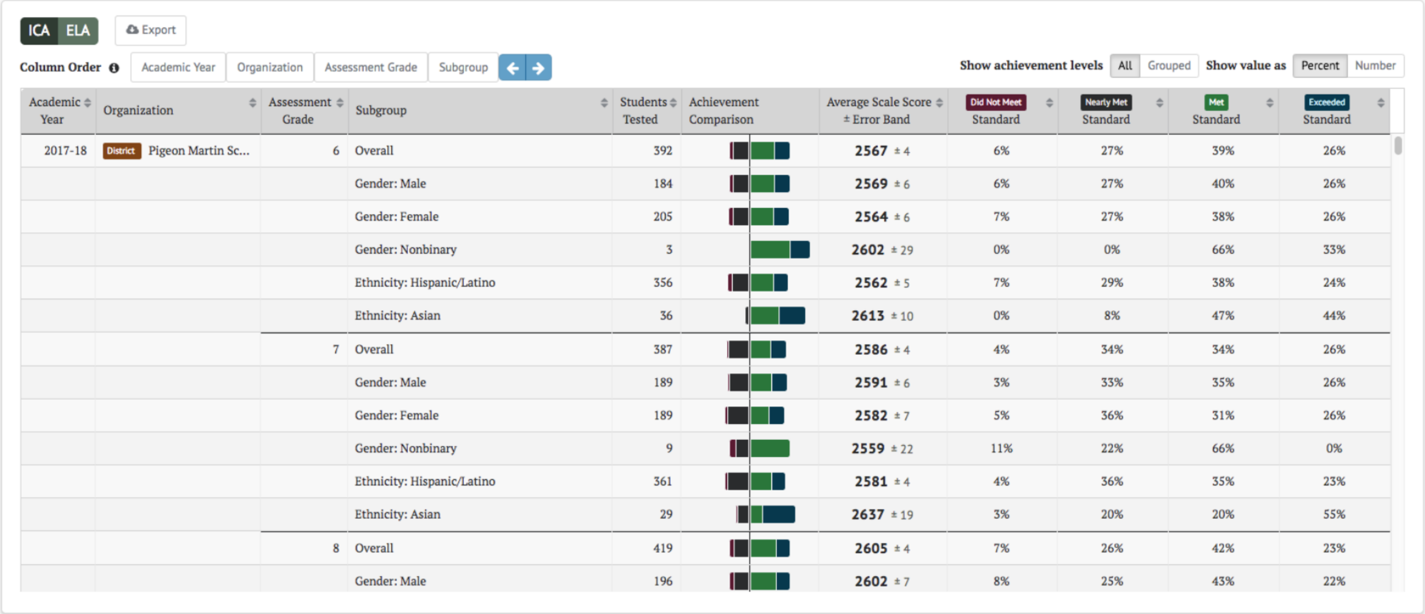


Student	Date	Session	Enrolled Grade	School	Status	Achievement Level	Scale Score / Error Band
Bell, Francis	May 11, 2018	MAT-69dd	G8	Woodcreeper Whale Intermediate School		Nearly Met Standard	**2543** ± 68
Bigler, Gary	May 11, 2018	MAT-69dd	G8	Woodcreeper Whale Intermediate School		Met Standard	**2600** ± 90
Brosius, Michael	May 11, 2018	PER-eb45	G8	Woodcreeper Whale Intermediate School		Exceeded Standard	**2671** ± 88
Brown-Payne, Brian	May 11, 2018	CUM-5268	G8	Woodcreeper Whale Intermediate School		Nearly Met Standard	**2565** ± 98
Bryan, Claude	May 11, 2018	MAT-69dd	G8	Woodcreeper Whale Intermediate School		Nearly Met Standard	**2584** ± 60
Burton, Geraldine	May 11, 2018	SIL-7b21	G8	Woodcreeper Whale Intermediate School		Exceeded Standard	**2650** ± 93
Carruth, Roger	May 11, 2018	SIM-5006	G8	Woodcreeper Whale Intermediate School		Nearly Met Standard	**2536** ± 91
Casey, Manuel	May 11, 2018	MAT-69dd	G8	Woodcreeper Whale Intermediate School		Met Standard	**2626** ± 82
Cochran, James	May 11, 2018	EME-7e0b	G8	Woodcreeper Whale Intermediate School		Met Standard	**2644** ± 69
Constant, Valarie	May 11, 2018	SIM-5006	G8	Woodcreeper Whale Intermediate School		Met Standard	**2598** ± 64
Cotton, Morgan	May 11, 2018	SIL-7b21	G8	Woodcreeper Whale Intermediate School		Exceeded Standard	**2672** ± 55
Drye, Scott	May 11, 2018	KEN-a709	G8	Woodcreeper Whale Intermediate School		Met Standard	**2622** ± 59

(sample OLTP -based report)

Data Analysis

Data analysis reports have two characteristics that make OLTP not such a good option: they require aggregation queries and the ability to query on any permutation of the data elements.



Academic Year	Organization	Assessment Grade	Subgroup	Students Tested	Achievement Comparison	Average Scale Score ± Error Band	Did Not Meet Standard	Nearly Met Standard	Met Standard	Exceeded Standard
2017-18	District Pigeon Martin Sc...	6	Overall	392		2567 ± 4	6%	27%	39%	26%
			Gender: Male	184		2569 ± 6	6%	27%	40%	26%
			Gender: Female	205		2564 ± 6	7%	27%	38%	26%
			Gender: Nonbinary	3		2602 ± 29	0%	0%	66%	33%
			Ethnicity: Hispanic/Latino	356		2562 ± 5	7%	29%	38%	24%
			Ethnicity: Asian	36		2613 ± 10	0%	8%	47%	44%
		7	Overall	387		2586 ± 4	4%	34%	34%	26%
			Gender: Male	189		2591 ± 6	3%	33%	35%	26%
			Gender: Female	189		2582 ± 7	5%	36%	31%	26%
			Gender: Nonbinary	9		2559 ± 22	11%	22%	66%	0%
			Ethnicity: Hispanic/Latino	361		2581 ± 4	4%	36%	35%	23%
			Ethnicity: Asian	29		2637 ± 19	3%	20%	20%	55%
		8	Overall	419		2605 ± 4	7%	26%	42%	23%
			Gender: Male	196		2602 ± 7	8%	25%	43%	22%

(sample OLAP-based report)

Aggregation queries in OLTP

Yes, ANSI SQL supports most common aggregate functions such as SUM, COUNT, AVG, MIN/MAX, etc. There are also SQL 'windowing' functions to help with performing aggregate calculations across a set of related tables. With a small data set, that may be sufficient. But, as the data size grows, OLTP systems would be difficult (if not impossible) to optimize for these types of queries. Just Google "slow count" in MySQL or Postgres (or any other OLTP engine), and you'll find lots of references.

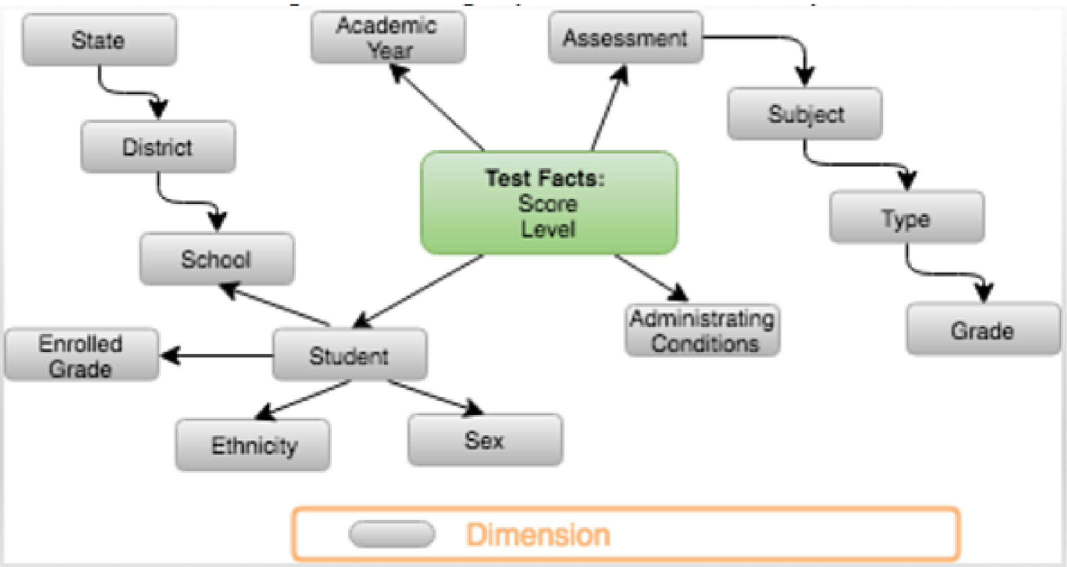
Query on any permutation of the data elements with OLTP

RDBMSes use indexes to optimize the queries. This works well with a pre-defined combination of search data elements. But, our project needed to support ad-hoc queries on any permutation of the data elements. This becomes problematic with a large volume of data. Creating many indexes with different permutations will confuse the query optimizer and cause all kinds of problems. Some databases are more flexible with combining multiple indexes at runtime (PostgreSQL, for example, is better than MySQL) but they still have limitations.

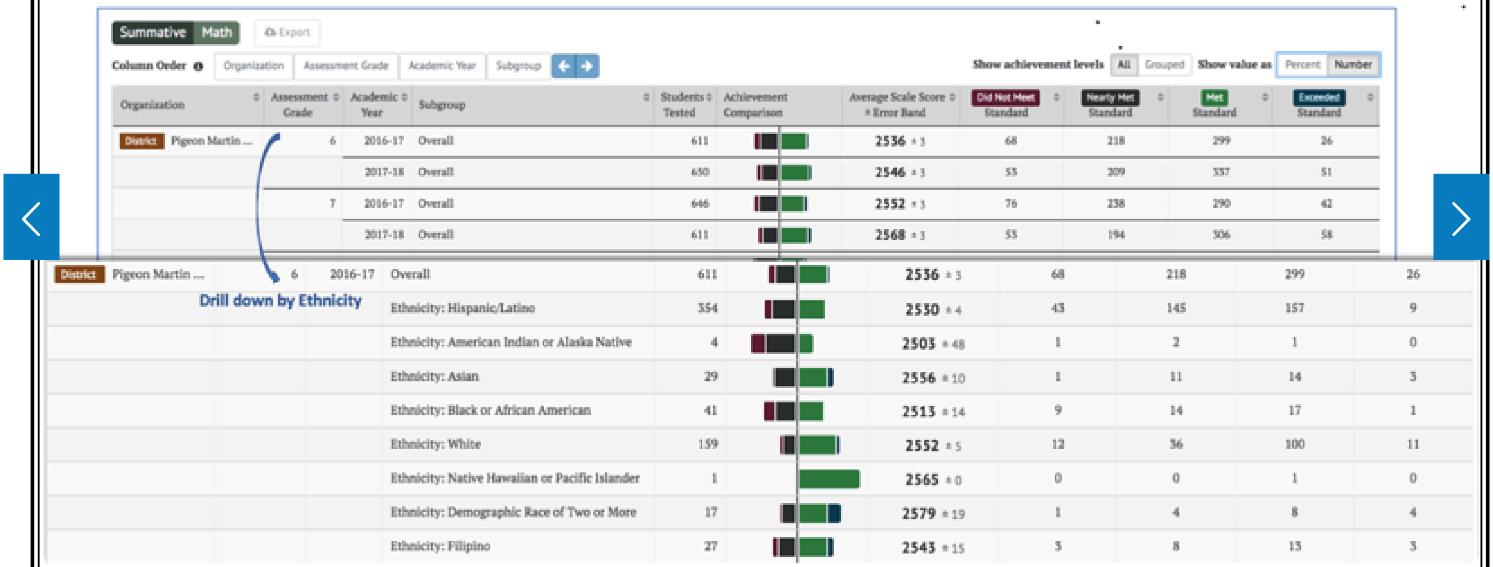
OLAP (Online Analytical Processing) solutions were created to address OLTP challenges with data analy

Online Analytic Processing (OLAP) - Brief Overview

In OLAP, data is organized into "dimensions" and "measures" (aka "facts"). The former represent attributes of the reported entities, while the latter is derived from the number-based characteristic of the entities. Users can build their own reports by organizing dimensions into tables (column headers/or row titles) and viewing the measures (aka slicing & dicing). Dimensions can also be organized into a hierarchy that allows for reporting the measures at different levels of a hierarchy: roll-up (consolidation), and drill-down. Think Excel Pivot table.



Facts and Dimension Example



Organization	Assessment Grade	Academic Year	Subgroup	Students Tested	Achievement Comparison	Average Scale Score ± Error Band	Did Not Meet Standard	Nearly Met Standard	Met Standard	Exceeded Standard
District Pigeon Martin ...	6	2016-17	Overall	611		2536 ± 3	68	218	299	26
		2017-18	Overall	650		2546 ± 3	53	209	337	51
	7	2016-17	Overall	646		2552 ± 3	76	238	290	42
		2017-18	Overall	611		2568 ± 3	53	194	306	58

Drill down by Ethnicity

District Pigeon Martin ...	6	2016-17	Overall	611		2536 ± 3	68	218	299	26
			Ethnicity: Hispanic/Latino	354		2530 ± 4	43	145	157	9
			Ethnicity: American Indian or Alaska Native	4		2503 ± 48	1	2	1	0
			Ethnicity: Asian	29		2556 ± 10	1	11	14	3
			Ethnicity: Black or African American	41		2513 ± 14	9	14	17	1
			Ethnicity: White	159		2552 ± 5	12	36	100	11
			Ethnicity: Native Hawaiian or Pacific Islander	1		2565 ± 0	0	0	1	0
			Ethnicity: Demographic Race of Two or More	17		2579 ± 19	1	4	8	4
			Ethnicity: Filipino	27		2543 ± 15	3	8	13	3

Facts (Students Tested Count, Average Score, Students Tested Counts in four achievement levels) organized by District, Assessment Grade, and Academic Year dimensions. Then, an example of a drill-down using the Ethnicity dimension.

Traditional OLAP databases usually store pre-aggregated, historical data in multi-dimensional schemas. https://en.wikipedia.org/wiki/Comparison_of_OLAP_servers.

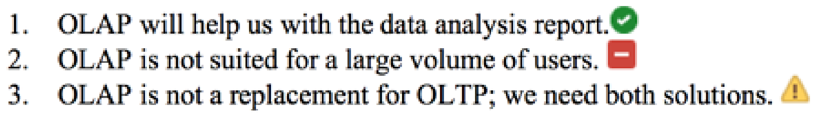
Columnar data stores,as the name implies, store the data in columns. They have characteristics that are well-suited for OLAP-like workloads. Storing data in columns allows for reduced and compressed storage, and because of that, these type of DBs perform particularly well with aggregation queries (such as SUM, COUNT, AVG, etc). Columnar databases are also very scalable and are well suited to massive parallel processing (MPP), which involves having data spread across a large cluster of machines. A few notable solutions in this category are:

Open Source: Greenplum Postgres-XL, Druid, Presto (with Parquet file store)

Managed: Google BigQuery, Amazon Redshift, Snowflake

Commercial: Vertica, Oracle, IBM Db2.

OLAP systems are generally optimized for a lower rate of concurrent transactions with complex queries (reads), and their effectiveness is measured by query response time. This statement highlights the following three points:



1. OLAP will help us with the data analysis report. ✓
2. OLAP is not suited for a large volume of users. ⊖
3. OLAP is not a replacement for OLTP; we need both solutions. ⚠

With OLAP, we can fulfill the requirement of providing complex customer-built reports to a small population of users (district and state administrators), but expanding this to all teachers is going to be problematic. This happened to be one of the tricky data problems our team faced. Our solution falls between OLAP and OLTP solutions, with neither of them being a good fit.

Shifting reporting responsibilities in the application

We solved this problem by dividing the reporting responsibilities between the DB layer and the application layer. The reports were reviewed and analyzed to define a database query constrained to a predefined subset of the data elements (e.g. one school, one subject, and x number of years). This allowed us to use the strength of OLTP indexes to quickly return the needed set of data to the application layer. Since these reports were focused on one class or one school, the returned data size was small enough for the application layer to handle. The application then performed the additional filtering/sorting in memory.

Summary

The Educational Reporting Data Warehouse presented the team with particular data reporting/analysis challenges, similar to issues encountered in many data warehouse systems. Neither an OLTP or OLAP solution by itself addressed business needs. The resulting product is a hybrid solution using Amazon Aurora (OLTP) for data reporting and Amazon Redshift (OLAP) for data analysis reports.

Bibliography

1. https://en.wikipedia.org/wiki/Online_transaction_processing
2. https://en.wikipedia.org/wiki/Online_analytical_processing
3. https://en.wikipedia.org/wiki/Data_reporting
4. https://datahero.com/blog/2017/08/14/data-reporting-and-analysis
5. https://en.wikipedia.org/wiki/ACID_(computer_science)
6. https://en.wikipedia.org/wiki/Pivot_table

For weekly tech updates ...