LIGHTHOUSE PROJECT SUMMARY

A (tiny) Data Warehouse

Redactions

Before we go any further I should mention that this document has numerous redactions to protect a client's privacy: all names have been changed and some information has been removed. Where necessary, substitute names and values have been used.

The Problems

As the analytics team's efforts evolved the team discovered that analysis was the easy part: this new data-centric work had two components that were proving to be highly labor-intensive:

- Merging disparate datasets (surveys, Salesforce, federal/state agency data).
- Managing formatting differences in datasets (some datasets vary their scale from year to year, like using 1-10 instead of 1-100, or changes to dataset conventions over the years).
- Managing inaccuracies/errors in datasets.

All of these processes were being done entirely manually in Excel for datasets that had thousands of rows encompassing sometimes millions of individual cells. Given the onerous nature of this approach it has been assigned a pejorative nickname in data and analytics circles: spreadmart (spreadsheet datamart).

The Solution

As a stop-gap solution I spun up a MySQL relational database for a one-time data project. This was found to be a huge time-saver, allowing the team to focus their resources on applying the analytics and developing a data strategy instead of manual data processing. When it came time to begin budget planning for the next year a data strategy began to form that involved wider investment in Tableau licenses along with staffing to support this new data infrastructure. Over the next couple years this strategy evolved into a rudimentary data warehouse named **Lighthouse**.

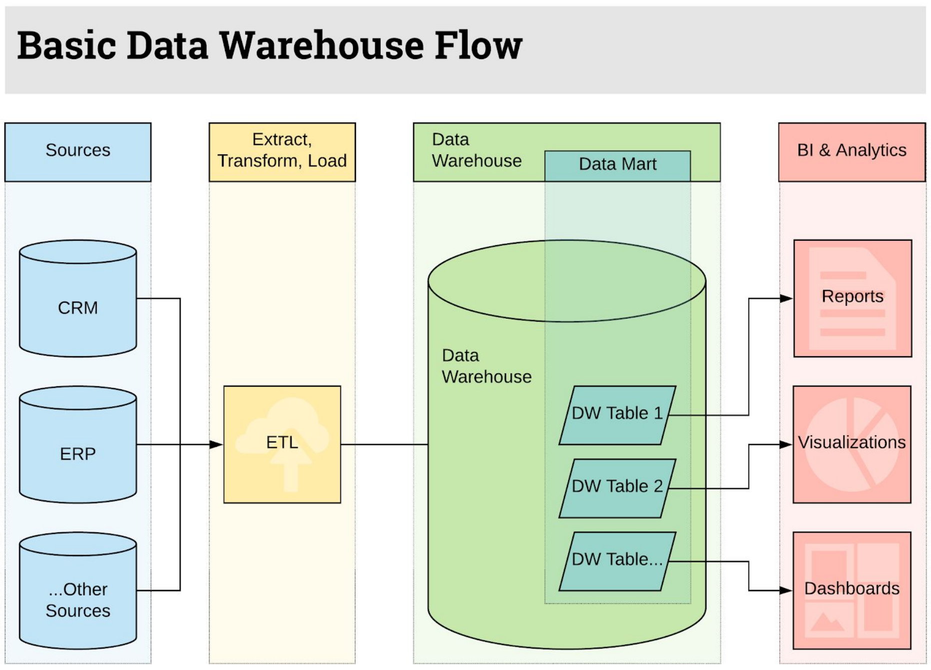
Data Warehousing

- *"…a system used for reporting and data analysis… DWs are central repositories of integrated data from one or more disparate sources. They store current and historical data in one single place"* -Wikipedia
- Stores data from multiple sources (vs siloed).
- Salesforce, state & federal agency, surveys… we can store and process almost any data source.
- Provides a workspace for batch processing, cleaning, and normalization operations.
- Prepares secure, easily digestible endpoints for visualizing, reporting, and dashboarding.

Data ingestion (collection and import of requisite data into the warehouse) is typically handled in one of three ways:

- Manual imports: State/federal .csv files manually imported into the warehouse once or twice a year using a database import tool or IDE, such as DataGrip or Valentina Studio.
- Pushed via API: some applications and services may provide options to automatically push their data to a database. For example, we could potentially find a Salesforce plugin that can replicate tables over to a Postgres database.
- Pulled via ETL: Extract, Transform, Load solutions simplify and automate movement of datasets between locations. This is the preferred method of data ingestion.

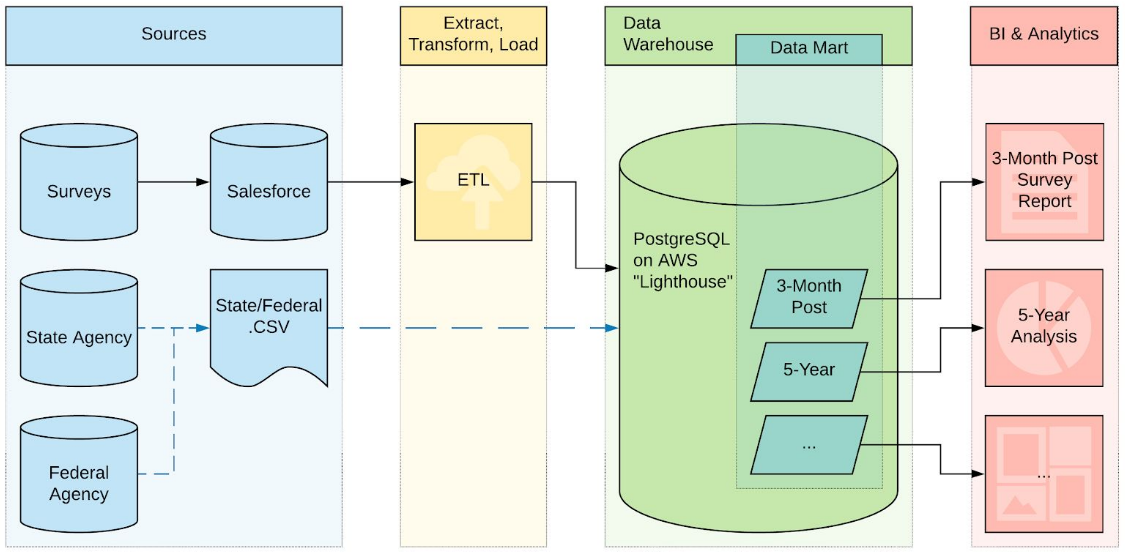
Once all of the requisite data has arrived in the warehouse we can begin to create additional tables in the warehouse that join and filter the data to cater to specific BI (business intelligence) and analytics purposes. The collection of these purpose-built tables we'll call the data mart.



Lighthouse Data Flow

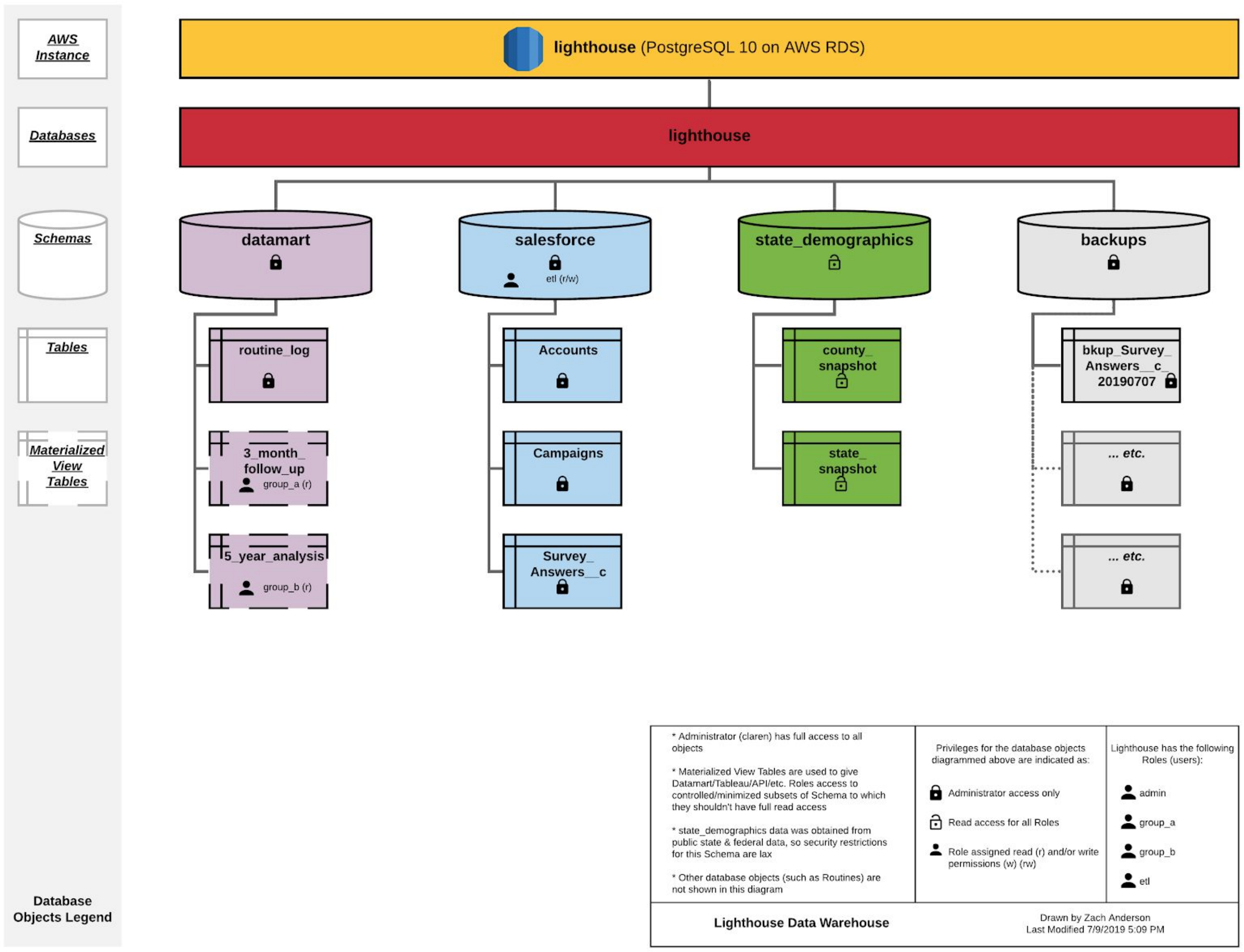
The specific flow of data for Lighthouse is only slightly more complex:

Lighthouse Data Warehouse Flow



- Surveys collected from client and program participants are automatically pushed to Salesforce (into the `Survey_Answers__c` object) upon submission. This tight Salesforce integration was a key reason we use the surveying tools we do.
- State and federal agency reports are provided as .csv files on an annual basis. Our **Data Preparation and Uploading Process** document dictates how these datasets need to be prepared prior to import into Lighthouse. Once the prepared .csv is ready it is manually imported into Lighthouse using Valentina Studio (after the impacted tables have been backed up).
- Our ETL solution is used to perform a one-way sync (aka replication) of our Salesforce data into Lighthouse. Salesforce objects that we want to report on (and merge against state/federal or other datasets) are all synced.
- Lighthouse itself is a PostgreSQL 10 relational database hosted on AWS (Amazon Web Services) RDS (Relational Database Service). We'll go into the specific schema/table structure in the Lighthouse Schema section of this document.
- Our BI and analytics endpoints are almost exclusively Tableau Packaged Workbooks. These files are openable by anybody with Tableau Reader (not unlike PDF files and Adobe Reader) but can only be created using Tableau Desktop. We've successfully built test BI & analytics endpoints that included live interactive maps (using tools like Carto and CanvasJS) and dashboards accessible via the web as either staff-only or publicly accessible visualizations, but have yet to officially deploy them.

Lighthouse Schema



* Administrator (claren) has full access to all objects

* Materialized View Tables are used to give Datamart/Tableau/API/etc. Roles access to controlled/minimized subsets of Schema to which they shouldn't have full read access

* state_demographics data was obtained from public state & federal data, so security restrictions for this Schema are lax

* Other database objects (such as Routines) are not shown in this diagram

Privileges for the database objects diagrammed above are indicated as:

- 🔒 Administrator access only
- 🔓 Read access for all Roles
- 👤 Role assigned read (r) and/or write permissions (w) (rw)

Lighthouse has the following Roles (users):

- 👤 admin
- 👤 group_a
- 👤 group_b
- 👤 etl

Drawn by Zach Anderson
Last Modified 7/9/2019 5:09 PM

- Lighthouse is an AWS RDS PostgreSQL 10 Instance.
- Our instance runs a single database, also named Lighthouse.
- Our database is composed of four **schema**. Schemas are used in PostgreSQL to organize **tables** and other database objects for organizational and privileging/access control purposes.
- Our `salesforce` schema is replicated via one-way sync daily by our ETL from our Salesforce. You can see that we have selected four Salesforce Objects to sync to tables in Lighthouse.
- The **materialized views** within the datamart schema are essentially calculated tables: they combine and structure data sourced from the other tables in Lighthouse into a format specifically tailored for a visualization endpoint, such as a specific Tableau Workbook. As such, the materialized views are typically named after the Workbook or visualization they provide data for.
- Tableau Workbooks (and other visualization/analytics endpoints) will need to log in to Lighthouse, typically using either the `group_a` or the `group_b` **roles** (roles == users in PostgreSQL). Both the `group_a` and `group_b` roles are only given access to their specific materialized views in observance of the principle of least privilege. These materialized views are comprised of the very minimum data required for their intended visualizations.
- The `backups` schema is used to make backups of tables and other database objects.

Routines & Triggers

ROUTINES

For each Materialized View in `lighthouse.datamart` there is a corresponding custom User-Defined Function (UDF) which refreshes the content of that Materialized View when executed. At the time of this writing there are two Materialized Views (as shown in the diagram on the previous page) whose corresponding Functions are:

- `lighthouse.datamart.refresh_3_month_follow_up`
- `lighthouse.datamart.refresh_5_year_analysis`

All of these UDFs will also add a new Row to the `lighthouse.datamart.routine_log` Table. This allows us to track the history of these UDFs being executed.

TRIGGERS

Every time the ETL runs the Salesforce sync there are also Triggers on the `salesforce.Campaign` Table responsible for actually executing the UDFs outlined above. Any INSERT, UPDATE, or DELETE operation performed on the `salesforce.”Campaign”` Table will execute all of these Triggers thus keeping the Materialized Views up-to-date. At the time of this writing there are five Materialized Views, each with a corresponding UDF (outlined above) whose corresponding Triggers are:

- `lighthouse.salesforce.”Campaign”.trigger_refresh_3_month_follow_up`
- `lghthouse.salesforce.”Campaign”.trigger_refresh_5_year_analysis`

This timeline of these events looks like this:

1. The ETL runs the Salesforce Replication (one-way sync) process every day around 9:00am.
2. If any rows are added, modified, or removed from the `salesforce.”Campaign”` Table by the ETL then the Triggers outlined above are all executed.
3. Each of these Triggers has one responsibility: execute its corresponding UDF.
4. Each of these UDFs has one responsibility: update its corresponding Materialized View.

Additional Routines & Objects

USER-DEFINED FUNCTIONS

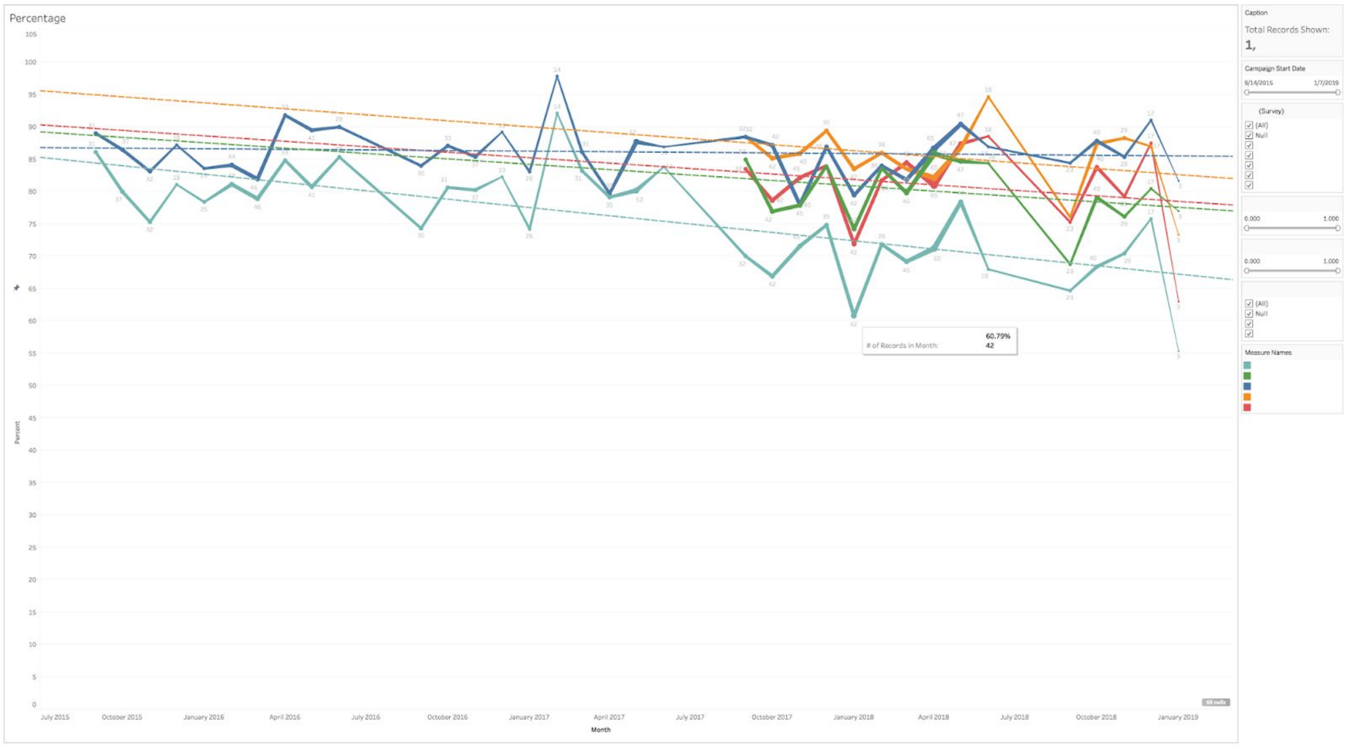
`lighthouse.salesforce.salesforce_15_to_18_id()`
Converts 15-character length Salesforce IDs into their 18-character length equivalent.

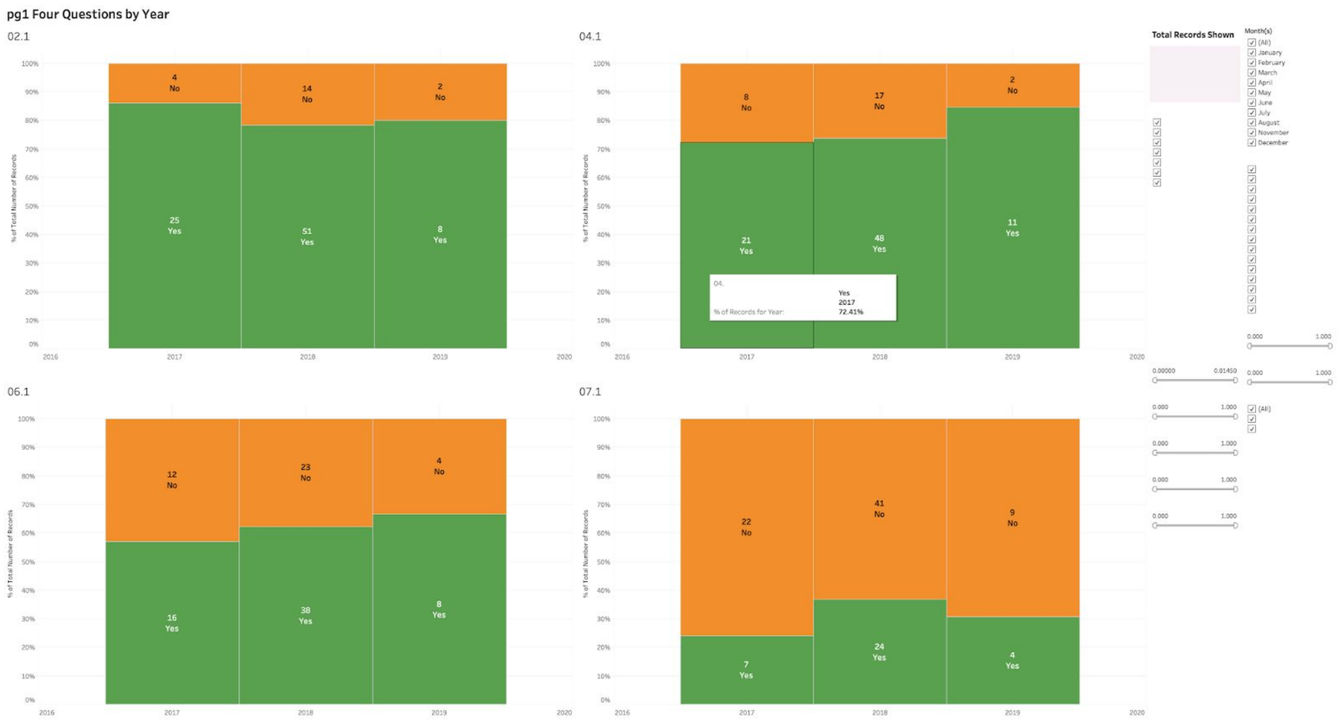
USER-DEFINED DOMAINS

`lighthouse.student_demographics.fiscal_year`
Restricts `fiscal_year` Fields to `smallint` between 1970 and 2100.

`lighthouse.student_demographics.government_or_nonprofit`
Restricts `government_or_nonprofit` Fields to `varchar(1)` that is either 'Government' or ‘Nonprofit’.

Gallery (Groups A & B Tableau Workbooks)





Potential Use Cases / Expansions

Lighthouse provides a wealth of data and analytics capabilities than could expand the current use cases of all these departments and groups that are performing this sort of data analysis. As with most things the ‘sky's the limit’... provided financial and staffing resources can afford the ‘sky’. Before I so much as enumerate some possible examples I recommend that we be clear on the direction of data strategy and conscious of how much resources everyone is willing to dedicate to said strategy.

Projected Costs

The table below estimates annual costs (365 days) based on expenses incurred over the last 12 months calculated against 2019 contract rates and budgeted hours:

Expense Item	Expense Details	Estimated Yearly Cost
AWS	RDS db.t2.micro in us-west-2c region	$191.88
ETL	Introductory Plan	$600.00
Zach**	Data Warehouse Administration**	$xx,xxx.xx
	Group A Report Building**	$xx,xxx.xx
	Group B Report Building**	$xx,xxx.xx
		TOTAL
		$xx,xxx.xx

***Estimates based on hours spent on these tasks from June 03rd, 2018 - July 07th, 2019. These are effectively based on last year’s averages and assume the projects in these areas continue at the same pace.*