Migrating to Snowflake the Easy Way

Adam Boscarino
Karla Goodreau
Mallori Harrell

DevotedHealth

Data Engineering Team

Karla Goodreau

Adam Boscarino

Mallori Harrell

Osman Qamar

WE ARE HIRING!!!!

Team Data, early 2019

- Devoted Health's Medicare Advantage Plan just went live on ***1/1/2019***!
- Team of ***9*** highly skilled ***Data Scientists*** working on workstreams for the business building lots of pipelines
- ***3 Data Engineers***
- Rapid development to support business

Data Infrastructure, January 2019

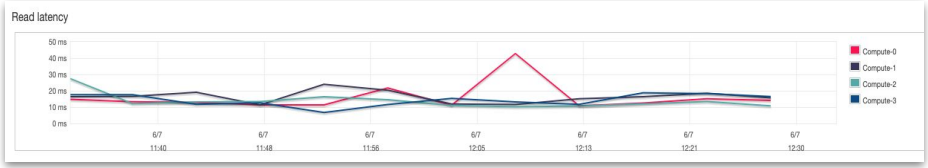
Source Data	Job Scheduler	Storage/Data Lake	Data Warehouse	Reporting/BI
	[~]$ crontab	Amazon S3	amazon REDSHIFT	Periscope Data by Sisense

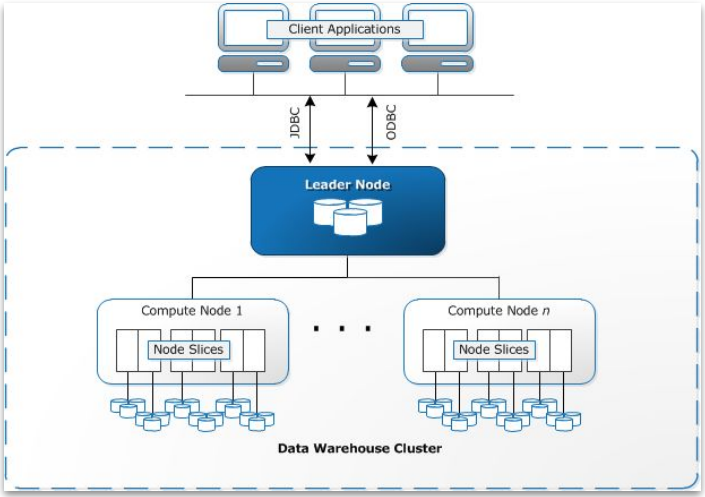
Data Infrastructure, May 2019

Source Data	Job Scheduler	Storage/Data Lake	Data Warehouse	Reporting/BI
	NEW	Amazon S3	NEW	

Why the rush to get off of Redshift?

- Increased Latency
  - Long wait times for key processes
- Concurrency Issues
  - Analysts blocked by ETL
- Scaling for increasing membership/complexity
  - In Redshift, this means adding more nodes (storage and compute are combined)
  - We would need to scale everything for our slowest processes

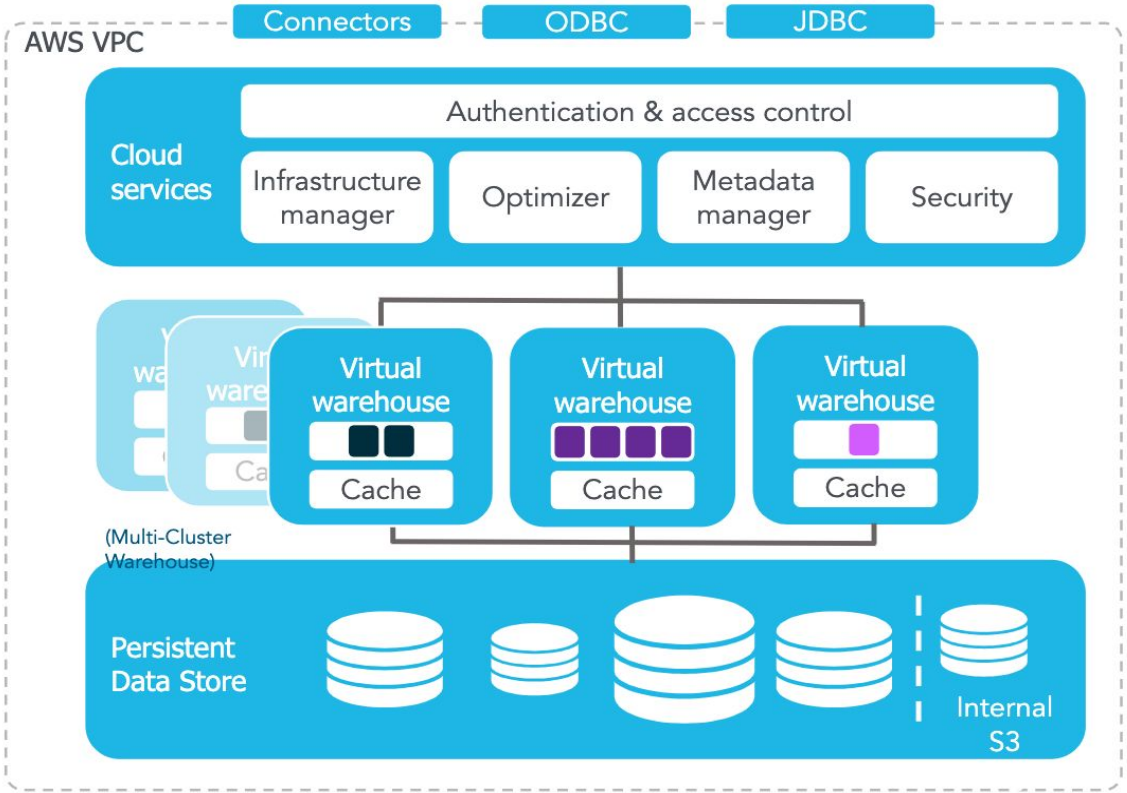




WE KNEW THERE WAS BETTER!!!

Why Snowflake?

Snowflake's Multi-Cluster, Shared Data Architecture



- **Cloud Services**
  Scalable, resilient cloud services layer coordinates access & management
- **Independently Scalable Compute**
  Multiple "virtual warehouses" compute clusters scale horsepower & concurrency
- **Centralized Storage**
  Instant, automatic scalability & elasticity

External Customer S3 Stage

Lessons Learned: Pro-Tips from Previous Migrations

- Redshift and Snowflake are different!
  - SQL Syntax Differences
    - Dates - DATEADD vs. INTERVAL
    - Join Conditions - USING vs. ON
    - TIMESTAMP casting converts to TIMESTAMP_NTZ - need to cast to TIMESTAMP_TZ if has a timestamp
    - GET_DATE vs. CURRENT_DATE or CURRENT_TIMESTAMP
    - BOOL_AND/BOOL_OR to BITAND_AGG/BITOR_AGG
    - !~ to NOT REGEXP_LIKE
  - Snowflake's default timezone is PST - Need to update to UTC!
  - Ensure NULLs and empty strings are handled in the same way in both
- Automate as much as you can to avoid human error
  - Very important for data validation!!
- Limit scope of changes
  - Avoid additional changes during migration like updating naming conventions or the data model

Migration to Snowflake: The Plan

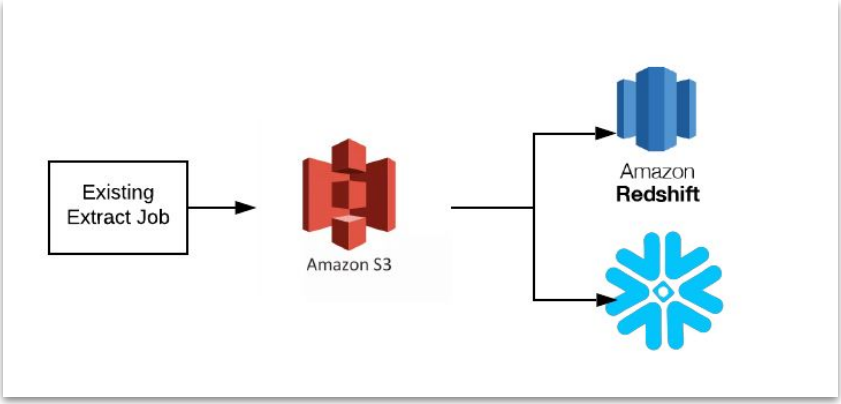
1. ***Migrate*** the data
   - Move ALL THE DATA from Redshift to Snowflake (SnowShift)
   - Repoint Scheduled pipelines (DAGs)
2. ***Verify*** the data
   - Compare Snowflake and Redshift tables (Checkup)
3. ***Migrate*** dashboards
   - Repoint Periscope dashboards to Snowflake (SnowSkulpt)
4. ***Finalize*** with stakeholders
   - Data Validation-athon™

All Roads Lead to the Redshift Sync

~1200 tables across 14 services

Migrate Data: Phase 1

- Started with hourly sync of Production data (~1200 tables)
- Dual wrote to Snowflake and Redshift from S3 files generated by existing job
- Quick validation since same data was written to both data warehouses

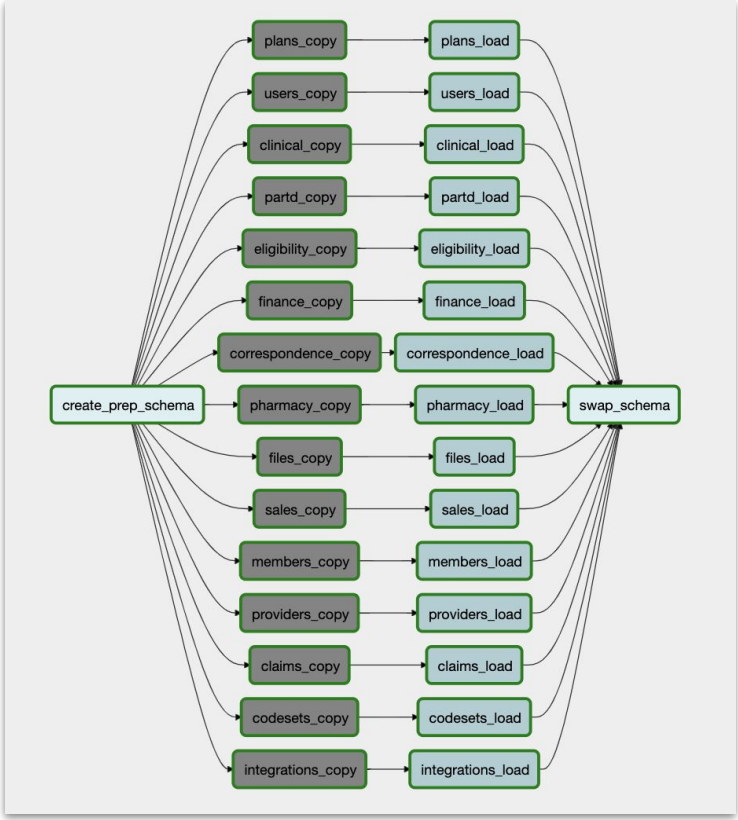


End Result: Production data in Snowflake, same issues with existing sync as before

Migrate Data: Phase 2

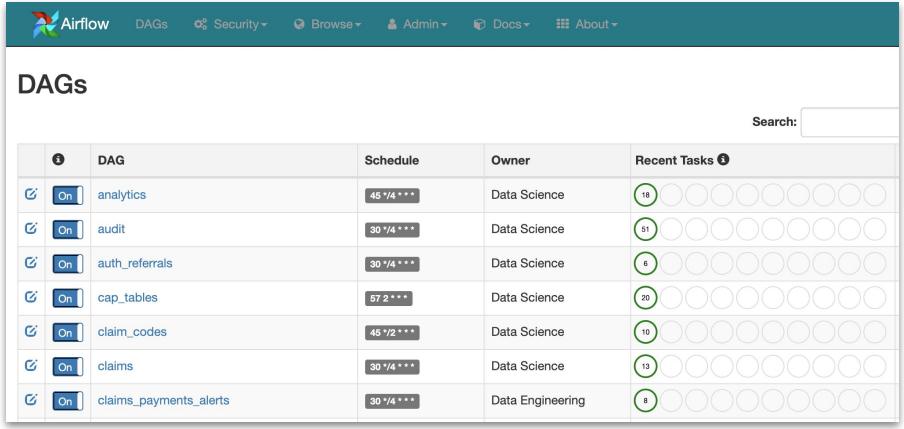
- Redesigned Production Sync for Airflow
  - Incrementally load append only tables
  - Automatically handle schema changes
  - Validate source to target record counts
- Now generating and writing data differently, requiring more validation

End Result: All Production data in Snowflake, improved reliability, but data requires validation



Migrate Data: Phase 3

- SnowShift - a tool for migrating static datasets to Snowflake
- Worked with Data Science to migrate transformation/analytics pipelines used to populate data warehouse
- Ran parallel DAGS for these pipelines, which resulted in two maintaining two sets of code



End Result: All data in Snowflake, but data requires validation

Validate Data: Checkup

Checkup - a tool to automate data validation between Redshift and Snowflake

- Copies a Redshift table to Snowflake
- Executes a series of queries to compare the two tables
- Can compare a single table or entire schema
- Stores results in an additional table for analysis and historical comparison
- Used to validate over 1,200 tables and multiple schemas

End Result: All data in Snowflake and fully validated!

```
python checkup -h

usage: checkup [-h] --env ENV [--redshift-table REDSHIFT_TABLE]
                   [--snowflake-table SNOWFLAKE_TABLE]
                   [--tables-to-compare TABLES_TO_COMPARE]
                   [--tables-file TABLES_FILE]
                   [--redshift-schema REDSHIFT_SCHEMA]
                   [--snowflake-schema SNOWFLAKE_SCHEMA]

tool for validating Snowflake tables against Redshift tables

optional arguments:
  -h, --help            show this help message and exit
  --env ENV, -e ENV     the environment to compare tables in (staging or prod)
  --redshift-table REDSHIFT_TABLE, -r REDSHIFT_TABLE
                        the name of the redshift table to use for validation
  --snowflake-table SNOWFLAKE_TABLE, -s SNOWFLAKE_TABLE
                        the name of the snowflake table to use for validation
  --tables-to-compare TABLES_TO_COMPARE, -t TABLES_TO_COMPARE
                        a dictionary of the tables to compare. Expected format
                        is {snowflake_table: redshift_table}.
  --tables-file TABLES_FILE, -f TABLES_FILE
                        a file of the tables to compare. Expected format is
                        {snowflake_table: redshift_table}.
  --redshift-schema REDSHIFT_SCHEMA
                        a Redshift schema to compare.
  --snowflake-schema SNOWFLAKE_SCHEMA
                        a Snowflake schema to compare.
```

Migrate Dashboards

Periscope Usage Facts:

- **88%** of Devotees have a Periscope account
- **90%** are active on a monthly basis (MAU)
- **55%** are active on a weekly basis (WAU)

THIS IS THE USER FACING PART OF THE MIGRATION!!

Migrate Dashboards

Periscope Makeup:

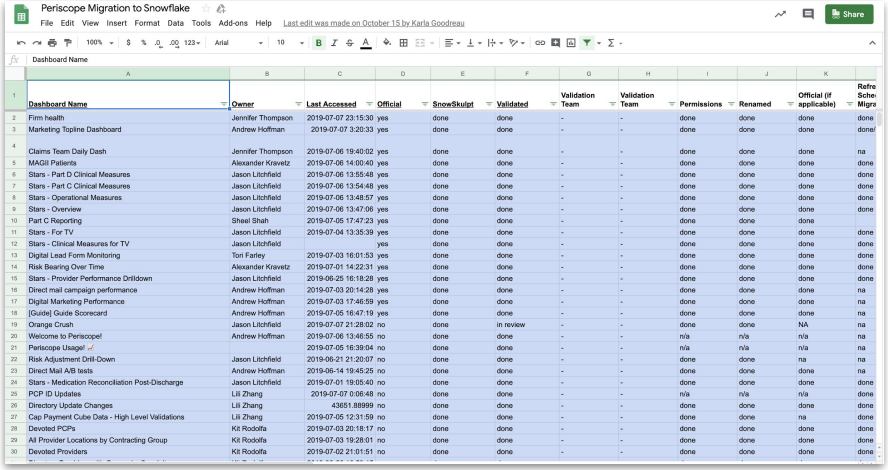
- We have more than **1400 Dashboards**
- We have over **7000 Charts**
- Over **13000 lines of SQL** for Periscope reports
- Over **25000 lines of SQL for Periscope** views

O M G

What do we do???

Periscope Dashboard Tracking

- Created a dashboard that showed which tables were validated
- Connected this to the chart level
- Rolled up to Dashboard level
- Let’s us know which dashboards are ready to be migrated



Dashboard Name	Owner	Last Accessed	Official	SnowSkulpt	Validated	Validation Team	Validation Team	Permissions	Renamed	Official (if applicable)	Refre Sche Migra
Firm health	Jennifer Thompson	2019-07-07 23:15:30	yes	done	done	-	-	done	done	done	done
Marketing Topline Dashboard	Andrew Hoffman	2019-07-07 3:20:33	yes	done	done	-	-	done	done	done	done
Claims Team Daily Dash	Jennifer Thompson	2019-07-06 19:40:02	yes	done	done	-	-	done	done	done	na
MAGII Patients	Alexander Kravetz	2019-07-06 14:00:40	yes	done	done	-	-	done	done	done	done
Stars - Part D Clinical Measures	Jason Litchfield	2019-07-06 13:55:48	yes	done	done	-	-	done	done	done	done
Stars - Part C Clinical Measures	Jason Litchfield	2019-07-06 13:54:48	yes	done	done	-	-	done	done	done	done
Stars - Operational Measures	Jason Litchfield	2019-07-06 13:48:57	yes	done	done	-	-	done	done	done	done
Stars - Overview	Jason Litchfield	2019-07-06 13:47:06	yes	done	done	-	-	done	done	done	done
Part C Reporting	Sheal Shah	2019-07-05 17:47:23	yes	done	done	-	-	done	done	done	
Stars - For TV	Jason Litchfield	2019-07-04 13:35:39	yes	done	done	-	-	done	done	done	done
Stars - Clinical Measures for TV	Jason Litchfield		yes	done	done	-	-	done	done	done	done
Digital Lead Form Monitoring	Tori Farley	2019-07-03 16:01:53	yes	done	done	-	-	done	done	done	done
Risk Bearing Over Time	Alexander Kravetz	2019-07-01 14:22:31	yes	done	done	-	-	done	done	done	done
Stars - Provider Performance Drilldown	Jason Litchfield	2019-06-25 16:18:28	yes	done	done	-	-	done	done	done	done
Direct mail campaign performance	Andrew Hoffman	2019-07-03 20:14:28	yes	done	done	-	-	done	done	done	na
Digital Marketing Performance	Andrew Hoffman	2019-07-03 17:46:59	yes	done	done	-	-	done	done	done	na
[Guide] Guide Scorecard	Andrew Hoffman	2019-07-05 16:47:19	yes	done	done	-	-	done	done	done	na
Orange Crush	Jason Litchfield	2019-07-07 21:28:02	no	done	in review	-	-	done	done	NA	na
Welcome to Periscope!	Andrew Hoffman	2019-07-06 13:46:55	no	done	done	-	-	n/a	n/a	n/a	na
Periscope Usage! 📈		2019-07-05 16:39:04	no	done	done	-	-	n/a	n/a	n/a	na
Risk Adjustment Drill-Down	Jason Litchfield	2019-06-21 21:20:07	no	done	done	-	-	done	done	na	na
Direct Mail A/B tests	Andrew Hoffman	2019-06-14 19:45:25	no	done	done	-	-	done	done	done	na
Stars - Medication Reconciliation Post-Discharge	Jason Litchfield	2019-07-01 19:05:40	no	done	done	-	-	done	done	done	done
PCP ID Updates	Lili Zhang	2019-07-07 0:06:48	no	done	done	-	-	n/a	n/a	n/a	done
Directory Update Changes	Lili Zhang	43651.88999	no	done	done	-	-	done	done	done	done
Cap Payment Cube Data - High Level Validations	Lili Zhang	2019-07-05 12:31:59	no	done	done	-	-	done	done	na	done
Devoted PCPs	Kit Rodolfa	2019-07-03 20:18:17	no	done	done	-	-	done	done	done	done
All Provider Locations by Contracting Group	Kit Rodolfa	2019-07-03 19:28:01	no	done	done	-	-	done	done	done	done
Devoted Providers	Kit Rodolfa	2019-07-02 21:01:51	no	done	done	-	-	done	done	done	done

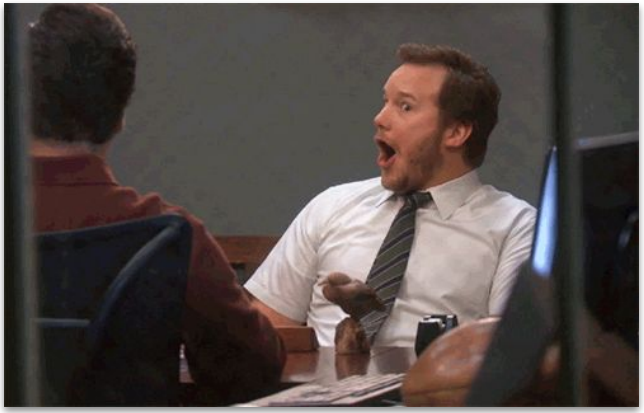
How do we migrate these CRAP TON of reports?

The thought of:

- Going through over a 1000 dashboards with
- 1000s of charts and
- Over 13000 lines of code to change

PAINFUL!!!

So painful we named that sprint Joffrey!



Then we realized, **EVERYTHING** in Periscope is committed to a git repo.

SQL Conversion: SnowSkulpt

- SnowSkulpt - automates converting Redshift SQL to Snowflake SQL and points all dashboards to the Snowflake warehouse.
- Point SnowSkulpt to a dashboard directory
- Uses regular expressions to convert SQL
- Handles most Redshift to Snowflake syntax differences
- Creates a duplicate directory for the Snowflake version of the dashboard

Data Validation-athon

We are done with Periscope Migration..... NOW WHAT? CAN WE SWITCH TO SNOWFLAKE???

- Paired with DS and Analytics
- Sat in a room on a video hangout for a full day
- Provided Pizza & snacks
- Managed a running Spreadsheet with status and owners and validated dashboards
- Once validated, we turned them live and archived the Redshift Dashboard

WE KILLED REDSHIFT!

"It's not a crime" - *Anonymous Data Engineer*

Lessons Learned: More Pro-Tips

- Maintaining two sets of the same code is bad
  - Some work was done to Redshift pipelines, but not Snowflake resulting in rework
- Migrating is a Team Sport
  - Involving stakeholders earlier would have made validation even easier
- We still **love** Snowflake!

8 months later...Thoughts from Data Science

How is it awesome for Data Science?

- That UI 😍😍😍
- Speed. We feel the need for speed.
- Enabled personal dev databases ->
  better Airflow experience ->
  better, faster work
- Built-in functions
- Working with tables is easier: clone, undrop, create or replace

Questions?