

---

---

# Transitioning to a Modern Data Platform

— Michael Ghen —  
Benefits Data Trust

---

---

# Big Data Sherpa, exploring the edges of info tech

## Academic

- B.S. Computer Engineer ('14)
  - Penn State
- M.S. Strategic Analytics ('16)
  - Brandeis University
- Adjunct Instructor
  - Saint Joseph's University

## Professional

- Systems Engineer
  - Brandeis University
- Data Platform Engineer
  - Cohealo, Inc.
- Data Science Project Manager
  - Benefits Data Trust

## Personal

- Entrepreneurship
- Python
- Computers

# Overview

Transitioning to a Modern  
Data Platform

Data Science at BDT

Legacy Data Science Infrastructure

Framework for a Modern Data Platform

BDT Data Platform Infrastructure

BDT Data Platform Operations

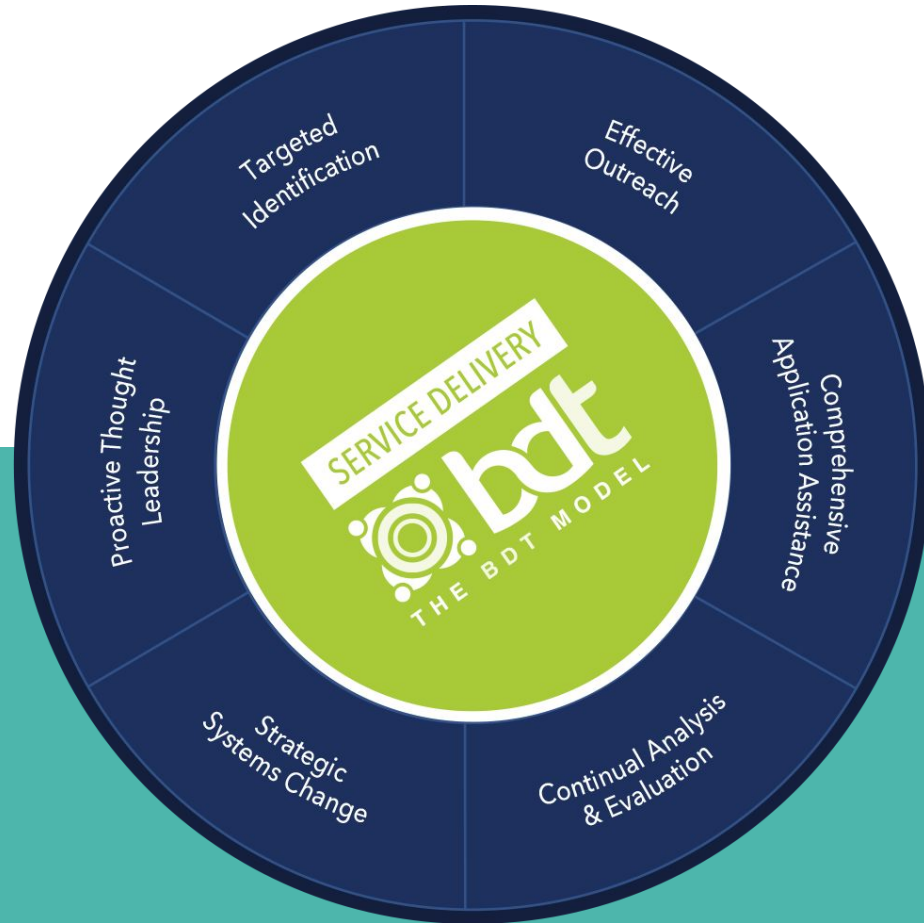
Questions

---

# Benefits Data Trust?

BDT seeks to make benefits access more simple, comprehensive, and cost-effective through **policy and systems change**

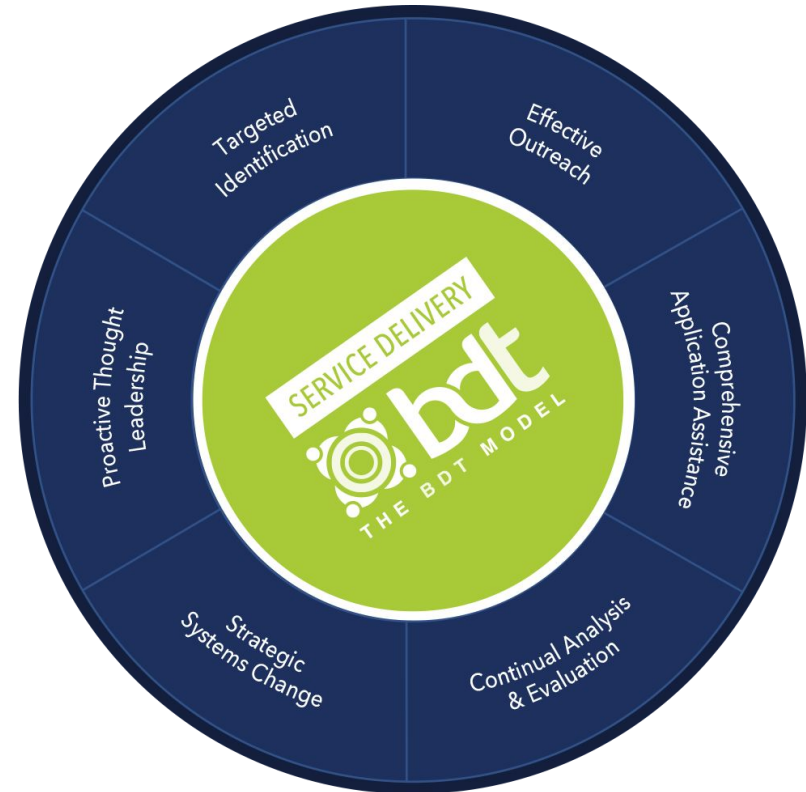
**BDT builds comprehensive and person-centered programs that connect individuals to all the benefits and services they need, while treating each person with dignity and respect**



# Data Science at Benefits Data Trust: data-driven initiatives across multiple states to connect low-income individuals to critical benefit programs

## *Main Things:*

- **Mail Outreach:** Process data needed to run outreach operations
- **Outcomes Reporting:** Summarize and report impact of our work on the clients and communities we serve
- **Internal Reporting:** Extract insights from data to empower tactical and strategic decision making



# Data Science at Benefits Data Trust: data-driven initiatives across multiple states to connect low-income individuals to critical benefit programs

## *Main Things:*

- **Mail Outreach:** Process data needed to run outreach operations
- **Outcomes Reporting:** Summarize and report impact of our work on the clients and communities we serve
- **Internal Reporting:** Extract insights from data to empower tactical and strategic decision making

## *New Main Things*

- Strategic Data Mining
- A/B Testing
- Machine Learning
- Automation
- Simulation

# Data Science at Benefits Data Trust: data-driven initiatives across multiple states to connect low-income individuals to critical benefit programs

## *New Programs*

- **Nudges:** Text message and outbound call campaigns to encourage clients to apply or finish applying for benefits
- **Software as a Service:** Sharing our proprietary software with partners
- **Social Determinants of Health:** Make sure patients are accessing benefits as part of their healthcare

## *New Main Things*

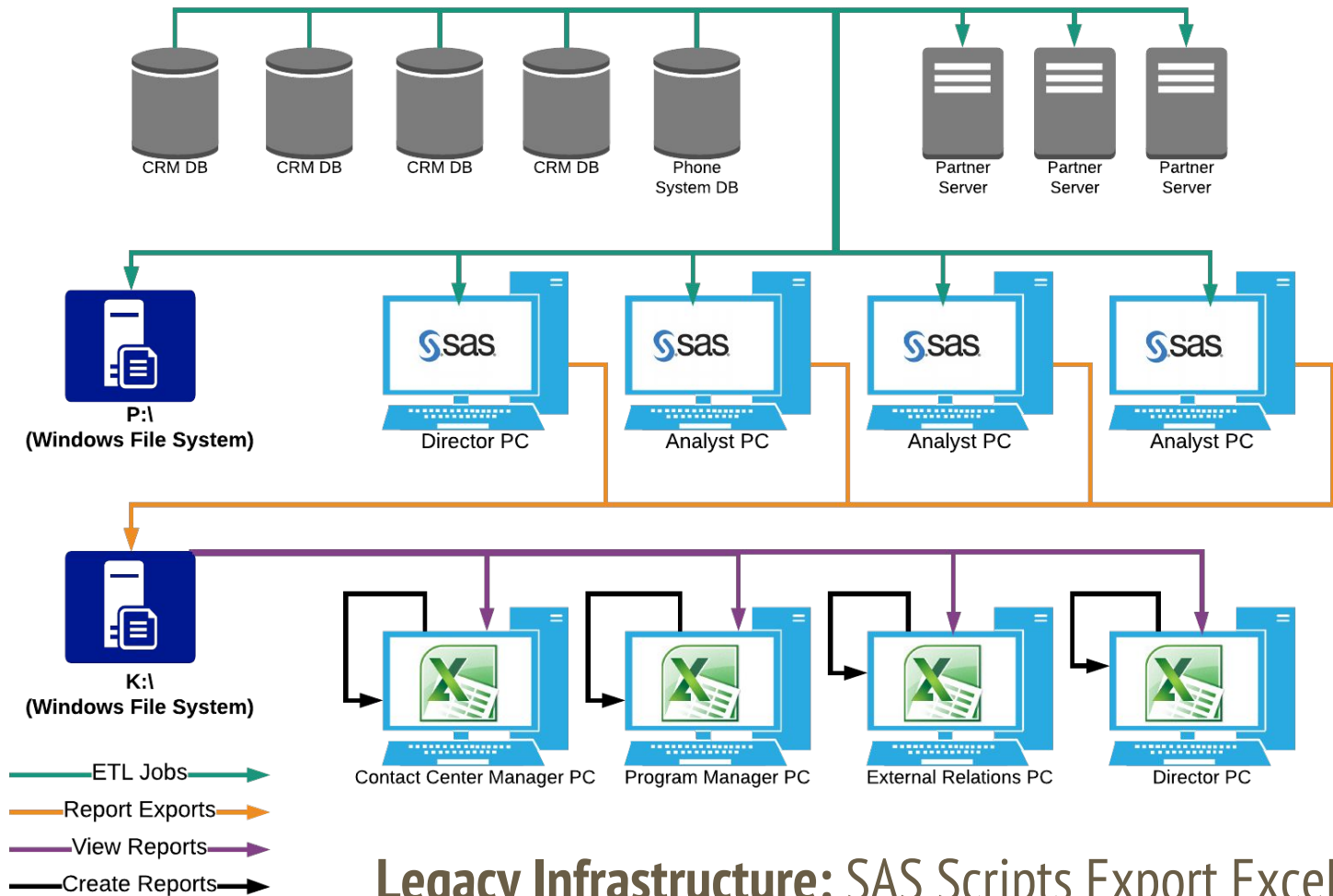
- Strategic Data Mining
- A/B Testing
- Machine Learning
- Automation
- Simulation

---

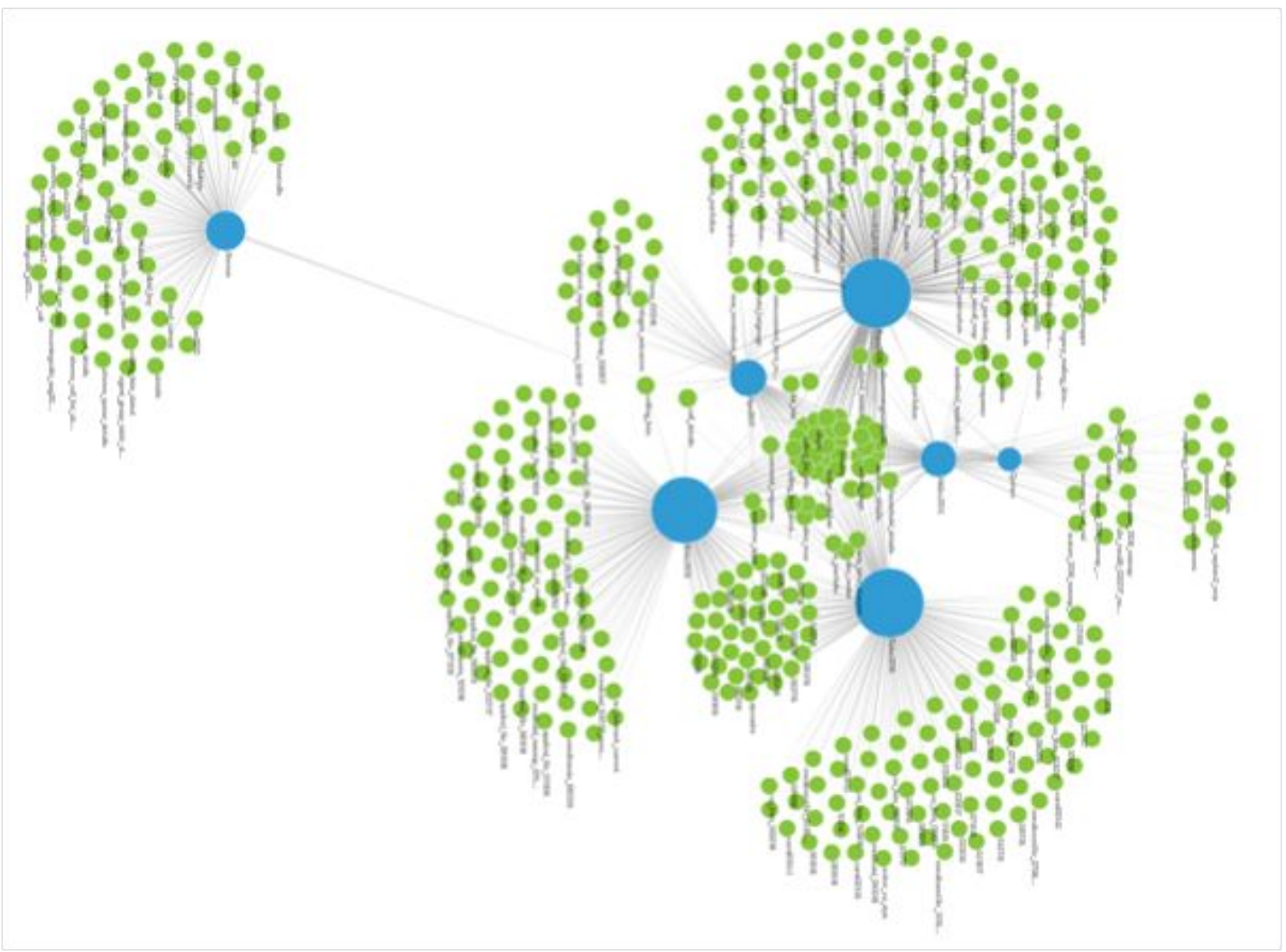
# Challenges with the Legacy System

Michael Ghen  
Benefits Data Trust

---



## Legacy Infrastructure: SAS Scripts Export Excel Reports



# Challenges stem from the tools we are using but also from the processes and culture we operate within

## Ambiguity

- What outcomes are we concern with? How they are measured?

## Centralization

- Data Science serves as the single point of contact between everyone and the answers to questions

## Repetition

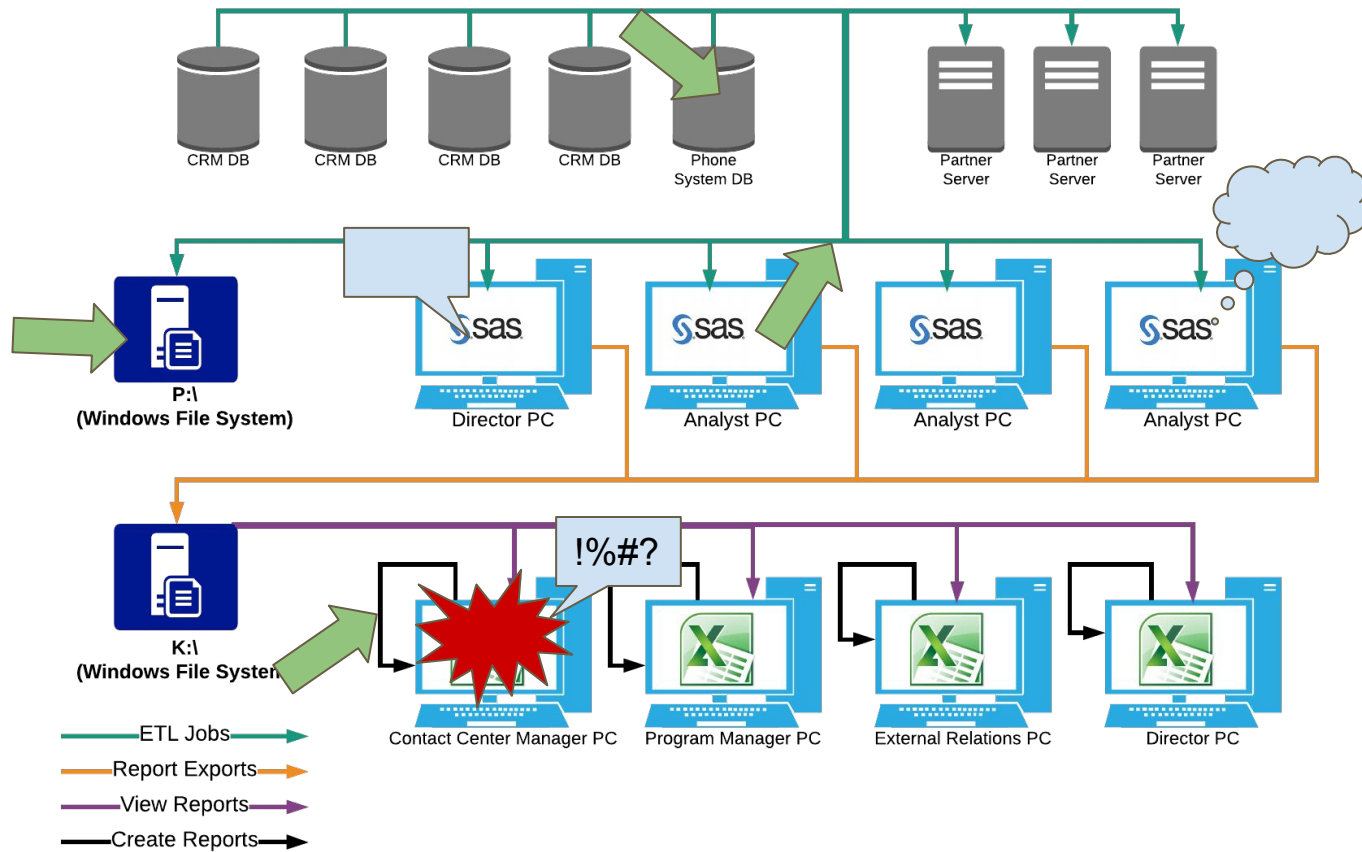
- Solution to a problem is recycled or slightly modified; small defects snowball

## Myopic

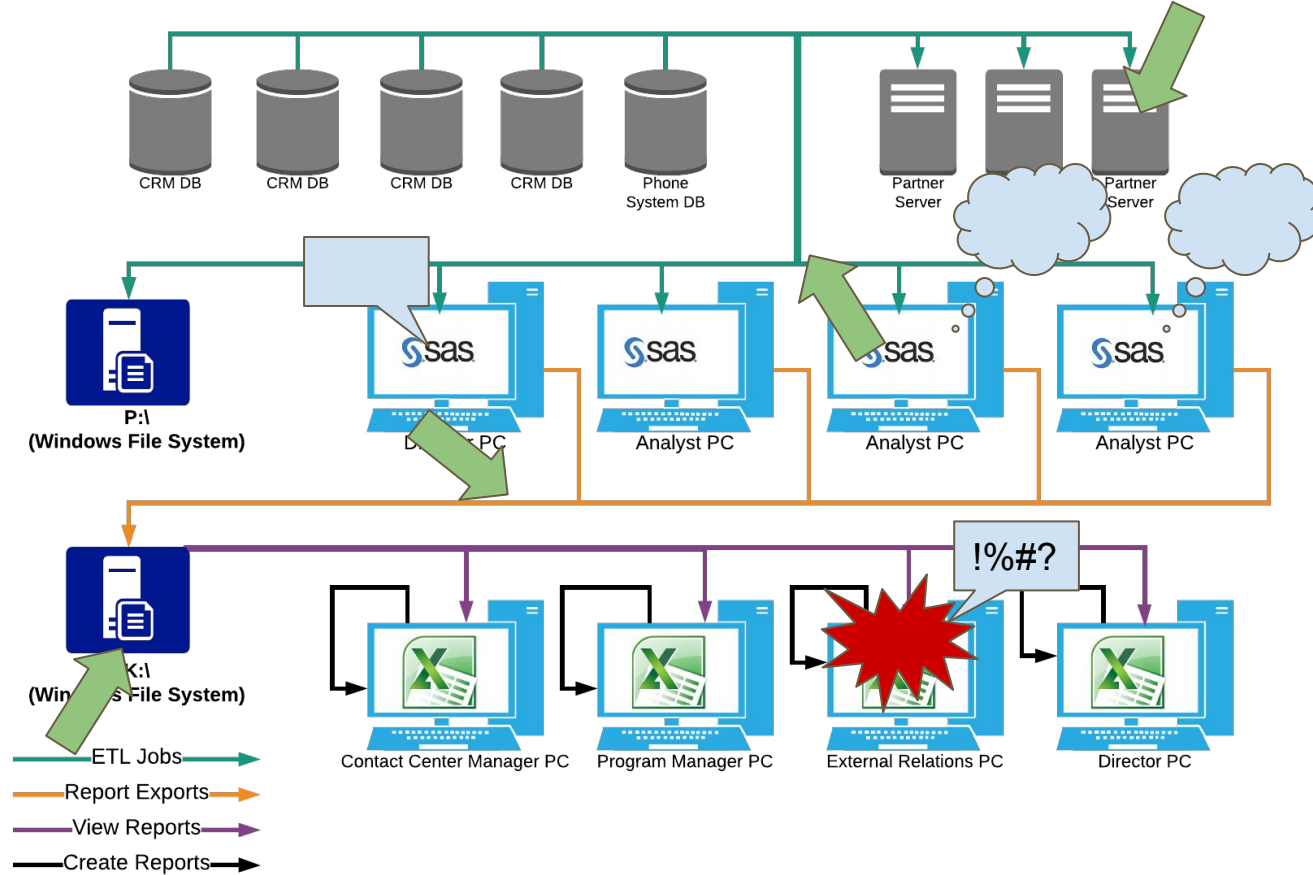
- Data Science is formalized as a final step; conversations about the optimal solution are deferred

## Primitive

- The tactics used, like creating multiple flat files to service requests, are primitive; do not scale well



**Unreliable:** CC Manager finds an error, Analyst runs it down



**Confusion:** External Relations wonders why the numbers are different than the numbers provided last year, Analyst runs it down

---

---

# Framework for a Modern Data Platform

— How do we support the  
organization's data needs? —

---

---

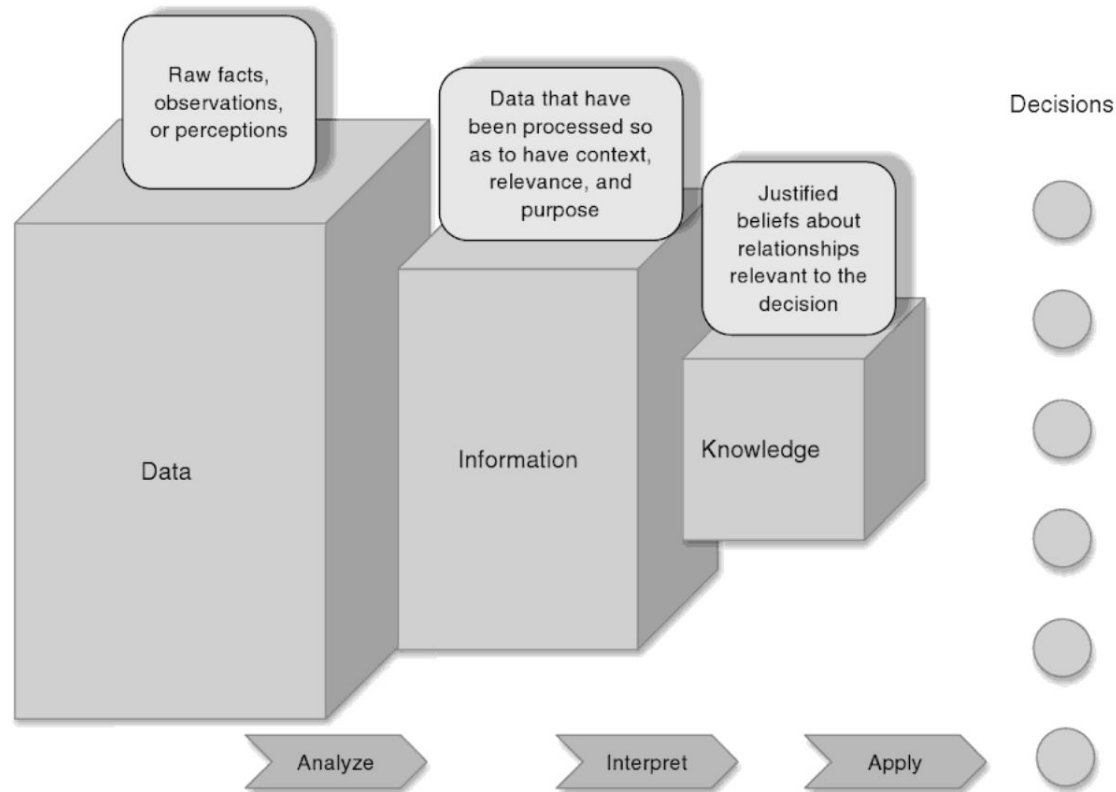
# Data Platforms are composed of technology, process, and people

- Definitions:
  - **Platform:** A platform is a group of technologies, processes, and people that are used as a base upon which other applications, processes or technologies are developed.

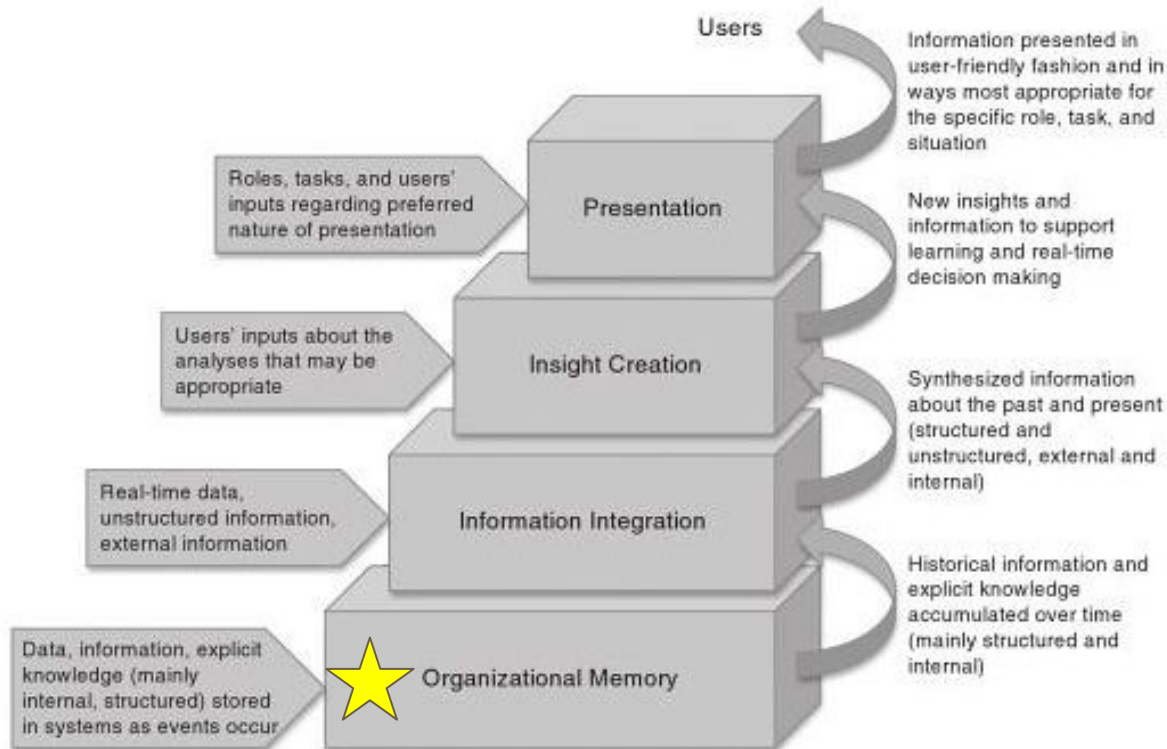
**Data Platform:** A group of technologies, processes, and people that are used as the base upon which data-driven application, processes, and technologies are developed upon.

- The real challenge was trying to establish a vision for what our infrastructure should look like

# Oil Refinery Analogy: Data, Information, Knowledge



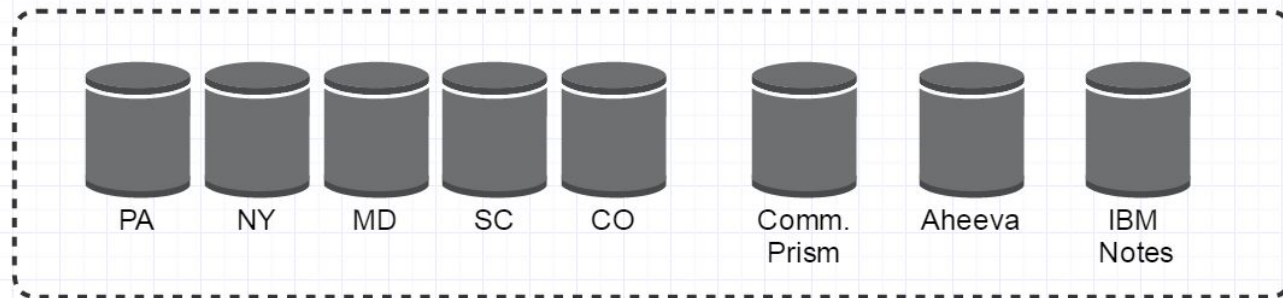
# Understand theory and put it into practice



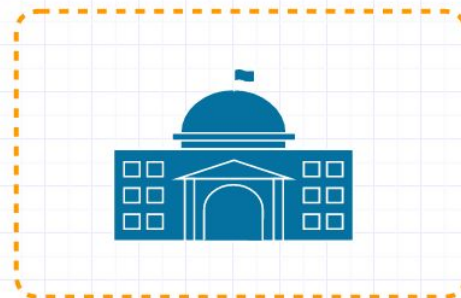
## Organizational Memory

- Represents an organization's accumulated history, including data, information, and knowledge
- Focuses on the storage of intellectual sources (data, information, and explicit knowledge) in such form that they can later be accessed and used
- **Capabilities**
  - Operational Databases
  - Data Lake
  - Data Warehouse
  - Knowledge Repositories

# Internal Databases, External Data from Partners

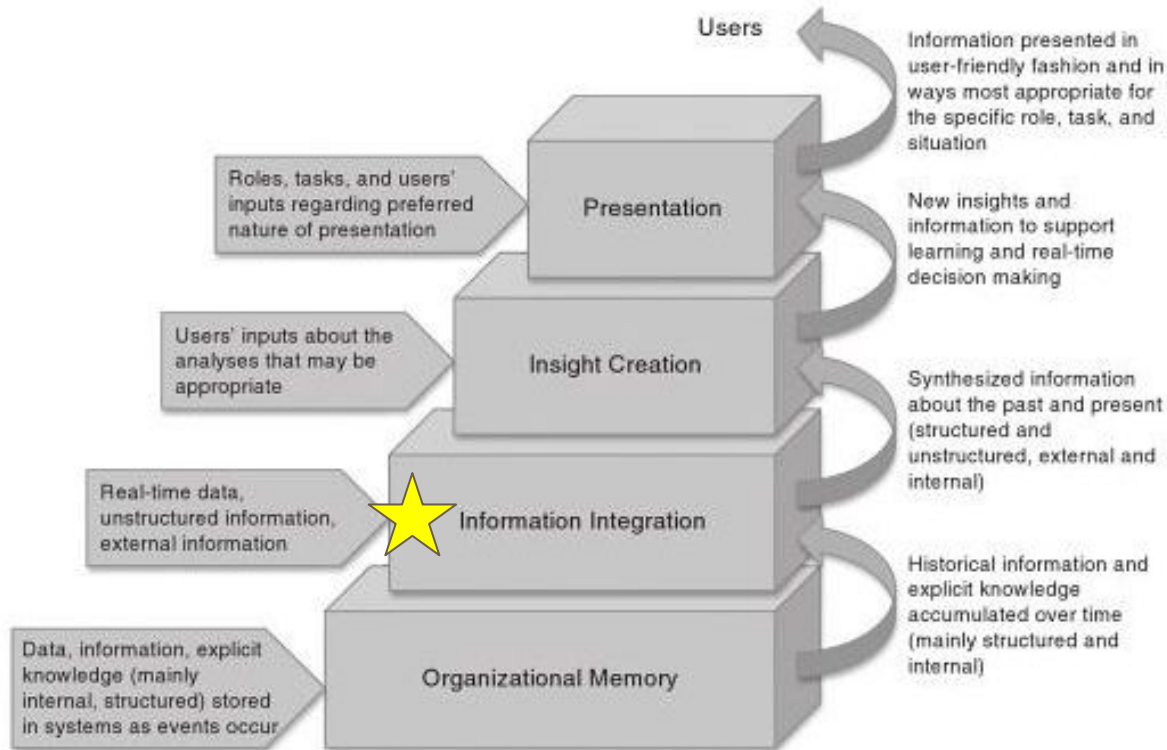


Data Center



Partners

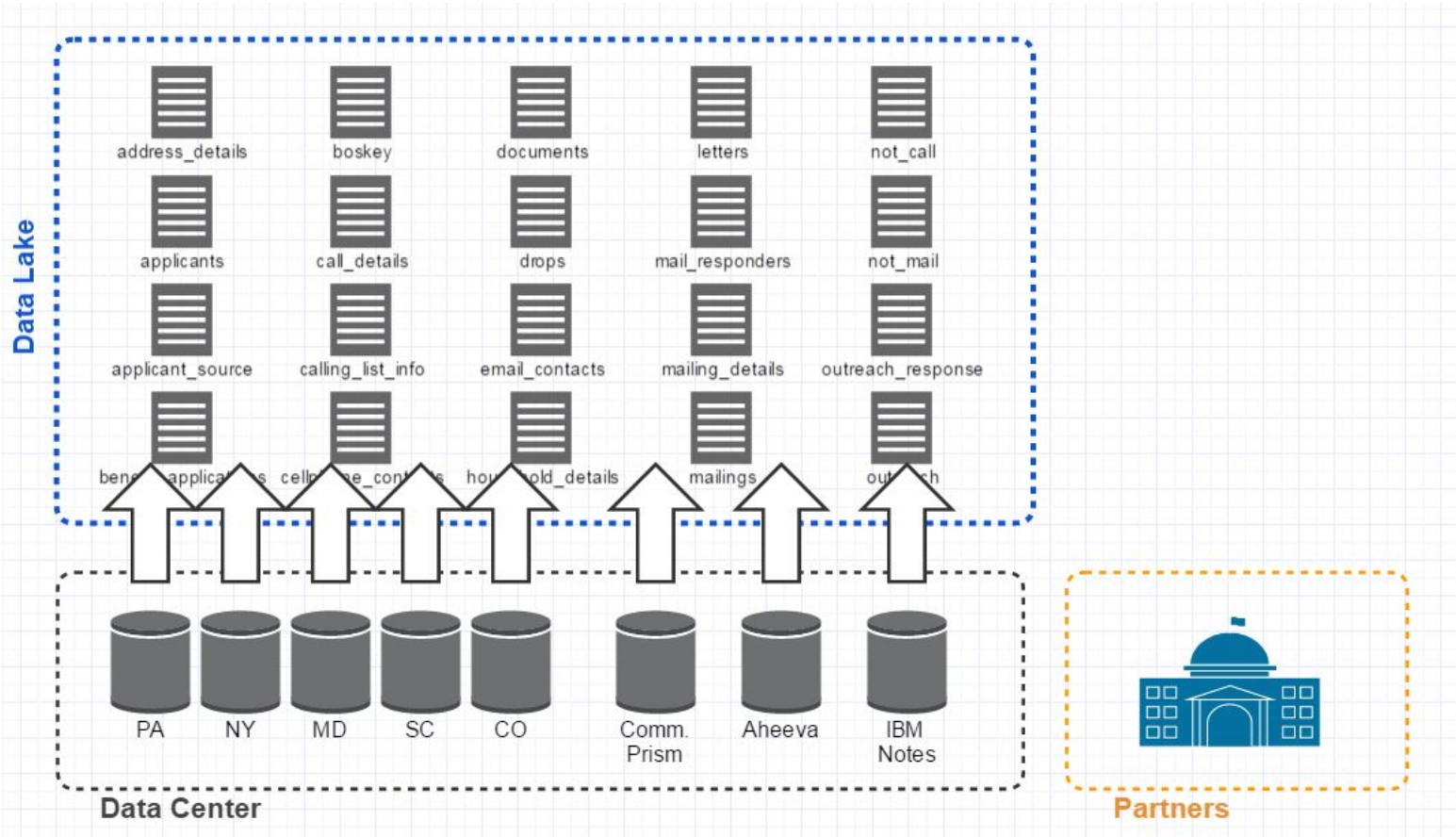
# Understand theory and put it into practice



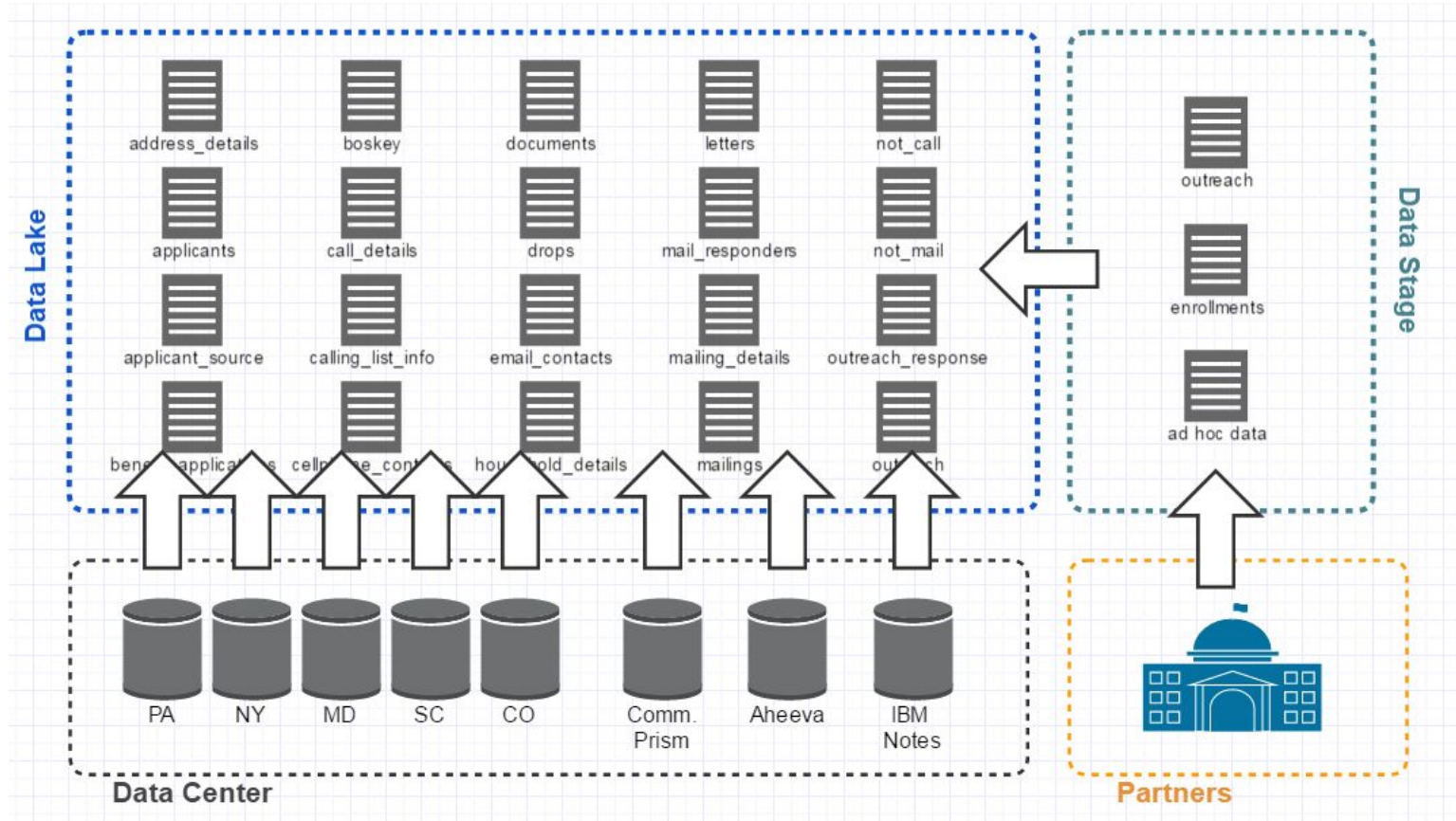
## Information Integration

- Represents the ability to link past structured and unstructured content from a variety of sources that comprise organizational memory with the new, real-time, content
- **Capabilities**
  - Integrating External/Internal Structured Data
  - Integrating External/Internal Unstructured Data
  - Environmental Scanning
  - Text/Web Mining

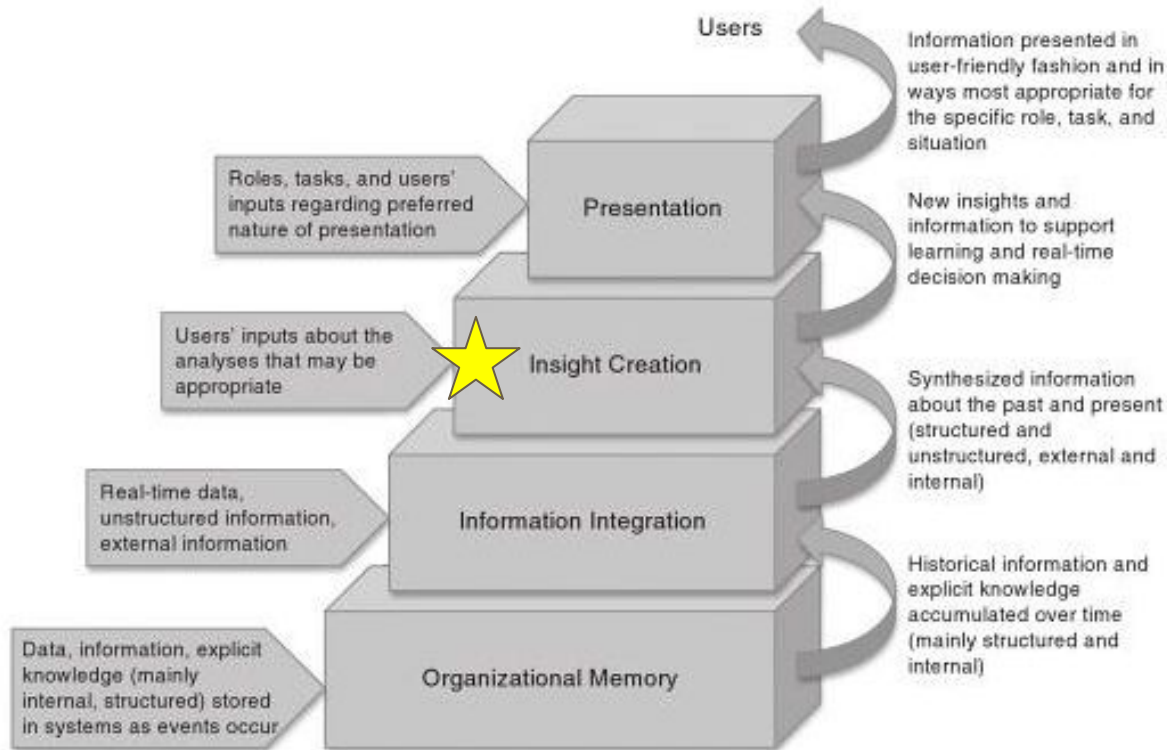
# All information flows into a central location



# All information flows into a central location



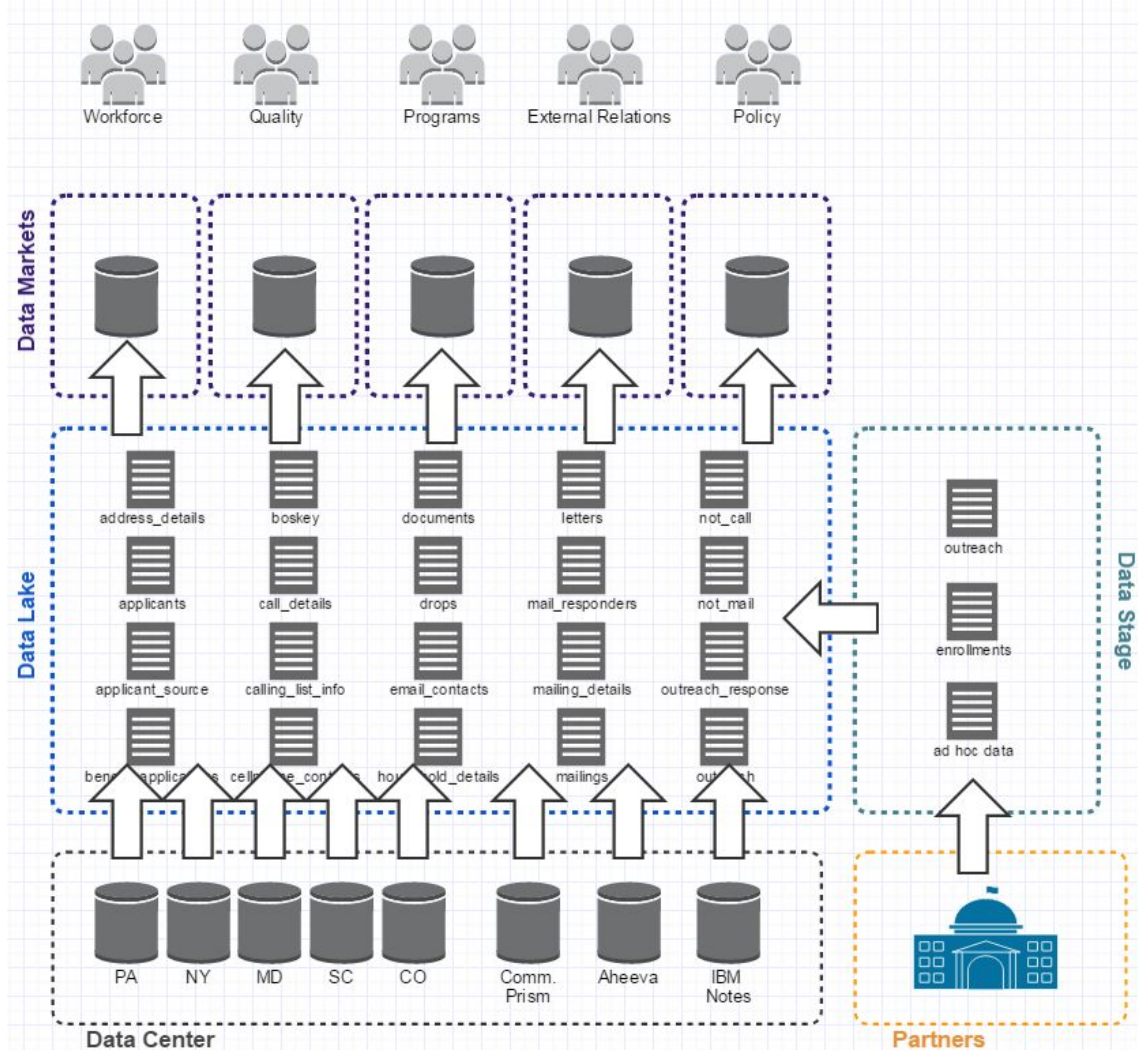
# Understand theory and put it into practice



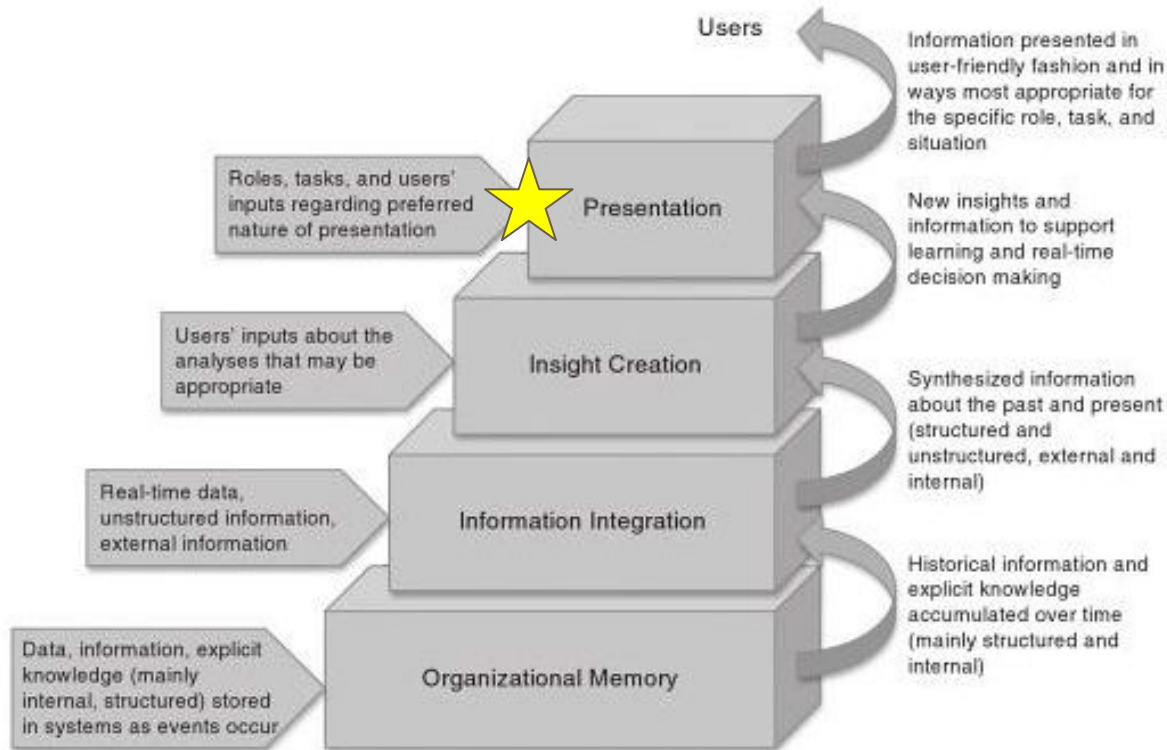
## Insight Creation

- Focuses on the utilization of “raw materials” to produce valuable new insights and enable effective decisions making based on continual rather than periodic analysis.
- **Capabilities**
  - Data Mining
  - Business Analytics
  - Real-time Decision Support

Information is  
integrated and  
distilled into  
smaller,  
subject-specific  
data sets



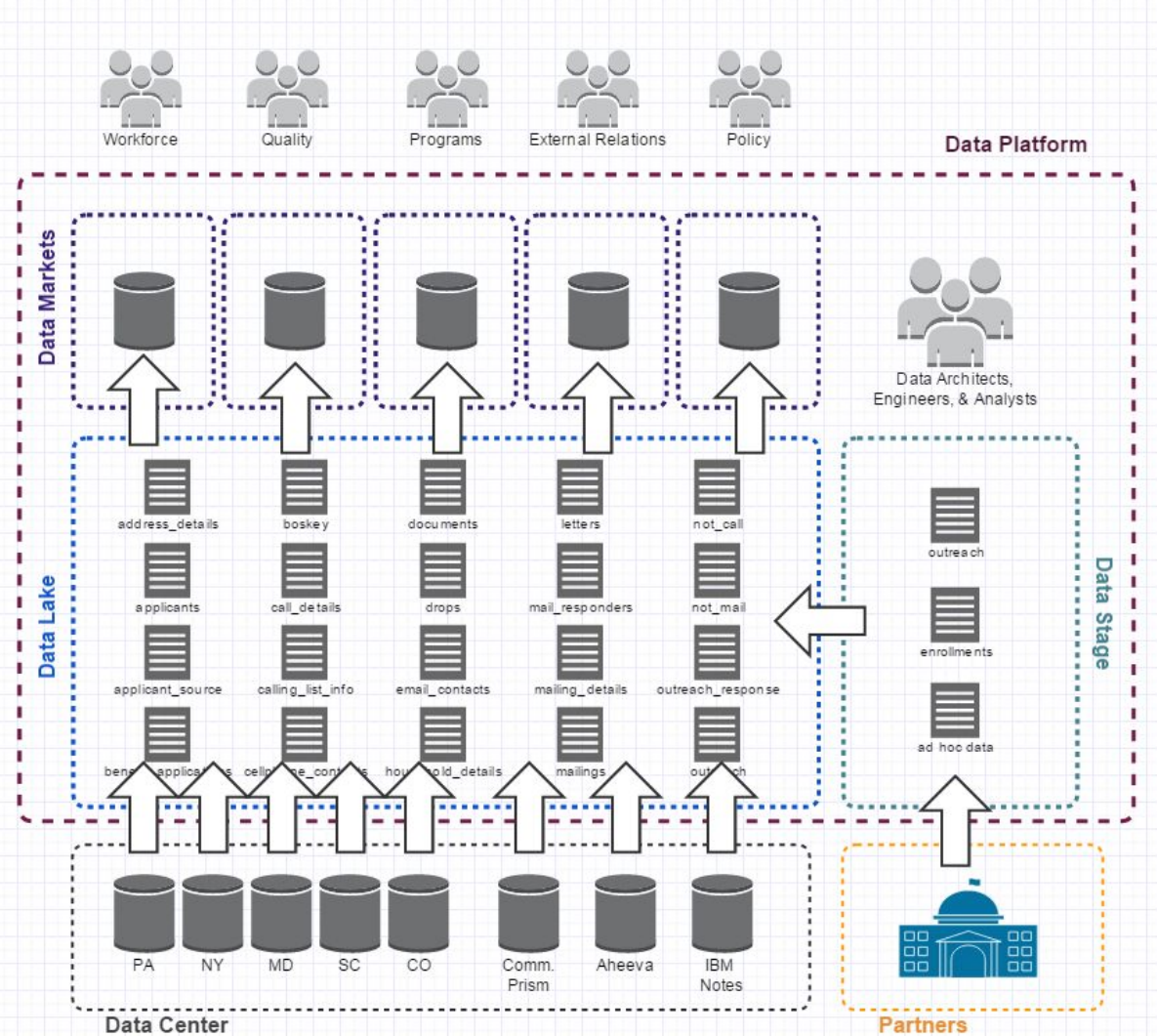
# Understand theory and put it into practice



## Presentation

- The point of contact between BI and the end user
- Focuses on presenting the appropriate information in a user-friendly fashion based on the user's role, the specific task, and the user's inputs regarding the nature of the presentation
- **Capabilities**
  - Enterprise OLAP
  - Visual Analytics
  - Performance Dashboards
  - Scorecards
  - Enterprise Key Performance Indicators

End users are supported by the technologies, processes, and people that support the platform



---

---

# Technologies for Creating a Modern Data Platform

— Business Intelligence Practice —

---

---

# Technology

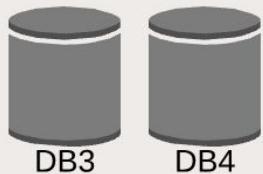
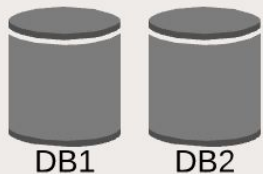
Operations

- Workflow Management with **Airflow**
- Data Warehousing with **BigQuery**
- Visual Analytics with **Looker**

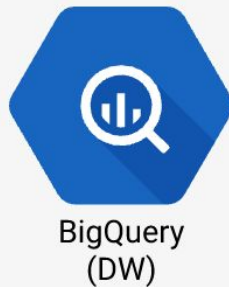
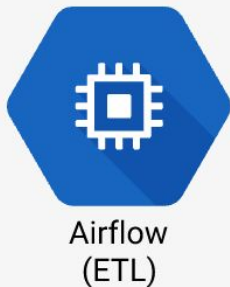
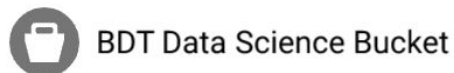
---

# Data Platform

## BDT Data Center



## Google Cloud Platform

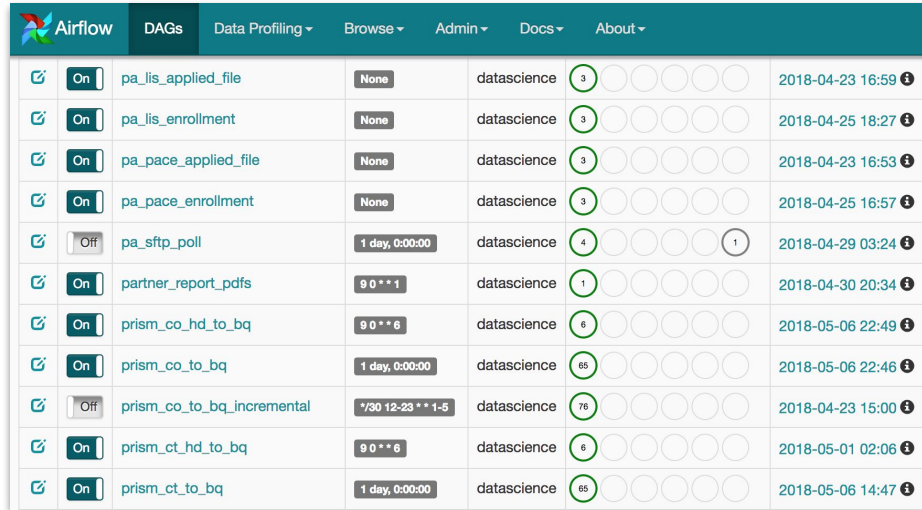


## External Systems



# Apache Airflow manages workflows written in Python

- Why we choose it:
  - Workflows are defined as code, are maintainable, versionable, testable
  - Uses a python programming model
- How we use it:
  - ELT Jobs; Scheduled Reporting
  - Application Integrations (e.g. SFTP to Ruby on Rails applications)



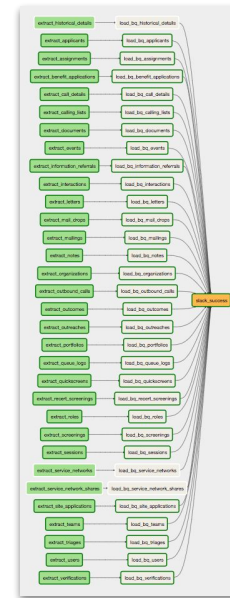
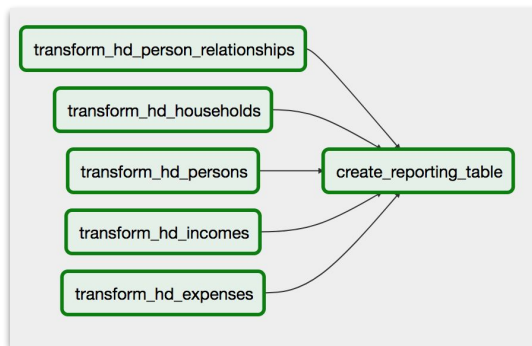
The screenshot shows the Apache Airflow web interface. The top navigation bar includes the Airflow logo and links for DAGs, Data Profiling, Browse, Admin, Docs, and About. Below the navigation bar is a table listing various DAGs. Each row contains a status icon (On/Off), the DAG name, a trigger type (None or a cron expression), the owner (datascience), a progress indicator (a circle with a number inside), a visual progress bar, and the last execution time with a status icon.

Status	DAG Name	Trigger	Owner	Progress	Last Execution
On	pa_lis_applied_file	None	datascience	3	2018-04-23 16:59
On	pa_lis_enrollment	None	datascience	3	2018-04-25 18:27
On	pa_pace_applied_file	None	datascience	3	2018-04-23 16:53
On	pa_pace_enrollment	None	datascience	3	2018-04-25 16:57
Off	pa_sftp_poll	1 day, 0:00:00	datascience	4	2018-04-29 03:24
On	partner_report_pdfs	9 0 * * 1	datascience	1	2018-04-30 20:34
On	prism_co_hd_to_bq	9 0 * * 6	datascience	6	2018-05-06 22:49
On	prism_co_to_bq	1 day, 0:00:00	datascience	65	2018-05-06 22:46
Off	prism_co_to_bq_incremental	* / 30 12-23 * * 1-5	datascience	76	2018-04-23 15:00
On	prism_ct_hd_to_bq	9 0 * * 6	datascience	6	2018-05-01 02:06
On	prism_ct_to_bq	1 day, 0:00:00	datascience	65	2018-05-06 14:47

Airflow is a platform to programmatically author, schedule and monitor workflows  
Airflow is used to author workflows as directed acyclic graphs (DAGs) of tasks

# Apache Airflow manages workflows written in Python

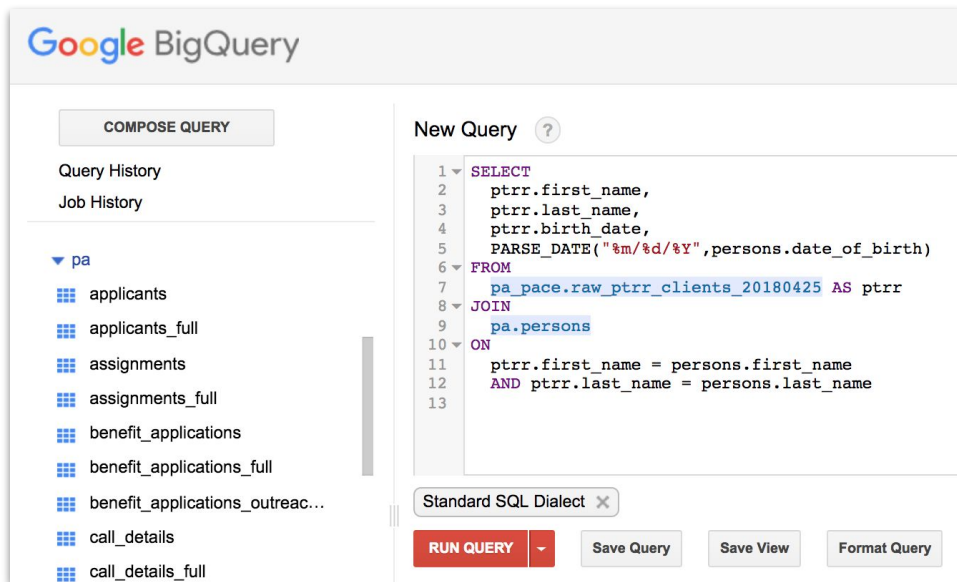
- Challenges:
  - Deployment, management
  - Testing workflows
  - Training analysts
- Advantages:
  - Everything is done with Python
  - Logging, monitoring
  - Simple to make improvements to the underlying codebase if needed



**Airflow is a platform to programmatically author, schedule and monitor workflows**  
**Airflow is used to author workflows as directed acyclic graphs (DAGs) of tasks**

# BigQuery allows us to access all our data with SQL

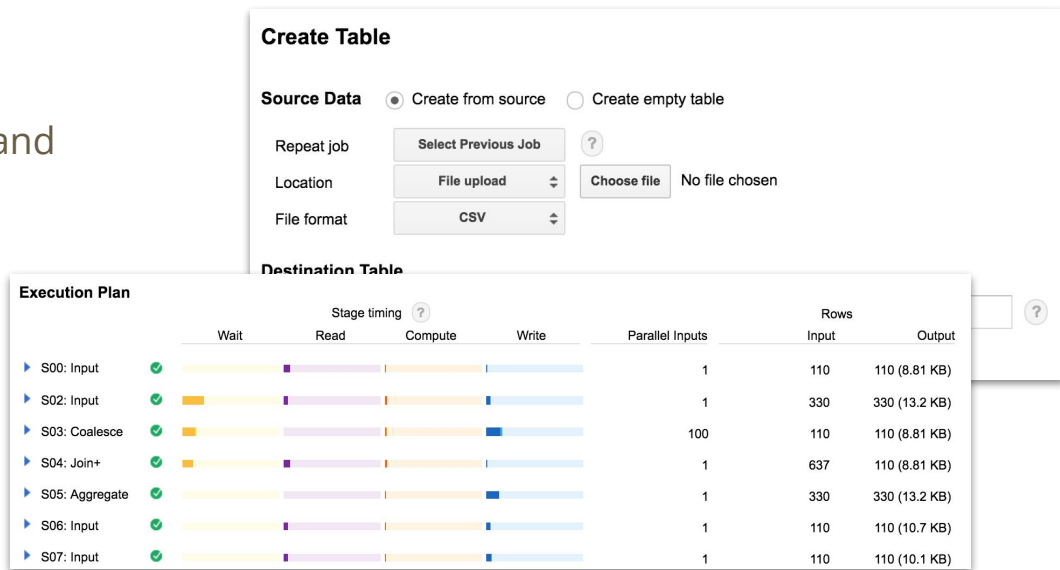
- Why we choose it:
  - Fully managed, always available
  - Queries are just fast
  - Easy import/export
  - Dirt Cheap
- How we use it:
  - Data Warehousing
  - All data flows into BigQuery, all reports based on BigQuery
  - Extract, Load, Transform



**BigQuery is Google's serverless, highly scalable, low cost enterprise data warehouse designed to make our data scientists productive**

# BigQuery allows us to access all our data with SQL

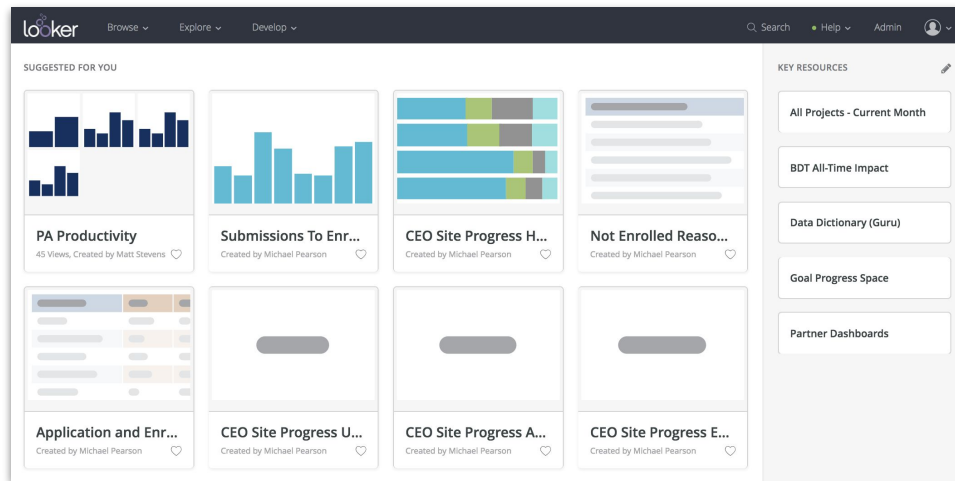
- Challenges:
  - Still requires data that is clean and structured before import
  - Does not support transactions
  - Requires strong SQL skills
- Advantages:
  - No database administration
  - No servers
  - Leverages SQL
  - Great documentation
  - Strong Software Development Kits



BigQuery is Google's serverless, highly scalable, low cost enterprise data warehouse designed to make our data scientists productive

# Looker provides control over how we are reporting

- Why we choose it:
  - Data model as code, version control, transparency with respect to data model/etl changes
  - Centralized data definitions
  - Usage tracking
- How we use it:
  - Dashboards
  - Internal and external reporting
  - Embedded reporting



**Looker is a business intelligence software and big data analytics platform that helps you explore, analyze and share real-time business analytics easily.**

# Looker provides control over how we are reporting

- Challenges:
  - Server-side/in-memory transformations, slow queries
  - Setting up drill-down
- Advantages:
  - Easy to troubleshoot (everything is SQL)
  - Dimensions and measures are centralized, sharable
  - We like to code, we like version control

The screenshot displays the Looker 'Explore' interface. The top navigation bar includes 'looker', 'Browse', 'Explore', and 'Develop' menus, along with search, help, and admin links. The main header shows 'Explore' and '56 rows · 2.5s · 1m ago' with a 'Run' button. The left sidebar contains a 'Reporting' section with a search bar and tabs for 'All Fields', 'Dimensions', and 'Measures'. Below this is a 'Close Outs' section for 'Co Benefit Applications' with a 'Closed Date' dimension. The main area shows a table with columns: 'Co Benefit Applications Created Date', 'Co Benefit Applications Completed Date', 'Co Benefit Applications Submitted Date', and 'Co Benefit Applications Closed Date'. A 'co\_enrollments' dimension is selected, and its SQL is shown in a code editor overlay.

```
122 dimension: expedited {
123   description: "Yes if the clients application was expedited in PRISM, no otherwise"
124   type: yesno
125   sql: ${TABLE}.expedited = 1;;
126 }
127
128 dimension: application_type {
129   description: "How the client's application was processed by BDT"
130   type: string
131   case: {
132     when: {
133       label: "Expedited - Complete"
134       sql: ${completely_submitted} = True and ${expedited} = True ;;
135     }
136   }
```

Looker is a business intelligence software and big data analytics platform that helps you explore, analyze and share real-time business analytics easily.

---

---

# Final Remarks

— Organizational Challenges  
& Benefits —

---

---

# Things got worse before they got better

- Departments need to own some of the analysts' tasks
  - Data Science needs to provide the tooling to support self-services
- Departments need data literacy and technical training
  - Data Science needs to be the experts supporting this
- IT needs to provide the infrastructure to support analytics technologies
  - Data Science needs to have the expertise to run these technologies
- Programs need to provide clarity about what they will report on
  - Data Science needs to seek clarity before building reporting

# Thank you!

