Interagency Data Team

October 21, 2021

https://octo.dc.gov/page/interagency-data-team

Agenda

Welcome, News & Updates
Michael Bentivegna
Program Manager, Data Visualization & Analysis

DC Health's Disease Surveillance Data Engine
Cindy Knoll
Accenture Engagement Lead

Johnny Nguyen
Accenture Data Engine Technical Architect and Application Lead

Data Lake Upgrade
Matthew Sokol
Program Manager, Big Data & Data Integration

Data Report
Matthew Crossett
Enterprise Mapping Project Manager

Business Intelligence

Tableau System Versions, Planned Upgrades & Site Administrators Meeting

![tableau • octo — GOVERNMENT OF THE DISTRICT OF COLUMBIA logo]

- **Internal Production**
  - Version 2020.4.2
- **Public Production**
  - Version 2020.4.2
- **Test/Dev**
  - Version 2021.2.3
  - Secure Auth
  - Data Lake Auth.
- **Events**
  - Tableau conference on November 9-11
 - Free for DC Gov.
  - **Register here**

- **Site Administrator Quarterly Meet**
  - November 5, 2021 (tentative)
- **Planned Upgrades**
  - Upgrade to Version 2021.2.3 10/28/21
    - Internal 6 pm-12 am
      - 4 hr. outage (8-12)
    - External 8 pm-12 pm
      - 2 hr. outage (8-10)
  - Creators should save all work and log off by 6 pm, Thursday October 28.

Business Intelligence

What new features will be available with a server upgrade to **version 2021.2.3**?

- **Tableau Desktop**
  - **Map layer control and area calculations**
  - **Ask Data integration into dashboards**
  - **Quick level of detail calculations**

- **Tableau Server**
  - **Web authoring formatting for filters, parameters, and legends**
  - **Unified notification experience**

- **Tableau Prep**
  - **Bulk rename in web authoring**
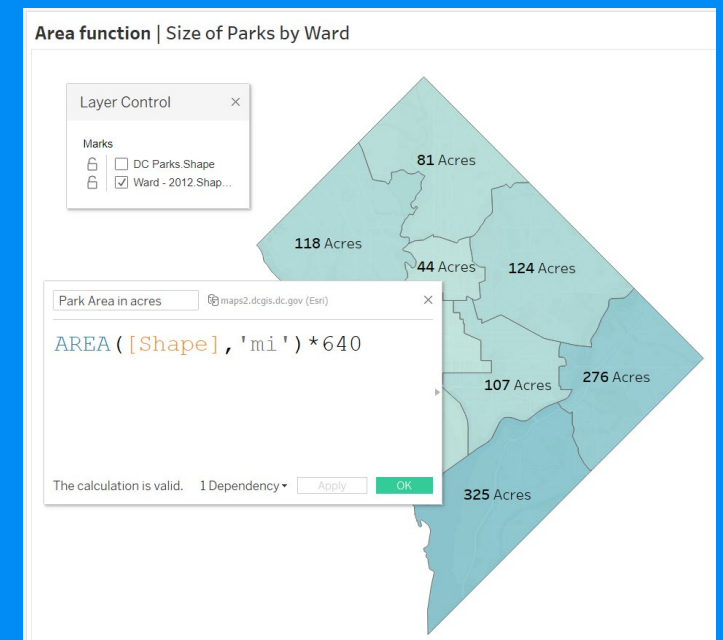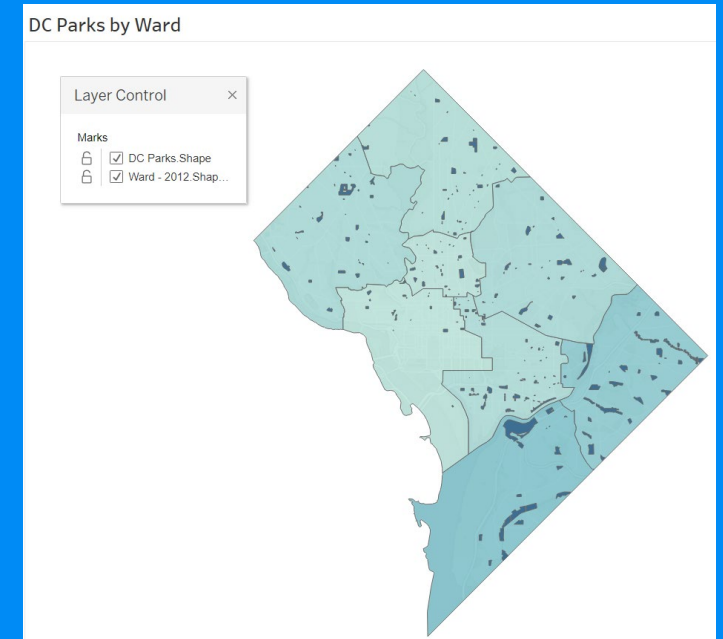
Business Intelligence

MicroStrategy System Versions, Planned Upgrades & Stake Holder Meeting

MicroStrategy® **octo**
GOVERNMENT OF THE DISTRICT OF COLUMBIA

- **Production Server**
  - Version 2020 Update 2
  - Deploying new physical servers
  - Upgrade to Version 2021.3 10/22/21
    - Internal 5 pm-12 am
      - 4 hr outage (5-9)
    - External 8 pm-12 pm
      - 2 hr outage (8-10)
- **Lower Environment**
  - Upgrade to Version 2021.3
    10/21/21 9 am –1 pm
  - Enterprise/Self Service
  - Deploying physical IS server for Development

- **Public/Sharepoint**
  - COVID-19 Public Dashboard

  - Recovery Plan Dashboards
  - Back to School Dashboard

- **Stake Holders Meeting**
  - November 4th 2:00-4:00 pm

Big Data & Data Integration

System Versions, Planned Upgrades & Projects

https://data.in.dc.gov

- **Middleware Team**
  - SOA - 5 million transactions
  - API GW – 20 million transaction
- **API Gateway Upgrade**
  - 9.4 --> 10 in Q2
  - Additional API Management tools
  - Supported OS
- **Data In DC**
  - Now updated with latest EDI
  - Searchable by dataset name, level, agency, PII

- **Cloudera Platform Upgrade**
  - Enhanced Security
  - Easier Kerberos Authentication
  - Apache Spark 2.x
  - Phase 1 Upgrade – Completed 10/9
  - Phase 2 Upgrade – TBD
- **Data Lake**
  - Docs for R-Studio, Excel, PowerBI
  - More at https://octo.in.dc.gov/node/47271
- **Data Anonymization**
  - Flat Files
  - Data Warehouse
  - Data Lake

Geographic Information Systems

DC GIS System Versions, Planned Upgrades & Data Highlights

Open Data DC

- **ArcGIS Server Production**
  - Currently @ 10.8.1 for Maps2
- **ArcGIS Portal**
  - Currently @ 10.8.1
- **Esri GeoEvent**
  - Internal Version 10.8.1
  - External Version Going to 10.8.1 late Sept
  - Integrated with Data Lake
- **New Hardware**
  - Updated Servers for Maps2 from OCTO Data Center Migration

- **Annual Ortho & Planimetric**
  - Flight completed in one lift on March 11th
  - Ortho's will be available first week in Nov, Planimetrics undergoing QC. Expected Release: End of Nov.
- **Pictometry Aerial Imagery**
  - Flown 100% Jan-March 2021
  - Data available via https://explorer.pictometry.com/login.php to Login-in request
- **Vaccination Sites**
  - https://coronavirus-data-dcgis.hub.arcgis.com

Enterprise Data Inventory

Application Updates and Upcoming meetings

- **EDI open for editing 9/16/21**

- **ADO Meeting**
  - November 3rd @10 AM
  - Will be recorded

- **EDI will be closed to editing mid-December**

- **Main Changes:**
  - Sensitivity Flags
  - Sort functionality is updated
  - Enhanced administrative pages

User Groups

GIS User Group

- October 28st, Topic TBA (Your presentation here?)
- Meetings via WebEx
- Contact Eva.Reid@dc.gov
- **Join the Group**

Data Science @ DC

- TBD
- Meetings via Teams
- Contact Matthew.Sokol@dc.gov

Virtual Training: Tableau

Register for a class:

https://octo.in.dc.gov/service/tableau-training-courses

Join the DC Gov Tableau Community Team!
Collaborate with fellow Creators, Site Administrators and search the Start Here! Wiki

- **Introductory Prep Builder**
  - Prepare your data for analysis in Tableau Desktop with a workflow
    - November 17

- **Introductory Desktop**
  - Create and publish dashboards using DC data
    - November 2-3

- **Intermediate Desktop**
  - Create an executive dashboard and use advanced analytical features
    - November 30 – December 1

Training
Virtual MicroStrategy

MicroStrategy **octo**
GOVERNMENT OF THE DISTRICT OF COLUMBIA

https://octo.in.dc.gov/service/microstrategy-introductory-training

- **Introductory Class**
  - November 16-17
  - December 14-15

- **Intermediate Class**
  - Q2 FY22

- **Architect & Developer Class**

  By invitation only, based on agency need

Training: Geographic Information Systems

All courses require ArcGIS Online Access and/or ArcGIS Pro locally.

https://octo.in.dc.gov/service/enroll-gis-training

- **DC Government GIS Fundamentals**
  - For new staff, managers, people interested in learning about GIS in DC Government
  - Next class: October 27th, 2021
  - Register via PeopleSoft

- **DC Government Learning Paths available via ESRI Training Portal**
  - Employees with ArcGIS Online access
  - View options

- **GIS User Group**
  - Open to all staff
  - Join

Disease Surveillance Data Engine
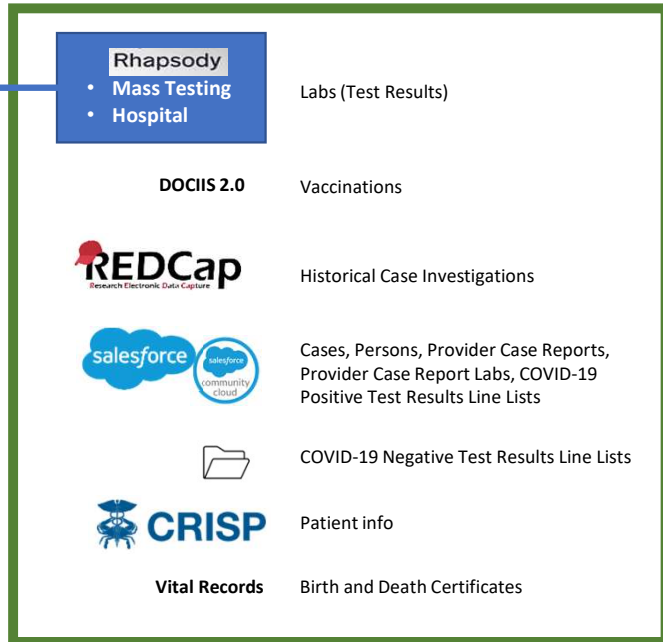## DC Health

October 21, 2021
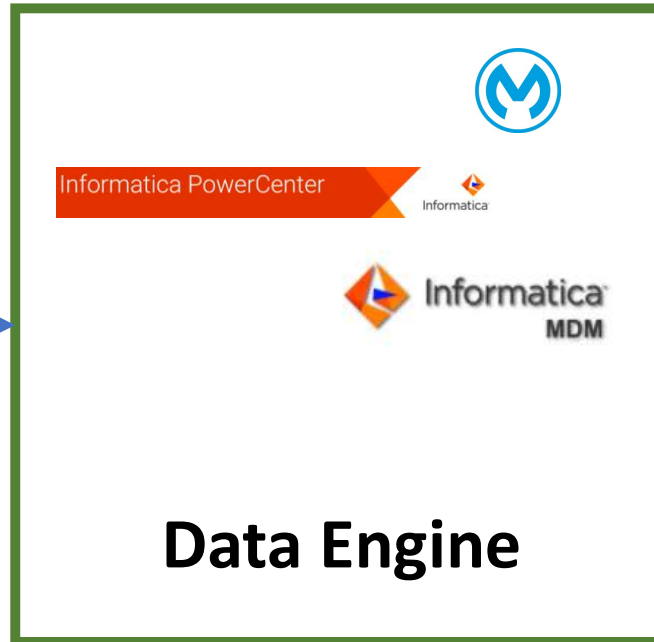
Data Engine Overview

Future state Disease Surveillance combines **agility, flexibility, scalability, and automation.** The increased automation involves building a **highway into the Data Engine**, which will drive enhanced reporting capabilities, result in time savings and help facilitate more informed health policy decisions.

Source Data - Highway Onramp

Rhapsody
- **Mass Testing**
- **Hospital**

Labs (Test Results)

DOCIIS 2.0   Vaccinations

REDCap Research Electronic Data Capture   Historical Case Investigations

salesforce community cloud   Cases, Persons, Provider Case Reports, Provider Case Report Labs, COVID-19 Positive Test Results Line Lists

COVID-19 Negative Test Results Line Lists

CRISP   Patient info

Vital Records   Birth and Death Certificates

Data Repository → **NBS**

Data Ingestion, Cleansing and De-duplication

Informatica PowerCenter

Informatica

Informatica MDM

Data Engine

Highway Offramps

+ableau

360 Views of Resident and Facility (Location) & Workflow Management, e.g. disease investigations and COVID-19 contact tracing

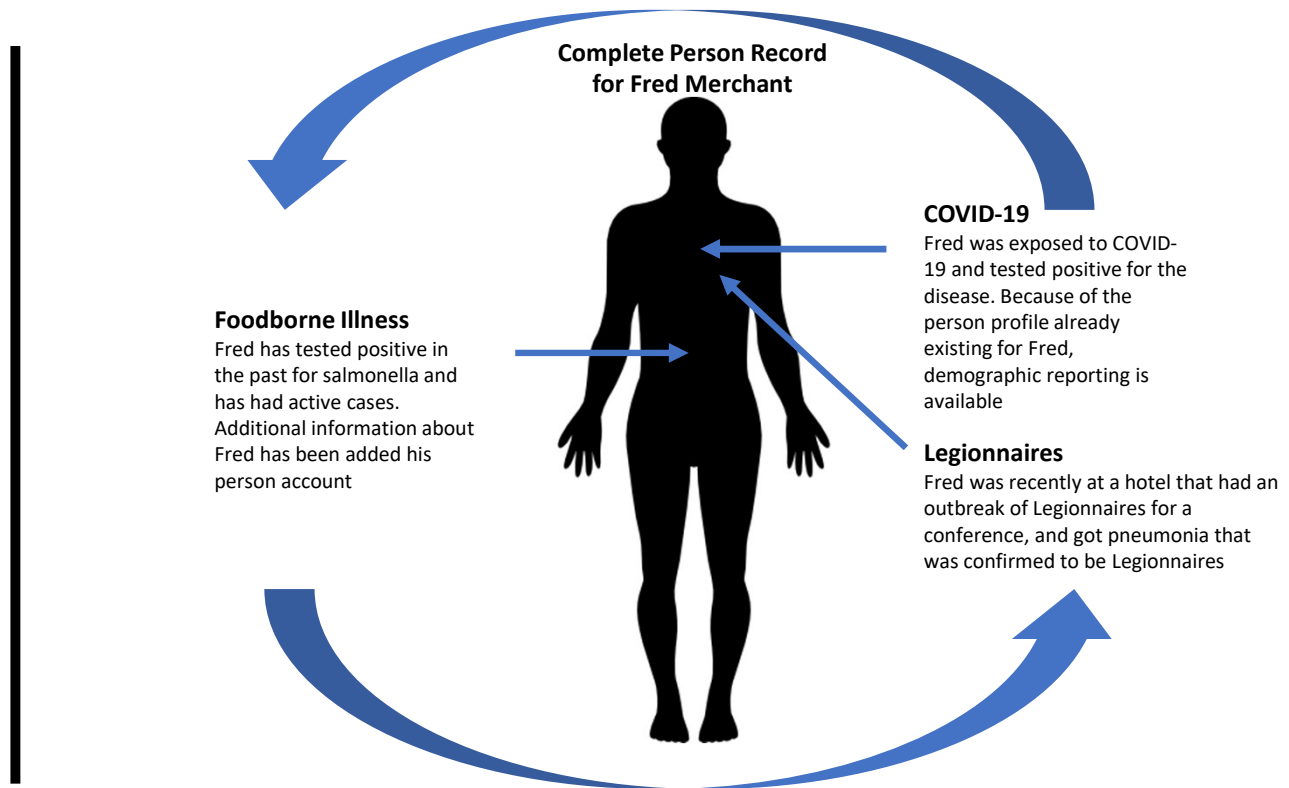salesforce

DC Resident 360 – Future State Reporting

The data engine feature a relational model to bring together data from various sources for diseases. Disease Surveillance will leverage the relational model to bring together all Disease Surveillance in one solution

Resident Persona

Below is fictional demographical information showing an example of what we'll capture and store on an individual

First Name: Fred
Last Name: Merchant
DOB: 9/30/1970
Home Phone Number: 202-555-1212
Address: 2 Lincoln Memorial Way
Sex: Male
Race: Asian

Complete Person Record for Fred Merchant

Foodborne Illness
Fred has tested positive in the past for salmonella and has had active cases. Additional information about Fred has been added his person account

COVID-19
Fred was exposed to COVID-19 and tested positive for the disease. Because of the person profile already existing for Fred, demographic reporting is available

Legionnaires
Fred was recently at a hotel that had an outbreak of Legionnaires for a conference, and got pneumonia that was confirmed to be Legionnaires
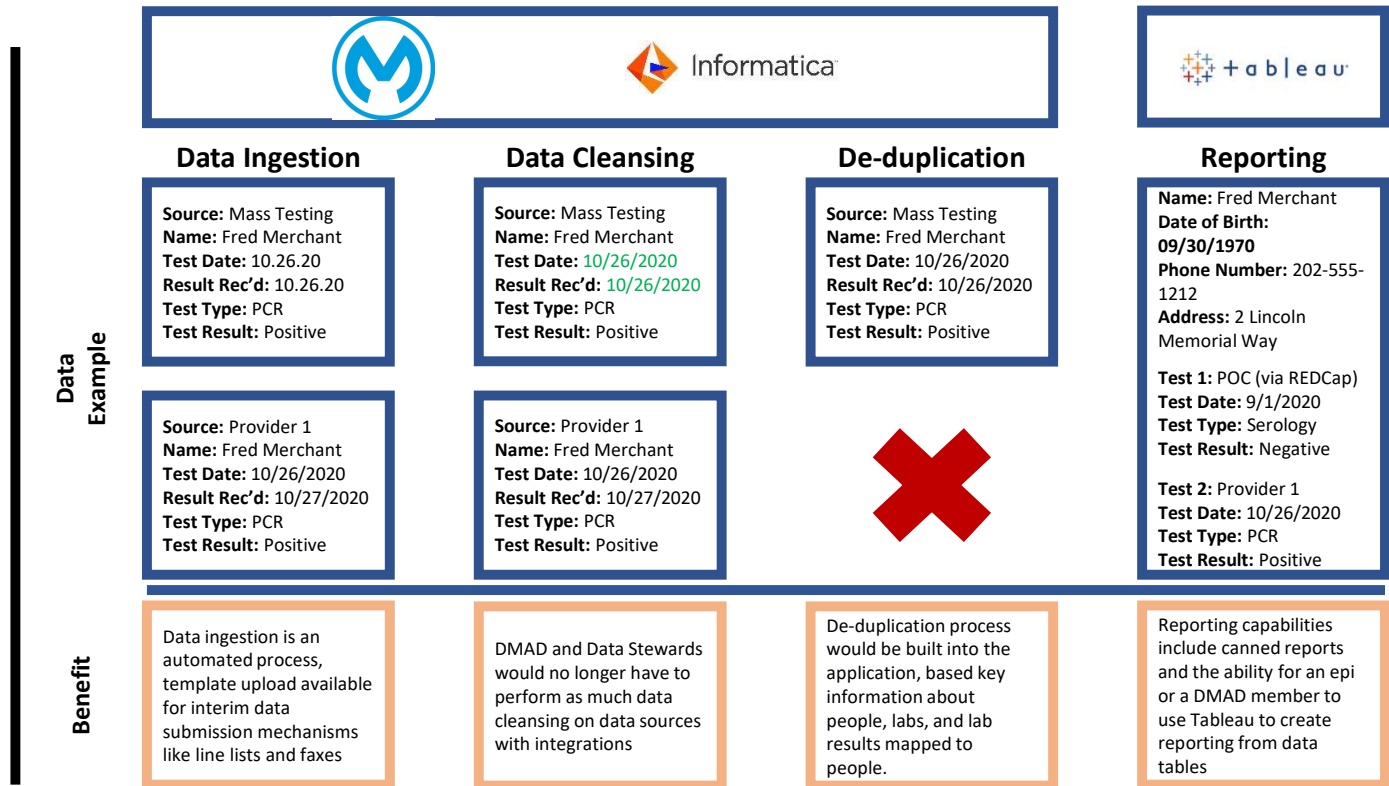
Data Engine Explained

The data engine focuses on **automation, data transformations, and reporting** (at the raw data, or cleansed data level) are available through this solution

Resident Persona

Below is fictional demographical information showing an example of what we'll capture and store on an individual

First Name: Fred
Last Name: Merchant
DOB: 9/30/1970
Home Phone Number: 202-555-1212
Address: 2 Lincoln Memorial Way
Sex: Male
Race: Asian

	Data Ingestion	Data Cleansing	De-duplication	Reporting
Data Example	**Source:** Mass Testing **Name:** Fred Merchant **Test Date:** 10.26.20 **Result Rec'd:** 10.26.20 **Test Type:** PCR **Test Result:** Positive	**Source:** Mass Testing **Name:** Fred Merchant **Test Date:** 10/26/2020 **Result Rec'd:** 10/26/2020 **Test Type:** PCR **Test Result:** Positive	**Source:** Mass Testing **Name:** Fred Merchant **Test Date:** 10/26/2020 **Result Rec'd:** 10/26/2020 **Test Type:** PCR **Test Result:** Positive	**Name:** Fred Merchant **Date of Birth: 09/30/1970** **Phone Number:** 202-555-1212 **Address:** 2 Lincoln Memorial Way **Test 1:** POC (via REDCap) **Test Date:** 9/1/2020 **Test Type:** Serology **Test Result:** Negative **Test 2:** Provider 1 **Test Date:** 10/26/2020 **Test Type:** PCR **Test Result:** Positive
	Source: Provider 1 **Name:** Fred Merchant **Test Date:** 10/26/2020 **Result Rec'd:** 10/27/2020 **Test Type:** PCR **Test Result:** Positive	**Source:** Provider 1 **Name:** Fred Merchant **Test Date:** 10/26/2020 **Result Rec'd:** 10/27/2020 **Test Type:** PCR **Test Result:** Positive	❌	
Benefit	Data ingestion is an automated process, template upload available for interim data submission mechanisms like line lists and faxes	DMAD and Data Stewards would no longer have to perform as much data cleansing on data sources with integrations	De-duplication process would be built into the application, based key information about people, labs, and lab results mapped to people.	Reporting capabilities include canned reports and the ability for an epi or a DMAD member to use Tableau to create reporting from data tables

MDM Match/Merge Rules

MDM match/merge rule will determine the type of merge (Auto or Manual) that will take place based on certain matching criteria being exact or fuzzy.

As of now we only auto merge based on 3 rules listed below and the other 7 rules will need data stewards to review the data and determine if they are the same person or not.

Auto Merge (3 Total)
    Rule 1   (Date of Birth (Exact), First Name (Exact), and Last Name(Exact))
    Rule 2   (Date of Birth (Exact), Phone Number (Exact), and Person Name (Fuzzy))
    Rule 3   (Date of Birth (Exact), Address (Exact), and Person Name (Fuzzy))

Manual Merge (7 Total)
    Rule 4   (First Name (Exact), Last Name (Exact), Date of Birth (Exact) and null match null, Phone Number (Exact))
    Rule 5   (First Name (Exact), Last Name (Exact), Date of Birth (Exact) and null match null, Address (Exact))
    Rule 6   (Address_Part1 (Fuzzy), Address_Part2 (Fuzzy), Date of Birth (Exact), First Name (Exact),
                Last Name (Exact), Phone Number (Fuzzy))

Ways of Working

Two-week sprint delivery will be performed, and the sprint ceremonies will align as shown in the diagram below.

Monday	Tuesday	Wednesday	Thursday	Friday
Salesforce (w. AWS) Disease Investigation Stand Ups & Backlog Refinement (Tuesday and Friday). Backlog Refinement Sessions Throughout Week				
Data Engine & Tableau Stand Ups & Backlog Refinement (Monday, Wednesday and Friday)				
Scrums for Cancer Programs, Tobacco Control, and Chronic Diseases (Daily)				
Scrums for DOCIIS 1.0 to 2.0 (Daily)				
Stand Ups & Backlog Refinement for Cancer Registry (Tuesday and Thursday)				
SteerCo (Every Week)	Sprint Demos (Every Other Week)		Roadmap Planning (Weekly)	
	Stand Up for Vaccine Breakthrough Reporting (Weekly)			
			Deployment to Production (Every Other Week)	

Primary Users

"Stacey"
Disease Investigator

Interviews and supports cases of COVID-19 and Notifiable Diseases

- Responsible for contacting people who have tested positive for specific diseases. Validates patient's identity, shares details of results with patient and collects exposed contacts information
- Contacts and refers exposed people, businesses and places for testing and provides them access to other support material/information

"Jacob"
Provider

Submits COVID-19 and Notifiable Disease information as required by law

- Responsible for filling out forms including patient demographic and health information. Provides initial diagnosis of potential diseases related to the patient

"Rachel"
Data Management & Analytics

Leverages data to create reporting and analytics for infectious diseases

- Uses Data Engine to drive advanced analytics and reporting via Tableau capabilities to provide reporting and analytics to epidemiologists and health leaders to help inform decisions

"Julie"
Director of DC Health

Drives public health policy using data

- Consults metrics and dashboards daily to view DC public health at a macro level with the ability for ad hoc reporting and analytics
- Consults team for ad hoc reporting and analytics to drive public health policy to ensure population health and potentially save lives

Onboarding New Source into Data Engine

1. Identify the source data attributes

Work with product owners to gather requirements and create source to target mappings.

2. Create the table structures

Re-create the table structures in the data engine to store the data similar to the source system.

3. Load the data

Ingest the data into Landing and perform any preliminary cleansing on the data such as address enrichment and name/date standardizations.
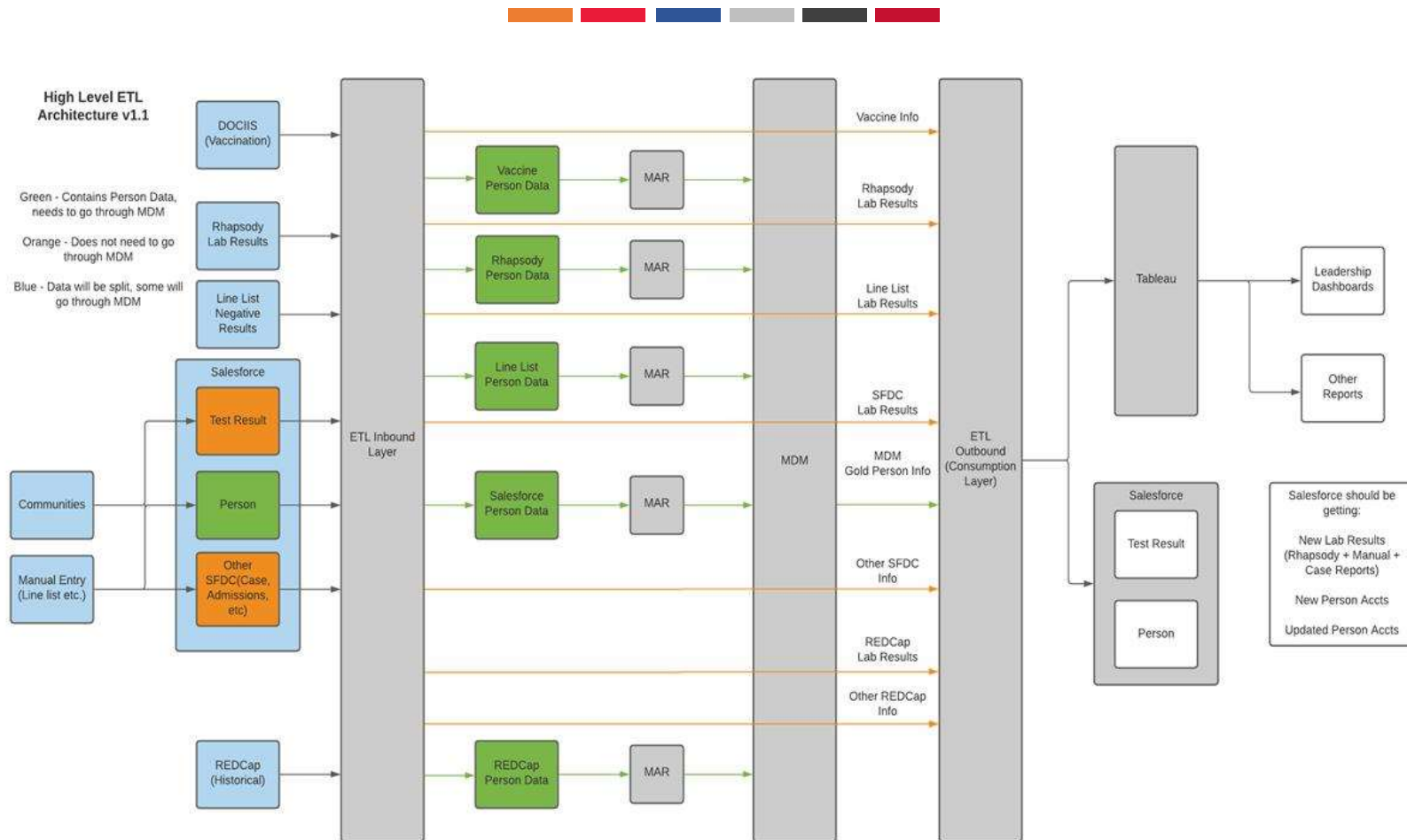
4. Match/Merge the data for Gold Person ID

Create golden records (best value of truth) for unique persons using the approved match and merge rules.

5. Publish the data to the Consumption Layer

Reconcile the source data sets and populate the golden person ids so we can identify the same person across another data sources.
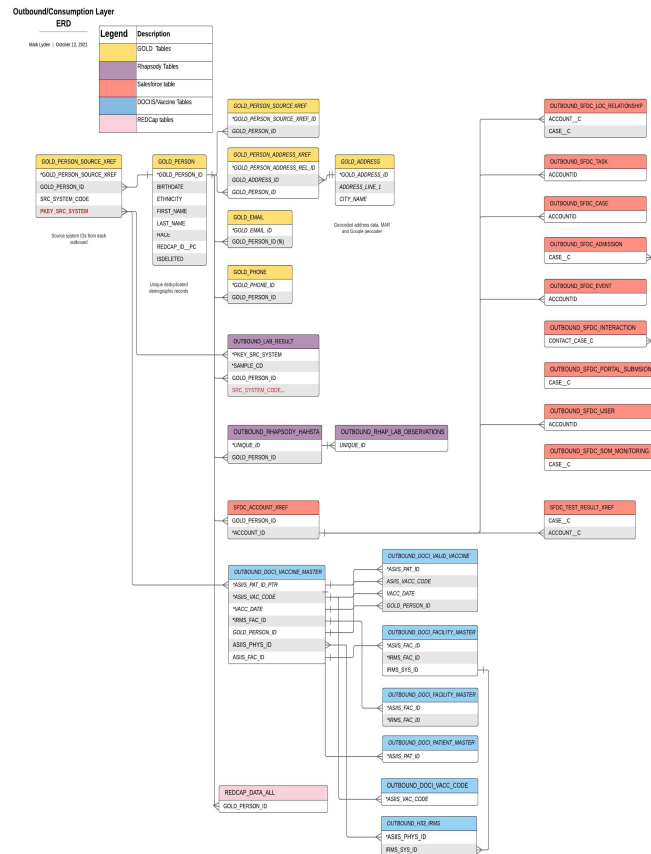
Data Engine Architecture

DC Health Disease Surveillance Conceptual Architecture
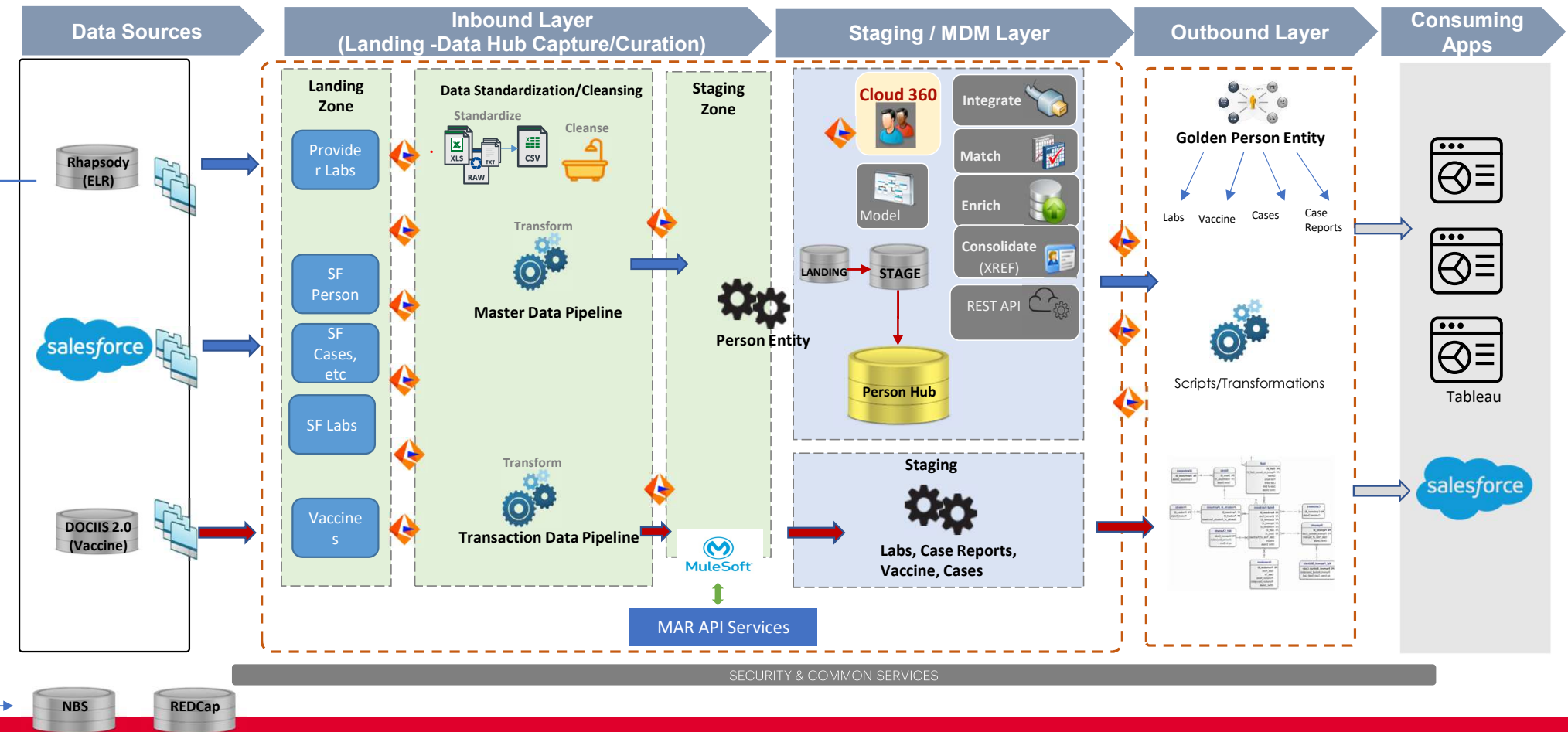
Outbound/Consumption Layer ERD

Key Focus Points

1. Core Platform

- Cloudera Data Hub (CDH) 5.13 End of Life --> Cloudera Data Platform (CDP) 7.1.7 Latest – New features, enhancements, bug fixes

2. Security

- Apache Sentry --> Apache Ranger – stronger, tag based security

3. Improved Data Governance and Auditing

- Cloudera Navigator --> Apache Atlas

4. Better Data Processing

- Apache Spark 1.x --> Apache Spark 2.x (and can upgrade to 3.x)

Post Upgrade

1.Kerberos Change

- MIT Kerberos --> AD Kerberos – allows you to automatically get a Kerberos ticket based on your Active Directory role

2.Expansion of Tools and Connected Platforms

- Requests for more Data Lake side tools --> R-Studio, etc.

3.Data Anonymization

- Integration of Kogni Data Anonymization tool with Data Lake

4.Project Support

- On-boarding more data and projects using machine learning

Data Report

Matthew Crossett
Enterprise Mapping Project Manager

matthew.crossett@dc.gov

New Datasets

- Comprehensive Plan 2021
  - Generalized Plan Policy
  - Future Analysis Areas
  - Resilience Analysis Areas
  - Future Land Use
  - State of Washington DC in 2021

- Traffic Signals

All available at
https://opendata.dc.gov
Published since last meeting, 9/16/2021
Contributors:
DOH
OP
DDOT

Brain Health Initiative
- Advanced Care and End of Life Planning
- Community-Based Dementia Care
- Case Management Care Planning and Care Navigation
- Connection to Services Helplines and Crisis Care
- Food and Nutrition
- Healthcare for Homebound Older Adults and End of Life Care
- Home-Based Homemaker Services
- Home-Based Personal Care Aides
- Individual and Home Safety
- Medication Support
- Money Management and Financial Literacy
- Non-Emergency Transportation
- Residential Long Term Memory Care
- Social Connection Engagement and Networking
- Brain Health Web-Based Resource

octo

Updated Datasets

- **Child Development Centers**
- **Camera Enforcement Locations**
- **Moving Violation August and Sept**
- **Parking Violation August and Sept**
- **Roadway Alleys**
- **Roadway Driveways**
- **Roadway Major Roads**
- **Roadway Ramps**
- **Roadway Service Roads**
- **Snow Removal Areas and Zones**
- **Wetland Inventory**

All available at https://opendata.dc.gov

Published since last meeting, 9/16/2021

Thank You for Joining

For notes and presentation material visit:

https://octo.dc.gov/page/interagency-data-team